Structure and variability of optogenetic responses identify the operating regime of cortex

Agostina Palmigiano[1] Francesco Fumarola [3,1,8] Daniel P. Mossing [5,8] Nataliya Kraynyukova [4] Hillel Adesnik [6,7] Kenneth D. Miller [1,2]

[1] *Center for Theoretical Neuroscience, College of Physicians and Surgeons and Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY*
[2] *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, NY*
[3] *Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa,Wako, Saitama, Japan*
[4] *Max Planck Institute for Brain Research, Frankfurt, Germany*
[5] *Biophysics Graduate Group, University of California, Berkeley, United States*
[6] *Department of Molecular and Cell Biology, University of California, Berkeley, United States*
[7] *The Helen Wills Neuroscience Institute*
[8] *These authors contributed equally to this work*

Contact Information:

Agostina Palmigiano
Center for Theoretical Neuroscience, Zuckerman Institute
3227 Broadway
New York,10027 NY
e-mail: ap3676@columbia.edu

Ken Miller
Center for Theoretical Neuroscience, Zuckerman Institute
3227 Broadway
New York,10027 NY
e-mail: kdm2103@columbia.edu

Abstract

Identifying the regime in which the cortical microcircuit operates is a prerequisite to determine the mechanisms that mediate its response to stimulus. Classic modeling work has started to characterize this regime through the study of perturbations, but an encompassing perspective that links the full ensemble of the network's response to appropriate descriptors of the cortical operating regime is still lacking. Here we develop a class of mathematically tractable models that exactly describe the modulation of the distribution of cell-type-specific calcium-imaging activity with the contrast of a visual stimulus. The model's fit recovers signatures of the connectivity structure found in mouse visual cortex. Analysis of this structure subsequently reveal parameter-independent relations between the responses of different cell types to perturbations and each interneuron's role in circuit-stabilization. Leveraging recent theoretical approaches, we derive explicit expressions for the distribution of responses to partial perturbations which reveal a novel, counter-intuitive effect in the sign of response functions.

Introduction

The presentation of a stimulus to a network of neurons elicits responses that qualitatively depend on the operating regime of the circuit[1,2]. Identifying it, is a prerequisite to predict neuronal activity levels in response to novel stimuli. Converging evidence from experimental and theoretical studies suggests that a defining feature of the operating regime of cortex is strong recurrent excitation that is stabilized and loosely balanced by recurrent inhibition[1,3–7]. This understanding was achieved through the discovery of a fundamental link between inhibition stabilization and response to perturbations established by classic models of recurrent networks with a single inhibitory type[1,2]. In these models, an increase in the input drive to the inhibitory population would elicit a simultaneous decrease of the excitatory and, *paradoxically*, of the inhibitory steady-state activity, if and only if the circuit was inhibition stabilized. This link was probed and verified in multiple experiments[1,3–5], demonstrating that the response of the circuit to controlled perturbations is a successful descriptor of the cortical operating regime.

Those earlier theoretical approaches suffer from three fundamental limitations that prevent linking the diversity of the network's response to the circuit's operating regime i) inhibition is not monolithic and is instead subdivided in multiple cell classes; how inhibition stabilization is implemented at the cell-class level is not understood; ii) cortical networks are rich in parameter heterogeneity; an understanding of the role of same-cell-type diversity in model predictions is currently lacking; iii) new models, if they hope to account for this emerging complexity, will be rife with parameter degeneracy; yet, a data-driven framework designed to subselect from the universe of such models has not been established.

The range and influence of the first issue has recently been quantified. The inhibitory sub-circuit is composed of multiple elements with three types – parvalbumin-(PV), somatostatin- (SOM), and vasoactive-intestinal-peptide (VIP) expressing cells that constitute 80% of GABAergic interneurons in the mouse primary visual cortex (V1)[8]. These inhibitory cell types do not act in unison, but differentially contribute to response tuning[9–12] and to contextual[13–17] and behavioral-state modulations[18–21]. In particular, two of these types, SOM and VIP, engage in competitive dynamics whose outcome directly regulates pyramidal cell activity via inhibition or disinhibition. It is currently debated how the stabilization of strong recurrent excitation is implemented by these interneurons, and whether increasing the input to a particular inhibitory type would lead to a paradoxical decrease of its activity as it is observed when the entire inhibitory sub-circuit (i.e. all GABAergic types) is perturbed. Importantly, these interneurons form a microcircuit characterized by a specific connectivity pattern[22–24], but the relation between the connectivity pattern and the circuit's response to perturbations has not yet been established.

The diversity of same-type cells responses is ubiquitous in neuronal activity[14,17] yet a theoretical framework elucidating how this broad distribution of responses could aid the identification of the circuit's operating point is lacking. Previous work on inhibition stabilization has assumed *homogeneous* populations[2,25], in which all same-type cells are identical and identically connected, fundamentally limiting the capacity of these models to be mapped to actual cortical circuits. Different theoretical approaches[26,27] have explicitly derived descriptors for the mean activity of heterogeneous populations, accounting for same-type cell diversity, and have provided a connection between those two levels of description, namely *homogeneous* or low-dimensional vs *heterogeneous* or high-dimensional description. Yet a framework that predicts the responses of heterogeneous circuits to perturbations, and relates these responses to the *homogeneous* case, is lacking. Furthermore, current optogenetic perturbations themselves are far from homogeneous. In most animal species the accessible toolbox for opsin expression is via local viral injection, infecting only a fraction of the cells in the relevant local circuit. The inconclusive outcome of partial optogenetic perturbations *in vivo*[5,28] is at odds with the expected *paradoxical* response of inhibitory neurons in ISN circuits, demonstrating that low-dimensional descriptions are insufficient to account for the response to concrete optogenetic manipulations, in which cell, cell-type, and perturbation diversity play an important role.

Finally, a fundamental trade-off prevents bridging concept-building theoretical constructs to the biologically realistic modeling necessary to identify a cortical operating regime. As biological realism increases, making parameter-independent predictions or even locating the parameters that situate a biologically insightful model in the correct network state becomes exponentially difficult. Current approaches range from detailed *heterogeneous* network models[29,30] which lack mathematical tractability, to simpler and insightful *homogeneous* models whose parameters are either chosen according to currently available connectivity data[31–34] or to account for stimulus-evoked responses[17,20], but that neglect within-cell-type diversity and are therefore limited in their predictive power.

The increasing versatility in the perturbation patterns introduced by holographic and cell-type-specific optogenetic techniques indicates that time is ripe to develop novel descriptors for the circuit's operating point suitable to characterize heterogeneous circuits with multiple cell types and data-compatible models in which to evaluate them. We have charted out a three-stage program to broaden the predictive power of cortical circuit models. Firstly, we used calcium-imaging recordings of the activity of each interneuronal type in response to stimuli of increasing contrast, to single out a family of low-dimensional system models that fit the data. We report that our fitting method, remarkably, provides sets of parameters endowed with key aspects of the structure of the connectivity matrix found in the mouse visual system[22,23]. By studying mathematically the implications of this structure for the response to cell-type-specific perturbations, we predict a parameter-independent symmetry between the responses induced by perturbation of VIP or of SOM, two interneuron types involved in a disinhibitory micro-circuit whose competition directly regulates pyramidal cell activity. We find that this hidden symmetry principle is respected with remarkable reliability in the models that fit the data. Furthermore, we establish a mathematical link between cell-type-specific response to perturbation and sub-circuit stability. By implementing those insights in these data-compatible models we provide new evidence, aligned with convergent experimental[15] and theoretical[33] arguments, that PV interneurons play a major role in circuit stabilization. Subsequently, we build upon the link between low-dimensional and high-dimensional models provided by mean-field theory[26,35], and are able to construct a family of high-dimensional rate models that fit the experimentally observed distribution of activity of each cell-type and its dependence on contrast. Finally, using recent results in random matrix theory[36], we obtain explicit expressions for the mean and variance of the distributions of responses to patterned optogenetic perturbations. We find, surprisingly, that even if the distribution of activity is spread out, the distribution of responses to optogenetic perturbations need not be. Furthermore, we find that when effecting cell-type-specific partial perturbations, the fraction of cells that respond paradoxically has a non-monotonic dependence on the fraction of stimulated cells. There is a range in which increasing the number of stimulated cells actually decreases the fraction of paradoxically responding cells, yielding a *fractional paradoxical effect* that can be linked to the loss of circuit stability in the context of partial perturbations, and which opens a new avenue for experimental inquiry.

Results

Classic models revealed that the response to controlled perturbations could be interpreted to characterize the operating regime of cortex. Because these models are low-dimensional (hereafter "low-D"), i.e. they have as many units as populations, and because they include one excitatory and only one inhibitory type, the circuit's response to perturbations is linked to the stabilization of the circuit (see Eq. (S11)). When recurrent excitation is strong and stabilized by inhibition (an inhibition-stabilized network or ISN, Fig. 1**a**), an increase in the external input drive to the inhibitory population results in a *paradoxical* decrease of its steady-state activity (Fig. 1**b**). The paradoxical response is then a signature of the ISN condition, occurring if and only if the network is an ISN (Fig. 1**c**). In circuits with multiple inhibitory types, the condition for inhibition stabilization (Fig. 1**d**) is identically defined as in the single-type inhibitory case, but it is linked to the paradoxical response of the total inhibitory *input current* to the excitatory population[37,38] and not to the response of any particular cell-type. In fact, the paradoxical response of a given cell-type is not a predictor of the ISN condition and its implications for circuit stabilization are not understood (Fig. 1**e,f**)[32]. Whether there is a link between the responses to perturbations of different inhibitory sub-types also remains unresolved. Low-D models neglect within-cell-type diversity, and therefore fail to predict the range of cellular responses found in the cortical circuit. Multiple configurations of high-dimensional (high-D) models, in which there are as many units as neurons in the network, are compatible with the L system response. Identifiers of high-D network structure need to be identified (Fig. 1**g**).

Thus, new descriptors, beyond the paradoxical response of the inhibitory sub-circuit, are needed to identify the operating regime from multi-cell-type, high-D activity. Finding those requires a data-driven approach that describes the typical behavior of models when constrained to reproduce the statistics of recorded neuronal activity. Here, by establishing a framework for the study of low- and high-D data-compatible models and their responses to perturbations, we determine three novel descriptors: i) we establish the link between paradoxical response and circuit stabilization, which fails in general scenarios, and identify which cell-types are responsible for stabilization; ii) we establish parameter-independent relationships between the response to two interneurons involved in disinhibitory control, and observe that fitted models fall in one and not the other category; iii) we link theoretically the responses of low-D and high-D systems to discover a novel type of paradoxical effect in the models that fit the data, which we link to the stability of the circuit in the context of partial perturbation. More generally, We demonstrate how the statistical structure of the high-D circuit can be determined from the statistics of neuronal responses.

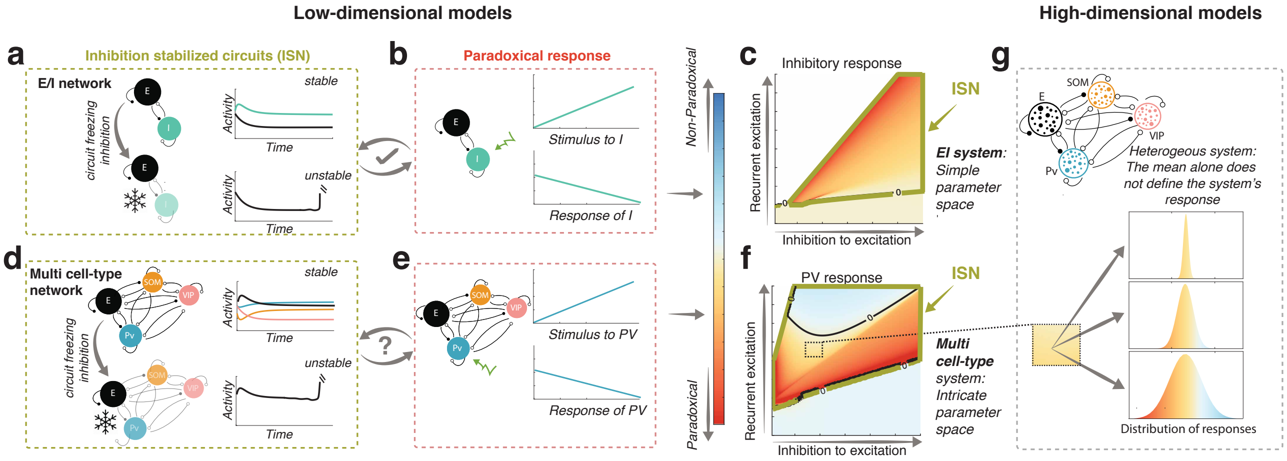

Figure 1 Locating the operating point of the cortex requires novel descriptors. a) Classic models are low-D models with only two units: one excitatory (E) and one inhibitory (I), EI circuits for short. We refer to fixed points of activity which mean a set of steady-state network firing rates in response to a fixed input to the network. The circuit is an inhibition-stabilized network (ISN) at a given fixed point when freezing the inhibitory rate at its fixed-point level renders the excitatory subnetwork unstable. **b)** The steady-state firing rate of the inhibitory population in an ISN paradoxically *decreases* in response to an increase in its input drive. **c)** Response of the inhibitory population to an input increase as a function of network parameters. The network is an ISN (area enclosed by green lines) precisely when the response of I to an increase in its input is paradoxical; in these classic models, there is a one-to-one link between paradoxical response and inhibition stabilization. **d)** In the circuit with multiple inhibitory types the definition of inhibition stabilization is a generalization of that in a): if all inhibitory input to the excitatory population is frozen at its fixed-point level, then the excitatory activity will be unstable. It has been shown[37,38] that an increase in the external drive to the inhibitory sub-circuit (*i.e.*to any or all GABAergic cell-types) will result in a paradoxical response of the steady-state inhibitory input to the excitatory population, *i.e.*either a decrease in excitatory activity accompanied by a paradoxical *decrease* of the total inhibitory input to the excitatory population, or a simultaneous *increase* of the excitatory activity and the inhibitory input to the excitatory population. Nevertheless, whether there are further relationships between stability of network sub-circuits and paradoxical responses in this case is not understood. **e)** Is there a condition on network stability that predicts paradoxical response of an individual population (here, the PV population), in response to an increase in its input drive? **f)** Response of the PV population to an input increase as a function of network parameters of the E-PV-SOM-VIP network. Non-paradoxical responses of the PV population are not restricted to a single region of parameter space (blue shaded area). Although regimes in which the network is inhibition stabilized (the area enclosed by green lines) and regimes in which PV has a paradoxical (red shading, black lines are contour lines separating paradoxical from non paradoxical response) response overlap, they are not predictive of one another. **g)** The response of the low-D systems described in panel e) corresponds to the mean population response in networks in which all neurons of a given cell type have the same properties and inputs (a *homogeneous* network). In systems in which inputs and connections are heterogeneous, the fraction of cells which respond paradoxically can vary depending on the operating point, even if its the response is not paradoxical on average.

Fit of a low-dimensional model recovers mouse V1 connectivity structure

We study the response to contrast manipulations in neurons of layer 2/3 of primary visual cortex (V1) of awake, head-fixed mice. Specifically, we study the responses of Pyramidal (E) neurons and of Parvalbumin (PV), Somatostatin (SOM) and Vasoactive Intestinal Polypeptide (VIP) -expressing interneurons while the animal is shown circular patches of drifting grating stimuli of a small size (5 degrees) at varying contrast. Contrary to what has been previously reported in E/I networks, both experimentally[10] and theoretically[37,39], the activity of inhibitory neurons can decrease with contrast. In particular, VIP activity has a non-monotonic dependence on contrast.

We begin by modeling the data as a low-D circuit with four units, each representing the mean activity of one cell-type population. Each unit has a supra-linear input-output function[39]. All four cell-types receive a baseline input to account for the spontaneous activity observed, while feed-forward inputs only target either E and PV or E, PV and VIP. These inputs are taken to be either a linear function of contrast (Fig. 2, S1) or proportional to the measured activity of L4 pyramidal cells (Fig. S1). To simultaneously find the synaptic connectivity parameters, the value of the baseline inputs, and the values of the stimulus-related inputs, we construct surrogate contrast-response curves for each cell-type by sampling from a Gaussian distribution with mean and standard deviation equal to the data mean and its standard error. We fit each model by finding

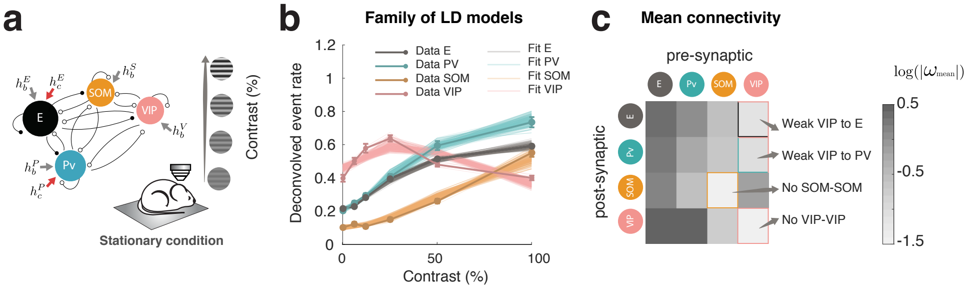

Figure 2 Unconstrained fit to contrast response data recovers mouse V1 connectivity structure. a) Low-D model with multiple inhibitory types. This circuit describes the mean activity of the E, PV, SOM and VIP populations in layer 2/3 of mouse V1. Inputs to each cell-type are composed of a spontaneous activity baseline h_b, and a stimulus related current, h_c, to E and PV, modeling the feedforward inputs from layer 4. In this example, the stimulus-related inputs are linear functions of the stimulus contrast, but similar results are obtained when the input is taken to be proportional to the measured layer 4 pyramidal cell activity instead (see Fig. S1 for different input configurations). **b)** Mean activity of the pyramidal-cell (E, black), PV (turquoise), SOM (orange) and VIP (pink) populations as a function of stimulus contrast, as measured with two photon calcium imagining (thick line, ± s.e.m.). After performing non-negative least squares (see text) to find the 22 parameters of the model (16 weights, 4 baseline inputs and 2 stimulus-related inputs) we find a family of possible models (thin lines, mean and s.e.m. over models, here we show the 300 models) that qualitatively reproduce the mean activity. Note that the activity of SOM and VIP as a function of contrast are mirrored. **c)** The logarithm of the mean connectivity over all possible models is shown (see Fig. S2 for the distributions). Notice that, as in experiments (see Fig. S2), the models lack recurrent SOM and VIP connections, and the connections from VIP to E and PV are small on average.

the non-negative least squares (NNLS)[16,40] solution to each surrogate data set, and select from these, several data-compatible model parameters for which the network steady-states provide the best fit. This optimization takes as sole input the response data and uses no prior information on the synaptic structure, hence it is not obvious that any meaningful synaptic structure should be recoverable from such a procedure. We report that, surprisingly, the structure of the inferred connectivity matrices has a striking resemblance to that reported experimentally (Fig. S2**b**). The family of models fitting the data are endowed with a broad distribution of possible weights that exhibit consistent features: recurrent connections within the SOM population and the VIP population were absent in most models, as observed in mouse V1[22,23] (Fig. S2). Furthermore, whenever inputs were chosen to target only E and PV, VIP interneurons have weak or absent connections to all other cell-types except SOM interneurons, also as reported in mouse V1[22].

We next asked whether the models that fit the data are inhibition stabilized networks. Inhibition stabilization is understood here as one fundamental descriptor of the operating regime[1]. We find that in the absence of stimulus, only few models are compatible with non inhibition-stabilized circuits, a fraction that becomes vanishingly small at full contrast as is consistent with experimental findings[1,5]. This conclusion is independent of the choice of inputs and holds for the four input configurations we analyzed (see Fig. S3). In a multi-cell-type circuit, the ISN condition implies that an increase in the input drive to any or all of the inhibitory populations results, in the new steady state, in either a simultaneous decrease or a simultaneous increase of both the inhibitory input to the excitatory population and of the excitatory activity.[37,38]. Therefore, if a perturbation to the entire inhibitory sub-circuit elicits a paradoxical decrease in activity in all GABAergic cells that project to excitatory cells and in the excitatory activity, then the circuit will be an ISN. The converse, that the ISN condition implies a paradoxical response of the inhibitory activity, is only true in an E/I circuit: in the multi-cell-type case, there are multiple ways in which the total inhibitory input current to the E population can decrease, so no specific cell-type needs to respond paradoxically. In our models, which are almost all ISNs, perturbing all GABAergic cells systematically implies a decrease in PV activity, but whether it causes an increase or decrease in the responses of SOM or VIP depend on the specific choices of stimulus-related input (Fig. S4).

Paradoxical effects in circuits with multiple cell types and link to sub-circuit stabilization

To systematically investigate the response to cell-type-specific stimulation. We focus on the linear response matrix, whose elements $\chi_{\alpha\beta}$ give the change in steady-state response of cell-type α per change in input to cell-type β, for small changes (see Eq. S10). The diagonal elements of this matrix, $J_{\alpha\alpha}$, can be written as a function of the Jacobian J of the entire circuit and the Jacobian J_α of the sub-circuit without cell-type α:

$$\chi_{\alpha\alpha} \propto \frac{-1}{\det J} \det J_\alpha \qquad\qquad \alpha = \{E,P,S,V\} \tag{1}$$

If the response of cell-type α at a given fixed point is paradoxical ($\chi_{\alpha\alpha} < 0$) and the fixed point is stable (implying $\det J > 0$), then the sub-circuit without that cell-type, linearized about the fixed point, will be unstable ($\det J_\alpha > 0$, see Eq. (S10) and Fig. 3**a**). This insight is a simple generalization of the two-population ISN network, in which the I unit shows a paradoxical response at a given stable fixed point when the circuit without it, *i.e.* the E unit, is unstable. It links cell-type-specific paradoxical response to sub-circuit stability in a more general setting.

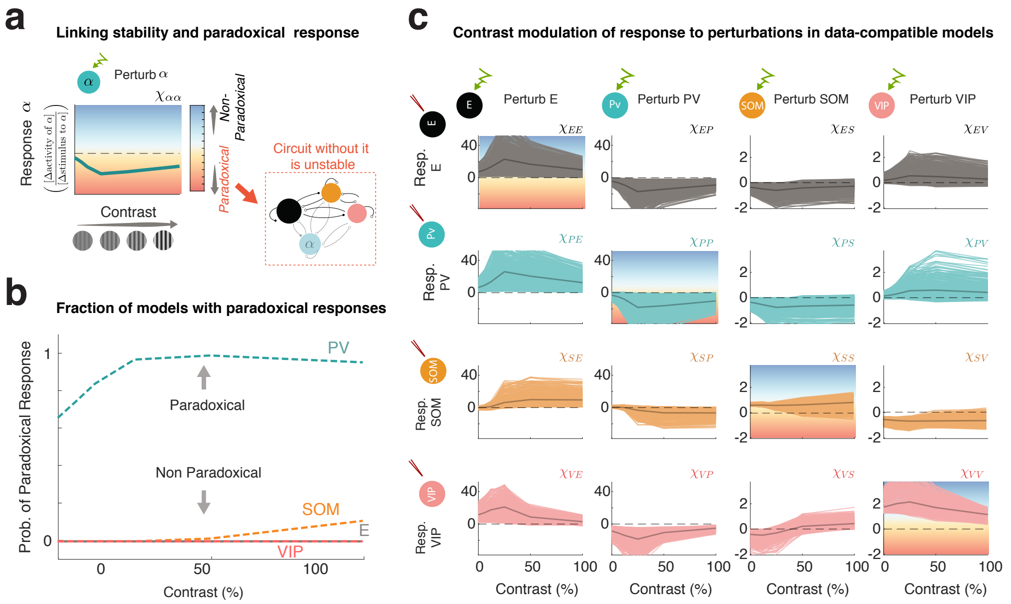

Figure 3 Paradoxical response and circuit stabilization in data compatible models. a) Graphic summary of the relation between stability and paradoxical responses. The response of a cell-type α, which is the change in activity normalized by the size of the perturbation, is shown as a function of contrast. When the response of the cell-type which is being perturbed is negative, the response is paradoxical. Mathematically, that means that the circuit without it is unstable. In the special case in which VIP projects only to SOM, SOM being paradoxical means that the E-PV circuit is unstable. **b)** Fraction of models that fit the data which exhibit paradoxical responses in each of the 4 cell-types, as a function of contrast. **c)** Linear response matrix $\chi_{\alpha\beta}$ quantifying the change in activity that a population α undergoes when extra current is injected in the population β, as a function of contrasts for the models fitted in Fig. 2. Note the paradoxical response of PV at all contrasts and the non-paradoxical response of SOM in most cases. In the multiple cell-type circuit and unlike in the EI system, excitatory activity can in principle also respond paradoxically. Nevertheless, none of the data-compatible models obtained had an excitatory paradoxical response.

We next compute the linear response matrix $\chi_{\alpha\beta}$ as a function of stimulus contrast for the family of data-compatible models. We find that a perturbation to the PV population elicits a paradoxical response in almost all models consistent with recent findings of optogenetic perturbations to the full PV population in mice[5] (Fig. 3**b**,**c**, see also Eq.S46 for a relation between the calcium activity used here and firing rate). When the projections from VIP to E, PV and VIP are very small, the response of SOM to its own perturbation is directly linked to the stability of the sub-circuit E-PV: a paradoxical response in the SOM population indicates that the E-PV sub-circuit is unstable (see Eq. S12).

Consistent with previous work showing that strong perturbations to PV destabilize the dynamics in V1[15], we find that in most models that fit the data i) SOM does not respond paradoxically, consistent with the E-PV circuit being stable, and ii) PV responds paradoxically, meaning that the circuit without it is unstable (Fig. 3**b**,**c**). We conclude that a first novel descriptor of the operating regime of circuits with multiple cell-types, is PV-stabilization, meaning that potential circuit instability is stabilized by PV cells and not by SOM cells (see also[33]). This describes most of our models that are consistent with the data, but about 20% show paradoxical SOM response at higher contrasts (in addition to paradoxical PV response) and so are stabilized by the combination of PV and SOM and not by either alone. In our models, its role is robust and supported by the paradoxical response of PV and the non-paradoxical response of SOM.

Symmetry principles of optogenetic response

We find consistent symmetries in the response matrix. The first set of symmetries involve responses of, or to, E vs. PV: i) the responses elicited in E and in PV are similar when perturbing any cell-type (first two rows of 3**c**, see also Fig. S6**a**); and ii) perturbing E and PV elicits a mirror change in activity in all cell-types (first two columns of 3**c**, see Fig. S6**b**). We hypothesize that these symmetries result from the similarity between the activity of E and PV. To test this, we repeat the fitting procedure in a simplified configuration in which PV activity is proportional to E (see Fig. S7**a**). We find that in that case, the connectivity matrix has linear dependencies: weights from a given cell type to E are a linear (affine) function of that cell-type's weight to PV, and similarly for weights from E vs. those from PV (see Fig. S7**e**). This allows a mathematical understanding of these symmetries: the change in activity in E and PV to perturbations are different linear combinations of the same elements, so that the change in activity elicited by a perturbation to E will be proportional to the one elicited by a perturbation to PV, with an offset and slope that are contrast-dependent (see Eq. S33).

A second set of symmetries involve responses to perturbation of SOM vs. perturbation of VIP. Based on our recovery of the structure of the connectivity matrix found in mouse V1 (Fig. 2), we examined the linear response matrix for a connectivity that obeys the condition that VIP projects only to SOM, and is otherwise arbitrary. We found that in this case, the linear response matrix has a symmetry between the response of E, PV and SOM to a VIP perturbation vs. to a SOM perturbation: for each cell type, the two responses will be negatively proportional to one another, with a common proportionality constant across the three cell types (Fig. 4**a**)). Specifically, if f'_V is the gain of VIP at a particular steady-state configuration and ω_{SV} is the synaptic weight from VIP to SOM then

$$\chi_{\alpha V} = -f'_V \omega_{SV} \chi_{\alpha S} \qquad\qquad \alpha = \{E,P,S\} \tag{2}$$

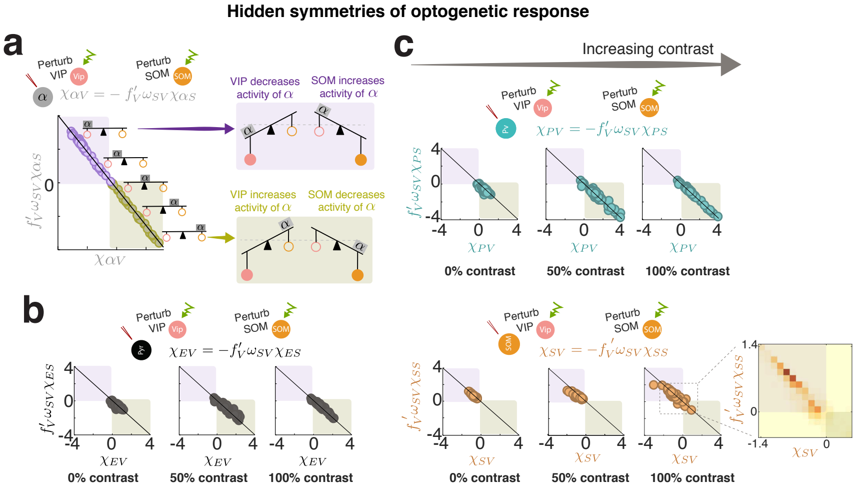

Figure 4 Hidden response symmetries (HRS). We reveal that a parameter-independent relation holds true between the response of E, P, S to perturbation of SOM and VIP. These relations or symmetries, described by Eq. (2), are derived under the assumption of VIP projecting only to SOM. **a)** Illustration of the response of cell-type α to VIP perturbation vs the response to a SOM perturbation multiplied by the coefficient in Eq. (2). Given a perturbation to the VIP population, the constraints imposed by the the HRS define the sign and the magnitude of the response to SOM, so that possible values lie on a line, as shown. Two regimes can be identified: One in which VIP disinhibits while SOM inhibits the cell-type α (lower right quadrant) and another one in which the opposite is true (upper left quadrant). **b)** HRS for E for the top 0.01% of models. Given that the data-compatible connectivities have only small values of connections weights from VIP to E, PV and SOM, this symmetry is evident in the models that fit the data. **c)** Same as b) for PV and SOM. Inset shows the joint probability distribution over models for the SOM case, for visual clarity. Together these results show that the best fit models support a clear disinhibitory motif in which a perturbation to VIP decreases SOM activity and increases both E and PV, and a perturbation to SOM does the opposite.

We call these equalities *"Hidden Response Symmetries" (HRS)*. Even though we fit the models unconstrainedly, allowing all possible connections, we find that the data-compatible models have only weak connections from VIP to other interneurons besides SOM, as in experiments[22,23]. Therefore, and as quantified in Figure 3b, this symmetry in the response is revealed in this family of models. We emphasize that this result is general in this type of low-D systems and formalizes a clear intuition: Because VIP only projects to SOM, a weak perturbation to VIP can only affect the rest of the circuit through SOM, relaying that perturbation with an opposite sign. The HRS defines two regimes: one in which an increase in the input to VIP increases the activity of a given population, and another one in which it decreases it, with SOM causing an opposite response in each case. Specifying this regime provides another good descriptor of the cortical operating point. Our data-compatible models almost all are in agreement as to this regime: in almost all, activation of VIP has a disinhibitory effect on E (see Fig. S5 for a detailed analysis of the effects of VIP on E), as in experiments[17,21,41,42], and disinhibits PV while inhibiting SOM. These effects of small VIP perturbations on PV and SOM, and the opposing, proportional effects on E, PV and SOM of small VIP vs. SOM perturbations, with the same proportionality constant for all, constitute predictions of our analysis. We note that the analysis also predicts that activation of PV suppresses all other cell types (Fig. 3c).

High-dimensional models recapitulate the dependence on contrast of the distribution of responses of all cell types

Understanding heterogeneity both in the cortical circuit and in the perturbation protocol, and analyzing partial perturbations from modern holographic optogenetic stimulation, require models capable of fitting and explaining the diversity of neuronal response. To describe contrast modulations observed within each cell-type population, we build high-D models with different numbers of cells in each population as measured experimentally[8] (Fig. 5**a**). Each neuron has a power-law input-output function[37,39] and receives a specific baseline input and a cell-type-specific stimulus-related input. The connectivity is heterogeneous with a mean and a variance dependent on both the pre- and post-synaptic cell-type. We emphasize that, due to the nonlinear transfer function, heterogeneity in the values of the synaptic connectivity will change the mean activity compared to the *homogeneous* case, in which all the neurons are the same. Consequently, the parameters needed to fit the data in the homogeneous case, which is mathematically identical to the low-D model, will be different than the ones needed to fit the high-D system.

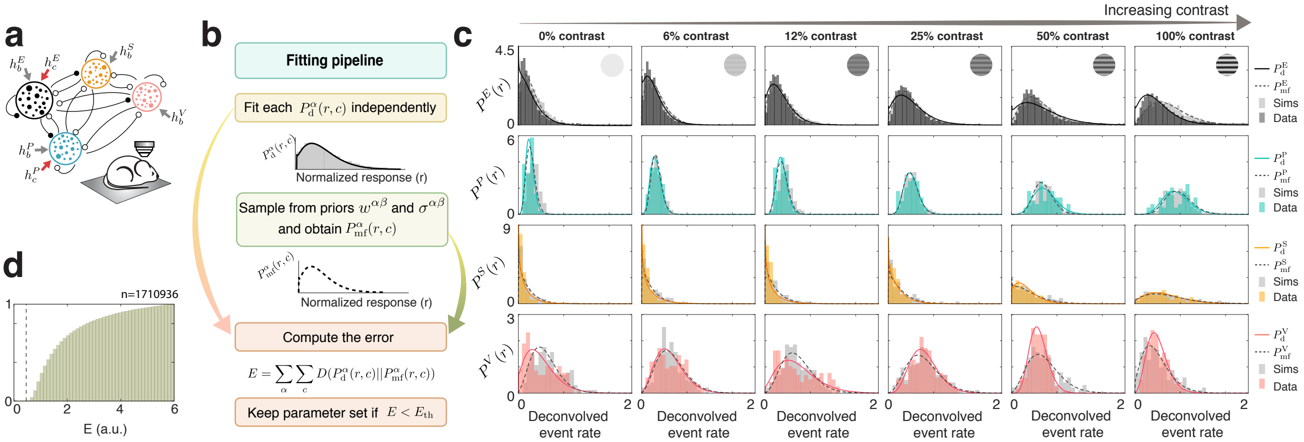

Figure 5 Fitting the distribution of responses to multiple contrasts with a mean-field model. **a)** Diagram of the high-D model. Weights and inputs are heterogeneous. As in the low-D case, inputs are composed of a baseline and a stimulus related signal, chosen to be linear in the stimulus contrast as in Fig. 2. **b)** Fitting pipeline: The distribution of activity generated by a model with a power-law non-linearity and recurrent and feed-forward inputs that are Gaussian distributed has an explicit mathematical form[35] (Eq. (S62)). We used it to fit that form to each of the distributions of activity for a given cell-type at a particular value of the stimulus contrast. A mean field model, with the appropriate parameters should be able to recapitulate the distributions of activity of all cell-types at all values of contrast. To find them, we generate high-D models by sampling from prior distributions of parameters, and compare them to the fitted data distributions using an error function, given by the sum of the Kullback-Leibler divergences of the distributions given by the model ($P^{\alpha}_{\rm mf}$) and the data ($P^{\alpha}_{\rm c}$) for all cell-types and contrast values, which can be found explicitly. By only accepting models with error less than a threshold of 0.45 (top 0.005%), we build a family of suitable models. **c)** Example of a parameter configuration within the threshold. Data (histogram, colored bars) and data fits (solid colored line) are in good agreement with the mean-field theory distributions (dashed gray line) and the simulations of the full high-D model (gray bar histogram). **d)** Distribution of KL divergences, indicating the 0.45 threshold. We used models below this threshold for the analysis in the remaining text (see Methods for details, and Fig. S5).

In order to fit the entire distribution of responses and its dependence to stimulus contrast, we build on two facts. First, given power-law input-output functions, there is a closed-form expression[35] (see also Eq. S62) that links the distributions of responses of a given cell type (α) to the distribution of inputs that population receives. Given that this expression is explicit, it allows inference from the data of the input distributions for each cell-type and each stimulus (contrast c).

Second, for a given high-D circuit model (for a fixed set of parameters), these distributions of inputs and activity can be computed self-consistently through mean-field theory[26,27] (see Eq. S58). To find high-D models that fit the observed data distributions, we build an error function (Fig. 5**b**) that measures the divergence (see Eq. S65) between the distributions fit to data, $P^{\alpha}_{\text{data}}(r)$, and the distributions produced by the mean field equations, $P^{\alpha}_{\text{mf}}(r)$. This error has an explicit expression. To generate candidate models, we sample from prior distributions on the parameters (the means and variances of the weights w and the external inputs h). We keep the solutions that have a sufficiently small error, to define a family of high-D models (Fig. 5**c,d**, see also Fig. S8). This family of models (see Fig. S8) recapitulates the dependence of the distribution of responses of all types on contrast, and captures both the spreading out of the distributions with increasing contrast and the heavy tails of the distributions seen in the calcium data.

Analytical approach to full and partial perturbations

To assess the predictive power of this new family of models that fit the data, we compute the distribution of responses of the network to perturbations, *e.g.* optogenetic activation or suppression of sets of cells. We assume that all neurons in a given population have the same firing rate at the network's fixed point, what we refer to as *homogeneous fixed point* (hereinafter HFP) approximation, see Eq. S72). This makes our system fit the assumptions needed to build on recent work on random matrix theory[36] to compute these response distributions. We thus obtain explicit analytical expressions for the behavior of the mean and the variance of the distributions of optogenetic responses to either full or partial, and either homogeneous or disordered, perturbations. In the following, we will distinguish analytics using the HFP approximation, from simulations of the fully nonlinear system, in which different cells of a given cell-type can have different firing rates at the network's fixed point.

First, we investigate which aspects of the distribution of responses to perturbations in the high-D system are linked to the low-D case. Importantly we find that under the HFP approximation, neurons belonging to a specific cell-type's population of the high-D heterogeneous system, have a mean response to cell-type-specific perturbations given by the response of the system without heterogeneity (see Eq. S138), allowing us to directly link the response of the low-D and high-D models. Figure 6**a** shows the distribution of responses to cell-type-specific perturbations of the network in Figure 5 in the absence of visual stimulation (see also Figs. S9 and S10). The distributions obtained under the approximation are in good agreement with the results of simulations of the fully non-linear system.

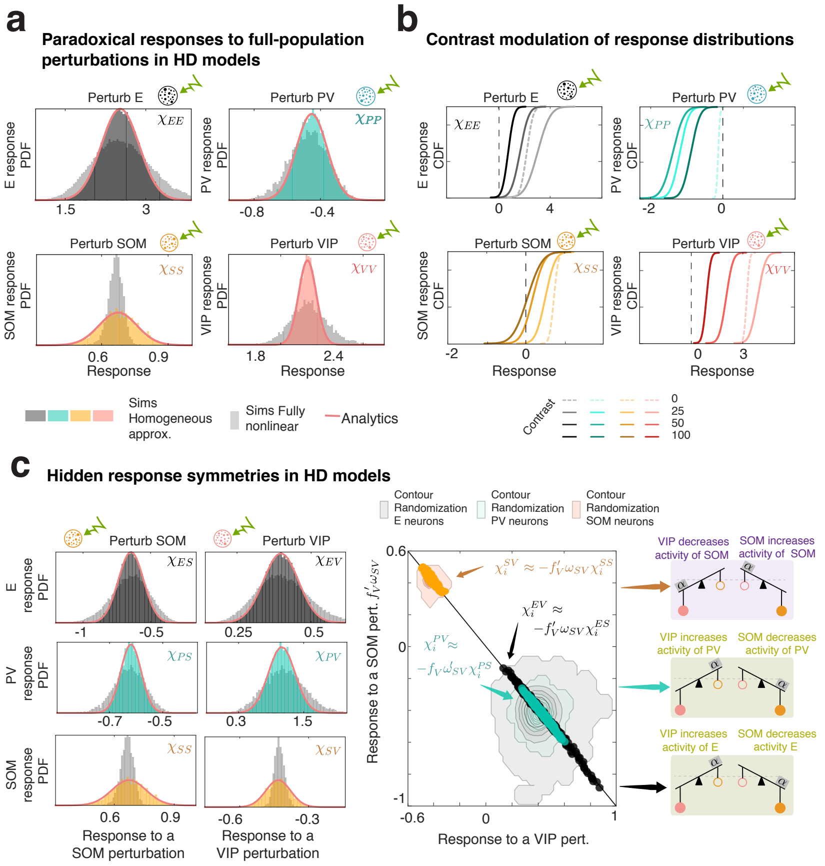

Figure 6 Distributions of responses to full-population perturbations in high-D systems: Here we consider perturbations to all cells of a given cell type ("full-population perturbations") in the high-D model shown in Fig. 5. **a)** Distribution of responses to full-population perturbations in the absence of visual stimulation (see Fig. S11 for the entire family of models). Gray histograms are the result of the simulation of a fully high-D nonlinear system, while colored histograms are simulations of the network within the HFP approximation. Orange lines are the analytical result, only possible with that same approximation. The responses of E, SOM and VIP cells are not paradoxical, while all cells in the PV population respond paradoxically to PV stimulation. **b)** Cumulative distribution of responses (analytics) to full-population perturbation in the presence of a visual stimulus for varying stimulus contrast. Increases in contrast modulate the variance of the responses in anti-intuitive ways. In some cases, contrast increases widen both the distribution of stimulus-driven activity and the distribution of responses to full-population perturbations spread out (SOM). Other cases either have a non-monotonic variance with contrast (PV) or become narrower with contrast, opposite to what was found for the stimulus-driven activity (compare Fig. 5, see also Fig. S11 for analysis on the variance) **C)** Hidden response symmetries in high-D systems: Opposite and proportional responses to perturbations of SOM and VIP. Left: Distribution of E, PV, and SOM responses to SOM and VIP perturbations. Right: responses of single cells of type E (black), PV (turquoise) and SOM (orange) cells to a VIP perturbation, vs the responses of those same cells to a SOM perturbation multiplied by the factor: $f'_V \omega_{SV}$ (see Eq. 2). The response symmetries that we had derived for the low-D system hold at the single cell level in the fully nonlinear high-D system. In experiments, it may be necessary to compare the response of one cell to a SOM perturbation and a different cell to a VIP perturbation. The contour lines show the distribution of such responses across pairs of cells, with VIP perturbed for one and SOM perturbed for the other. In this case, the response to VIP and to SOM perturbations are not correlated.

It is conceivable that, given broadening of the distribution of activity as a function of contrast (Fig. 5), an optogenetic perturbation would result in increasingly broad distributions. Counter-intuitively, we find that the spread of the distribution of activity to visual stimuli is rarely predictive of the spread of responses to stimulation. The responses to perturbation can become narrower with increasing contrast, even if the activity itself broadens (Fig. 6**b**, see also Figs. S9, S10 and Fig. S11 for a quantification). In all the models analyzed, the entire distribution of PV responses to a full PV perturbation was paradoxical whereas SOM has a non-zero fraction of paradoxically responding cells only at higher contrasts. We can relate these findings back to the stability of the network sub-circuit. The mean of the response distribution can be shown to be the response of an equivalent low-D system (under the HFP approximation), and therefore a paradoxical mean response of a give cell-type is

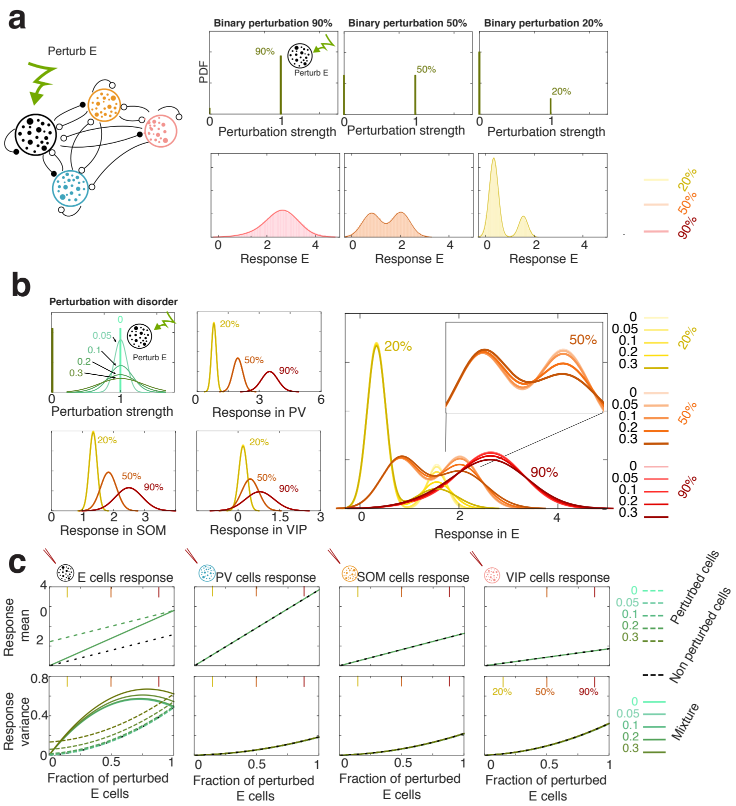

Figure 7 **Framework to describe the response to patterned perturbations. a)** Partial homogeneous perturbation to the E population: perturbation of a fraction of E cells with uniform strength. Bi-modality of the response distribution is only noticeable if the fraction of perturbed cells is small enough. **b)** Response to *partial heterogeneous* type perturbation: perturbation of a fraction of E cells with varying strengths. Top left: perturbation strength is distributed as a Gaussian, with mean 1 and normalized to integrate to 1, with varying standard deviations equal to 0 (a *partial homogeneous* perturbation), 0.05, 0.1, 0.2, 0.3. For each strength distribution, either 20%, 50%, or 90% of E cells are perturbed. The responses of cells that do not receive the perturbation are sensitive to the number of cells perturbed but not to the variance of the "opsin expression" distribution shown in the top-left panel (they are sensitive to its mean, not shown, see Eq. S136). **c)** Dependence of the mean and variance of the distribution of responses as functions of the fraction of stimulated E cells. As only the E cells are stimulated, they are the only population exhibiting a distribution of responses that is a mixture of two Gaussians. Top row: the means of all Gaussians contributing to the mixture (color-dashed for perturbed cells and black-dashed for non-perturbed cells). Bottom row: the variance of each Gaussian (dashed lines) and, for E, of the mixture of the two (full lines).

indicative of the instability of the circuit without that cell-type (see Fig. S12).

Finally we quantify to which extent the mathematical understandings offered by the *Hidden Response Symmetries* hold at the single cell level. Remarkably we find that although the distribution of responses to SOM and VIP are broad, the response of a single cell to a SOM and VIP perturbation is perfectly anti-correlated in the fully nonlinear system (Fig. 6**c**), and respect the relationships defined in the low-D case (Eq. 2). Our models suggest that this symmetry in the responses to VIP and SOM stimulation is a good descriptor of the circuit's regime both at the population and the single cell level.

Having found a model that reproduces the dependence of the distribution of activity on contrast, and established an analytical framework to study the distribution of responses to arbitrary perturbations, we now investigate the response to perturbations that do not affect all same-type cells equally. We studied two types of heterogeneous perturbations. In one, which we call *partial homogeneous* in which a perturbation affects only a fraction of cells, but those cells are stimulated identically. The second type of perturbation differs, in that those cells that are perturbed are not perturbed identically but instead in a heterogeneous way (akin to heterogeneity of opsin expression in optogenetic experiments). We name this last type *partial heterogeneous*. We find mathematically, that the distribution of responses of the population that is stimulated is bimodal, given by a mixture of Gaussians (one gaussian corresponds to the perturbed population and the other one to the unperturbed one; see

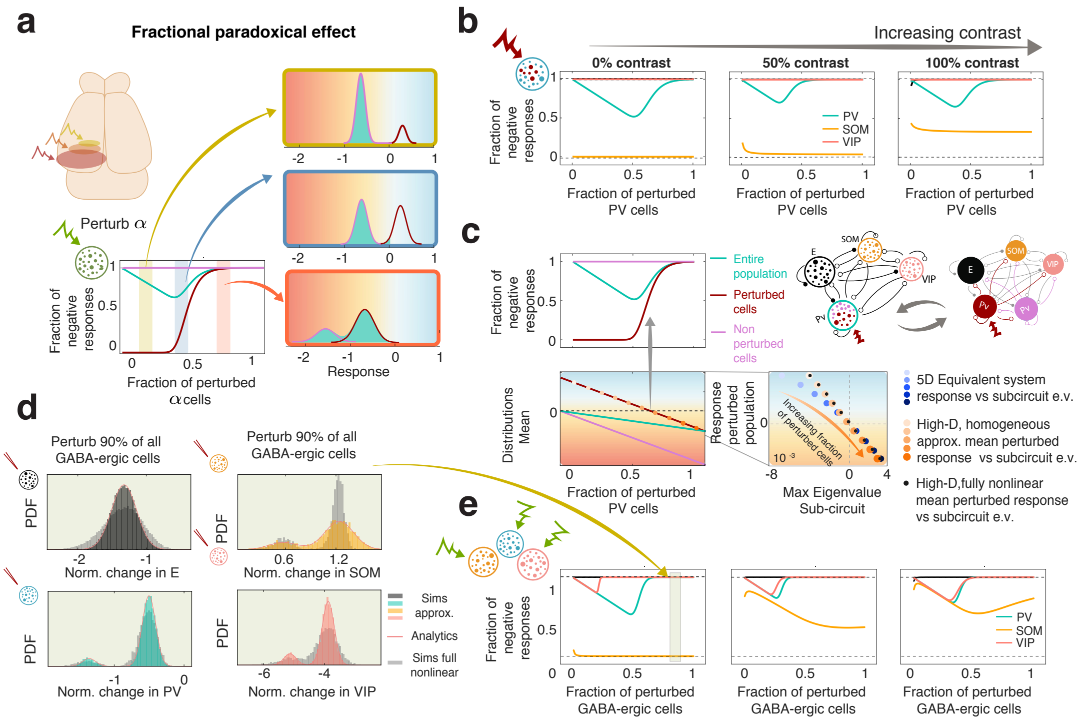

Figure 8 Fractional paradoxical effect. **a)** Top left: scheme of a mouse brain in which three different fractions of a cell populations of the local cortical microcircuit are stimulated. Top right: Partial perturbations result in a bimodal distribution of responses. The rightmost peak (red envelope) corresponds to the stimulated cells, while the leftmost peak (magenta envelope) corresponds to the non-stimulated cells. When stimulating a small number of cells, the bulk of the distribution of responses (cyan shading) is negative. Increasing the fraction of perturbed cells (middle and bottom right panels) first decreases, then increases the number of paradoxically responding cell. Bottom left: Fraction of negative responses as a function of the fraction of stimulated cells shows a non-monotonic dependence, which we name the *fractional* paradoxical effect. **b)** The fraction of negative responses of the PV, SOM and VIP populations as a function of the fraction of stimulated PV cells, for three different values of the contrast. PV shows a fractional paradoxical effect in all fitted models (see Fig. S13). **c)** For a given fraction of perturbed PV cells (x axis), the curves show the fraction of negative responses in each of the Gaussian distributions composing the mixture of PV cell responses (perturbed cells, dark red; unperturbed cells, magenta), and in the full PV distribution (turquoise) (see Eq. S111). Bottom left: The means of the same distributions, using the same colors. Note that the mean of the perturbed population crosses zero when its fraction of negative responses crosses 50% (arrow). Top right: The mean response of the perturbed population can be mapped to the response of a PV sub-population in a low-D system with two PV populations, a perturbed one (red) and an unperturbed one (magenta) see, Eq. (S123). Whenever the mean of the perturbed cell population becomes negative, so does the response of the perturbed cell type in the 5-D system, indicating that the fractional paradoxical effect is linked to a transition to instability of the non-stimulated 4-D sub-circuit of the 5-D system. Bottom right: The response of the perturbed population of the 5-D equivalent system (blue) and the mean response of the perturbed sub circuit in the high-D system under the HFP approximation (orange) and for the full nonlinear system (black) are shown as a function of the maximum eigenvalues of the non-perturbed sub-circuit in each case. The mean responses become negative when the maximum eigenvalue crosses zero, indicating instability of the 4-D non-perturbed sub-circuit. **d** Perturbation showing the response of all cells when stimulating 90% of all interneurons in the absence of visual stimulus, as computed analytically (line) the HFP approximation simulations (histogram in colors) or the fully nonlinear simulations (histogram in gray) **e)** Same as (b) but while stimulating simultaneously all GABAergic cells. The arrow to the gray region around fraction 0.9 links the distributions in d) with this panel. We note that when doing this manipulation the non-monotonic dependence on this fraction is a prominent feature, which is recovered in all analyzed models (see also Fig. S13).

Eq. S111 for *partial homogeneous* perturbation and Eq. S136 for *partial heterogeneous* perturbation). An example of a *partial homogeneous* perturbation can be found in Figure 7**a**, and an example of a *partial heterogeneous* perturbation in 7**b**. The mean of each of the Gaussians forming the response distributions is linear in the number of perturbed cells (Fig. 7**c** top row). The different contributions of the fraction of cells perturbed and the heterogeneity of the perturbation can be understood from the expression for the variance of each of the Gaussians in the mixture (Eq. S130 and S129). The variance (Fig. 7**c** bottom row) is nonlinear (and potentially non-monotonic) in the number of perturbed cells. Note that the behavior of non-stimulated cells strongly depends on the mean perturbation strength but not on its variance in the homegeneous-fixed-point approximation used to derive these equations (but does depend on the variance in the fully nonlinear system).

Fractional paradoxical effect

When analyzing responses to a perturbation of a fraction of cells of a given population, we find a surprising effect: with increasing fraction of cells perturbed, the fraction of cells of that type responding negatively (paradoxically) can show non-monotonic behavior. Over some range, increasing the number of stimulated cells decreases the probability that we will measure a cell showing paradoxical behavior. We name this the *fractional paradoxical effect* (Fig. 8**a**). This result extends the concept of critical fraction developed in Ref.[25] to the full rank case with multiple cell-types and heterogeneity.

Figure 8**b** shows the dependence of the fraction of negative responses of each inhibitory population on the fraction of perturbed PV cells. Intriguingly, in the models that fit the data, PV has a fractional paradoxical response. When stimulating about 50% of cells, about 50% show negative responses. This is consistent with what is observed experimentally in Ref.[5], where an optogenetic perturbation of PV interneurons with transgenic opsin expression (expressed in essentially all PV cells) elicits a paradoxical effect in most cells, whereas if the expression is viral (expressed in only a fraction of PV cells), only about 50% of cells show negative responses. The critical number of PV cells that are needed for the fraction of paradoxically responding cells to increase with the fraction of perturbed cells (i.e. the minimum of the turquoise curve in Fig. 8**b**) can be expressed, but only found numerically (the variance depends non-linearly on the fraction of stimulated cells).

The fractional paradoxical effect in a population is linked to the stability of the unperturbed sub-circuit (Fig. 8**c**). The mean response of the perturbed cells in the HFP approximation can be mathematically linked to the response of a low-D system with two PV populations, a perturbed one (red) and an unperturbed one (magenta), representing the perturbed an unperturbed population in the high-D system respectively. We show that whenever the mean response of the perturbed population becomes negative in the high-D system, the sub-system composed by all of the non-perturbed populations in the associated low-D system will lose stability, corresponding to a paradoxical response of the perturbed PV sub-network. The eigenvalues of the Jacobian of the non-perturbed sub-system, both in the high- and the low-D systems, become positive (indicating instability) at the same fraction of perturbed cells as the change in sign of the response of the perturbed cells. This understanding links stability of the non perturbed circuit to the fractional paradoxical effect: whenever the system exhibits a fractional paradoxical effect, the unperturbed neurons will form a stable circuit, which will lose stability only after a critical fraction of cells are stimulated. Finally, when increasing the fraction of perturbed cells in the entire inhibitory sub-circuit (Fig. 8**c-d**) we see that at high contrast the fractional paradoxical effect will be observed in all cells.

Discussion

We leveraged methods of nonlinear regression, mean-field theory, and random matrix theory to find low and high-D models of cell-type-specific response to contrast modulations, and to derive closed expressions for the mean and variance of the distributions of responses to heterogeneous and partial optogenetic perturbations. This framework, allowed us to identify and propose three useful descriptors of the circuit's operating point, suitable for multi-cell-type high-D circuits: i) cell-type response to an input increase and its link to sub-circuit stabilization, ii) the regimes of response to SOM and VIP manipulations identified via the hidden response symmetries, and iii) the existence of a fractional paradoxical effect, linked to the stability of the non perturbed network.

By fitting the mean activity of each interneuron type in response to contrast manipulations, we uncovered the underlying structure of the synaptic connectivity observed in mouse V1[22,23] (Figs. 2, S1 and compare Fig. S2). Some of those features, like the lack of recurrence within the VIP and SOM populations are independent of fitting choices, suggesting that there are certain features of the dynamics that implicitly carry information about the connectivity, and that those can be revealed by choosing a suitable model. We focused on small stimulus sizes in order to avoid the involvement of longer-range circuits evoked by larger stimulus sizes[6,14,15], which would presumably involve models with spatial structure[43]. Such models could offer further constraints to the synaptic structure found here.

Our mathematical analysis resulted in a number of insights on the response to weak, cell-type-specific perturbations. In our models, a strong similarity in the responses of PV and pyramidal cells to optogenetic perturbation (Figs. 3,S6,S7) stems from the similarity between the contrast tuning curves of E and PV in this dataset. We found indeed that forcing the activity of PV and E to be proportional, models that fit the data have parameters which are linearly dependent, leading to similarities in the optogenetic responses (Fig. S9). We expect these symmetries to hold only for certain stimulus configurations, but not generally. In contrast, in low-D circuits in which VIP only projects to SOM, the response of all cell-types to small perturbations to SOM or VIP are perfectly anti-correlated, independently of stimulus configuration or parameter choice. Responses of E, PV, or SOM cells to perturbation of SOM are proportional, with the same negative proportionality constant, to their responses to perturbation of VIP (Eqs. 2,S13). This mathematical prediction, the *Hidden Response Symmetries* held in most low-D models that fit the data (Fig. 3), we found it to further hold at the single cell level in high-D models (Fig. 6c) and would expect it to be found in *in vivo* optogenetic experiments. This prediction, showing with great generality that the independent manipulation of the activity of these interneurons elicits opposite effects on the network state, is in close accord with observations of SOM-VIP competition as has been observed in responses to multiple stimuli, or to behavioral or artificial manipulations[17,41,44].

Inhibition stabilization is well-defined in multiple interneuronal circuits[37,38], but how each interneuron contributes to circuit stabilization, and the link of stabilization with response to perturbations[32], has not been entirely understood (see also[33]). In this work, we offer a perspective that links responses of a unit to stability of the subcircuit without that unit: If the subcircuit is stable, then the unit will not respond paradoxically to a perturbation. Conversely, if the unit's response is paradoxical, the sub-circuit without it is unstable. Instability of a subcircuit lacking one unit does not guarantee the paradoxical response of that unit, just as the non-paradoxical response will not guarantee the stability of the sub-circuit. This is because the sub-circuit could lose stability by failing to satisfy other stability conditions than the determinant having the correct sign (see Eq. 1). In our family of models we find evidence in support of PV being the main circuit stabilizer (Figs. 3,6,S11), given that it shows paradoxical response (as in experiments,[5,32]) which guarantees that the circuit without it is unstable. This instability is consistent with experimental observations[15] and theoretical considerations[33]. The majority of models we analyzed did not show a SOM paradoxical response, consistent with the E-PV subcircuit being stable (Eq. S12). Nevertheless, we don't necessarily expect this insight to hold for all experimental configurations: in situations in which lateral recurrence through somatostatin interneurons plays a major role[6,14,15], it remains to be investigated how stabilization is performed throughout space.

To our knowledge this is the first time that a dynamical system model has accounted for the entire distribution of responses to stimuli of multiple cell types. Our high-D models belong to a family of recurrent rate networks[26,27,45] for which mean-field equations link the mean and variance of the inputs to the mean and variance of the activity (Eqs. S56, S58). These models have an explicit expression for the distribution of rates[35] that depends only on the mean and variance of the distribution of inputs (Eq. S62). Our approach to fit the distributions does not depend on the choice of non-linearity, but on the distribution of activity having an explicit form. With suitable simplifications, analogous methods could be used to fit models of multi-cell-type spiking networks, and extend it to account for other prominent cell-type-specific biological features, such as cell-type-specific gap-junctions or dynamic synapses as found in the mouse cortex[23]. Furthermore, current advanced machine learning methods[46,47] built to find the joint distribution of parameters that account for certain features of the data, could be combined with mean-field theoretical approaches to find specific manifolds in parameter space generating data-compatible models. In fact, the parameters that we found to be suitable for the high-D model revealed a comparatively small variance of the distribution of weights compared to its mean (see Fig. S2), unlike what is found experimentally (*e.g.*, Ref.[48]). It is conceivable that models with larger variances would be revealed by other methods, or by further incorporating biological detail in the model.

In this work we built on recent results in random matrix theory[36] to compute the mean and variance of the distribution of responses to partial and heterogeneous perturbations under a suitable approximation. In this case, the mean of the response matrix is the response of the linearized non-heterogeneous system (Eq. S138), offering a recipe to link the responses of high- and low-D systems. It is interesting to observe that, because the variance of the distribution of responses depends nonlinearly in a very intricate way on products of cell-type gain and the variance of the weights (Eq. S128), the spreading of the distribution of activity with contrast and the spreading of the responses to optogenetic perturbations are not predictive of one another (compare Fig. 5 and Fig. S9). This fact is not an artifact of the HFP approximation, given that via simulations we verify that it holds for the fully nonlinear system.

We find that partial perturbations of the network give rise to a bimodal distribution of responses described by a mixture of Gaussians (Fig. 8, Eq. S111). Increasing the number of stimulated cells leads to a non-monotonic dependence of the fraction of cells that respond paradoxically on the fraction of stimulated cells, an effect that we name the *fractional paradoxical effect*.

This does not depends on the HFP approximation, given that fully nonlinear disordered models show similar behavior (Fig. 8b), and does not depend on the specifics of the model, given that it is evident in the many thousand high-D models that fit the data (Fig. S10), and therefore is robust to strong parameter degeneracy. By mapping the mean response of the perturbed population to the response of a PV sub-population of an associated low-D system, we find an argument to link the *fractional paradoxical* effect to the stability of the non-perturbed sub-network, which is verified numerically in the full system.

One weakness of our current approach is that heterogeneity in the opsin expression and the heterogeneity in responses that contributes to heterogeneity of the linearized weights are not distinguished (Eq. S136), precluding an understanding of their interaction. In our system, because the variance of the response to perturbations is linear in the variance of the heterogeneity (Eq. S128), increased heterogeneity in the expression will tend to smear out the distribution of responses in this system. Future experiments that are able to control the number of perturbed cells, possibly through holographic manipulations of local circuits, will be able to determine the validity of this prediction.

Finally, all the work presented here is concerned with steady-state responses and perturbations. It is conceivable that temporal driving of the models developed here will have particular spectral signatures and stimulus dependencies on stimulus presentation[49], and that similar methods to the ones utilized here can be used to explore temporal fluctuations around the fixed points. Whether contrast modulations of the population's spectral signatures found in the monkey[50] and the mouse[51] visual cortex can be reproduced here, and the predictions for cell-type-specific temporal and spectral responses, are left for future work.

Acknowledgements

A.P. would like to acknowledge the support of the Swartz Foundation Fellowship for Theory in Neuroscience 2019-4.K.D.M., H.A., A.P., and D.P.M. would like to acknowledge funding from NIH 5U19NS107613. K.D.M. and A.P. also acknowledge funding from NIH U01-NS108683 and R01-EY029999, from NSF NeuroNex 1707398, and from Gatsby Foundation GAT3708. D.P.M. was supported by an NSF Graduate Research Fellowship. H.A. is a New York Stem Cell Foundation-Robertson Investigator. H.A. and D.P.M. acknowledge the funding of NEI grant R01EY023756-01. All authors would like to thank B. Doiron, G. Handy, A.L. Kumar, and L. Mazzucato for useful feedback on this manuscript.

Author Contributions

A.P. F.F. and K.M. conceived the study. A.P. and N.K. designed the low-dimensional fit approach. A.P. designed the high-dimensional fit approach, performed the numerical simulations and the analytical calculations with the advice of F.F. and K.M.. D.M. recorded and analyzed all the experimental data under the supervision of H.A.. A.P., F.F and K.M. wrote the paper. All authors discussed the results and contributed to the final stage of this manuscript.

Supplementary Figures

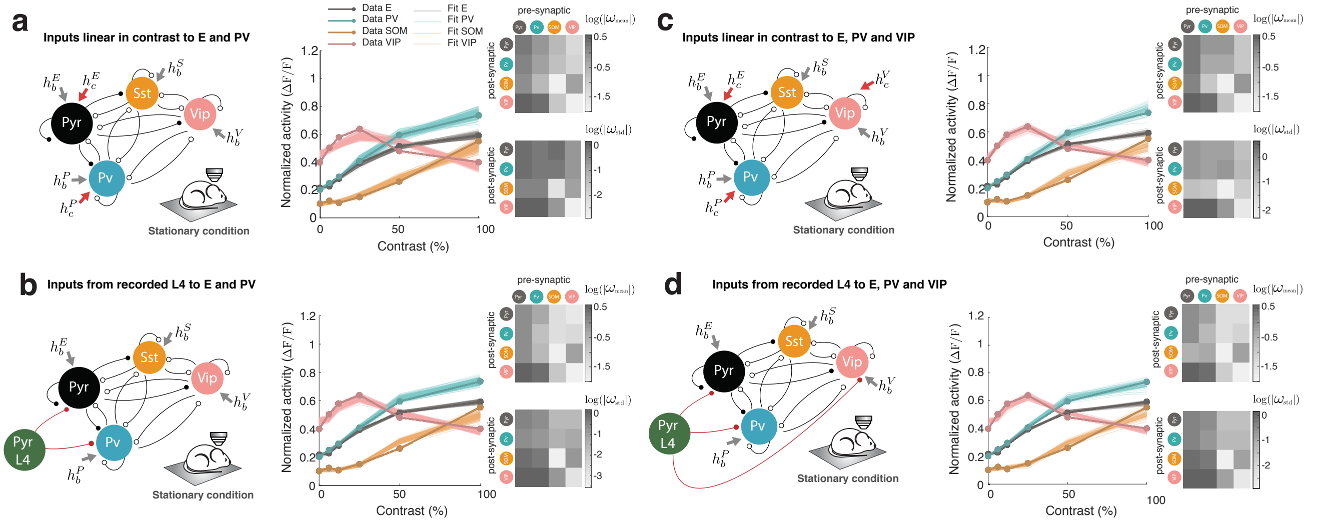

Figure S1 **Lack of SOM and VIP recurrence are independent of model choices.** **a)** Same as Figure 2, reproduced here for comparison. The inputs are linear in the contrast and target only the E and PV populations with constant factor h_c^{α} which are fitted together with the weights. **b)** Same as a) when the inputs are taken as the recorded L4 pyramidal cell activity. In this case the weights from the L4 population to E and PV populations are also fitted together with the rest of the weights and baseline inputs. The model fits the data less accurately, specially at the highest activity points of VIP at 25% contrast and of SOM at 100% contrast. **c)** The inputs are linear in the contrast with constant factor h_c^{α} which are fitted together with the weights, this time targeting also VIP interneurons. We see that in this case the solution that the NNLS finds, VIP projects to E and PV although the recurrence within VIP and SOM stays absent.**d)** Last case in which the inputs are taken as the recorded L4 pyramidal cell activity and target E, PV and VIP populations. Identically as before the VIP-VIP and SOM-SOM connections stay absent.

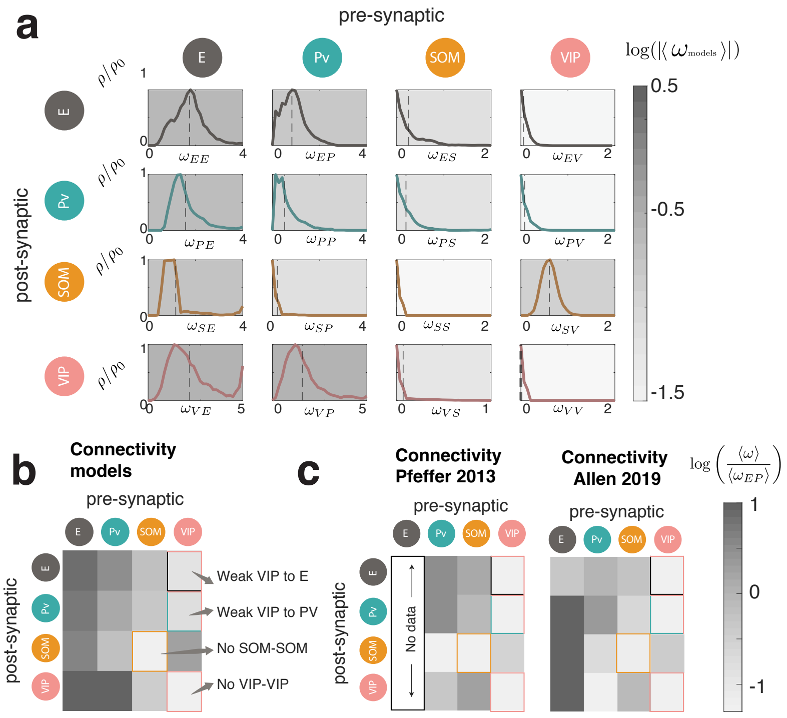

Figure S2 Direct comparison with experimentally reported synaptic connectivity. a) Distribution of connectivity weights for the top 0.1 percent of models that fit the data. **b)** Mean of the distributions of weights shown in a) normalized by the synaptic weight from PV to E, for comparison with the available experimental data. **c)** Left: Synaptic weight connectivity as obained in[22]. Right: Publicly available connectivity data from the Allen institute (https://portal.brain-map.org). The shown matrix is mean synaptic weight of the distribution of connections times the connection probability times the fraction of neurons belonging to the pre-synaptic cell-type, normalized by the synaptic weight from PV to E, as done originally in[22]. These two matrices are shown here for comparison. There is currently no agreement on the strength of the connection from PV to VIP.

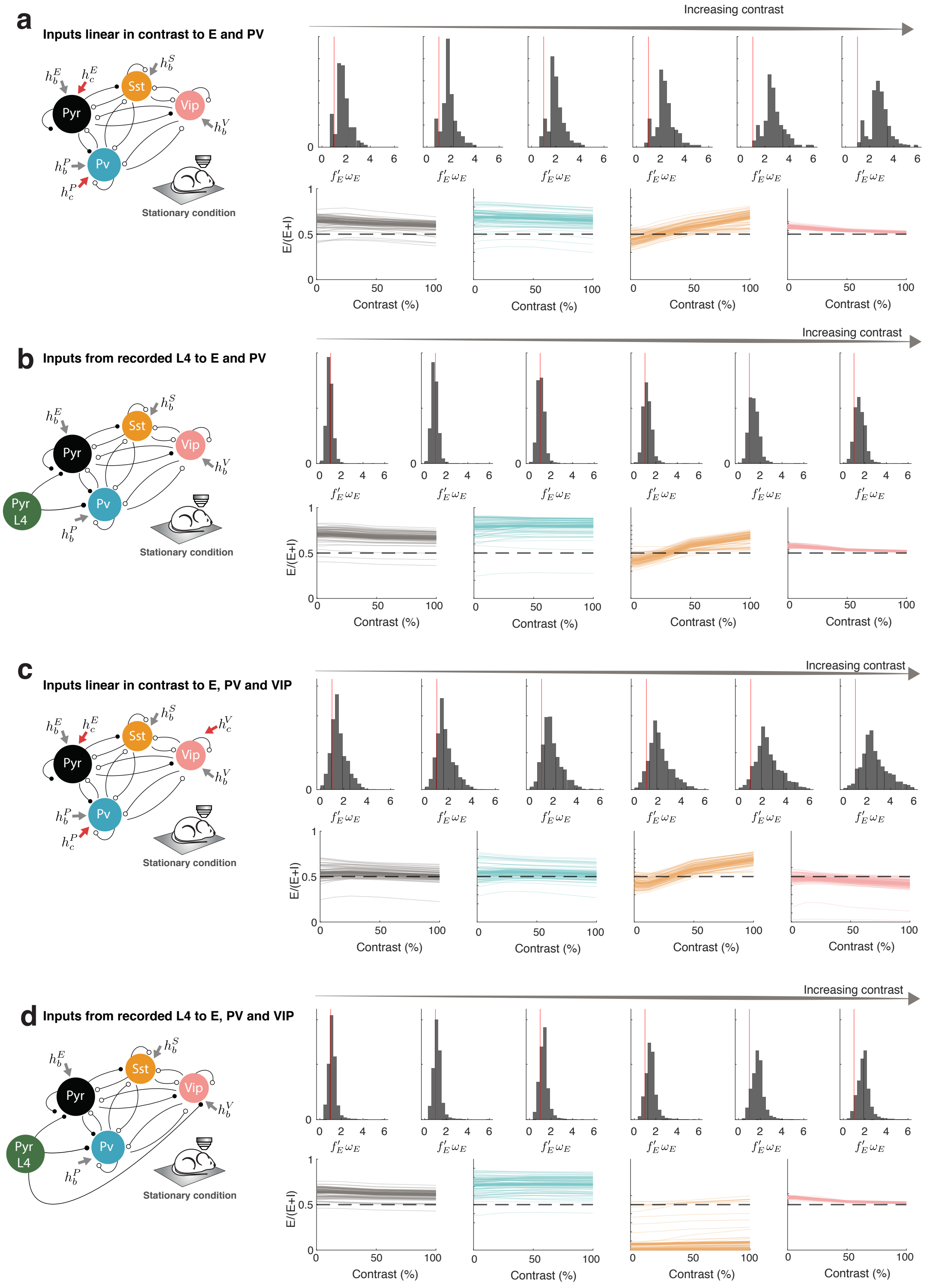

Figure S3 Inhibition stabilization is an input-independent property of models that fit the data. a) Case in which contrast inputs are linear in the contrast and project only E and PV. Top: ISN coefficient for different values of the contrast. When this coefficient is greater than 1 the network is inhibition stabilized. Although for low contrast, in which the excitatory activity is low, there are models compatible with a non ISN network, there distribution of weights shifts to values higher than one for larger contrast. Bottom: E-I ratio computed as the excitatory input current (i.e. $\omega_{\alpha E} \cdot r_E$) over the total recurrent input, for the E, PV, SOM and VIP population. **b)** Idem a) for the case in which the inputs are taken to directly by the mean pyramidal cell activity recorded in layer 4. **c)** Idem a) when feed-forward inputs target E,PV and VIP. **d)** Idem c) when feed-forward inputs target E,PV and VIP.

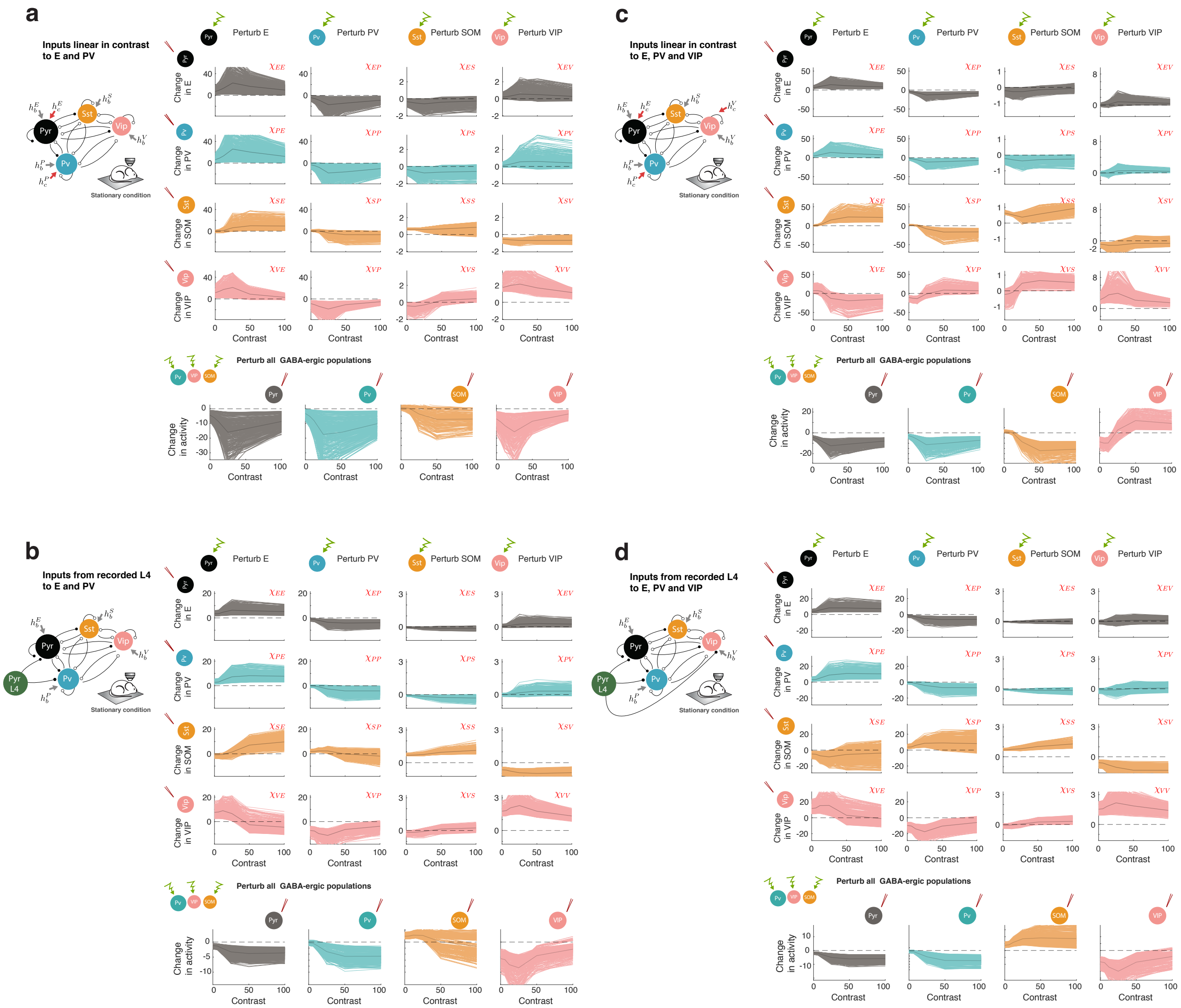

Figure S4 Paradoxical response to a cell-type perturbation and and not to a inhibitory-subcircuit perturbation is robust against model input choices. a) Top: Linear response matrix as a function of contrast for the network configuration receiving contrast inputs linear in the contrast targeting only the E and PV populations, same as Figure 3, reproduced here for comparison. Bottom: Response of E, PV, SOM and VIP populations to perturbation of all interneurons. All interneurons at sufficiently high contrast response with a decrease in their activity with this manipulation. **b)** Top: same as a) but when the contrast inputs target only E and PV and are taken to be proportional to recorded L4 E activity. Differently from panel a), VIP and SOM always have an inhibitory effect on each other, and PV has an always inhibitory effect on VIP. Bottom: idem as before, response of E, PV, SOM and VIP populations to perturbation of all interneurons. **c)** Top: same as a) but when the contrast inputs target E, PV and VIP instead of only E and PV, and are taken to be linear in the contrast. Bottom: The response of a perturbation to all interneurons is remarkably different from that observed in a), and VIP being excited by the stimulation of all interneurons **d)** Top: same as a) but when the contrast inputs target E, PV and VIP instead of only E and PV, and are taken to be proportional to recorded L4 E activity. Generally we find that there are similarities in the responses of the models that fit the data under these different input configurations. The response of SOM and VIP to a PV perturbation seems to be the only inconsistent response across these examples, and the response that ultimately defines whether that cell-type will respond paradoxically to a full sub-circuit perturbation.

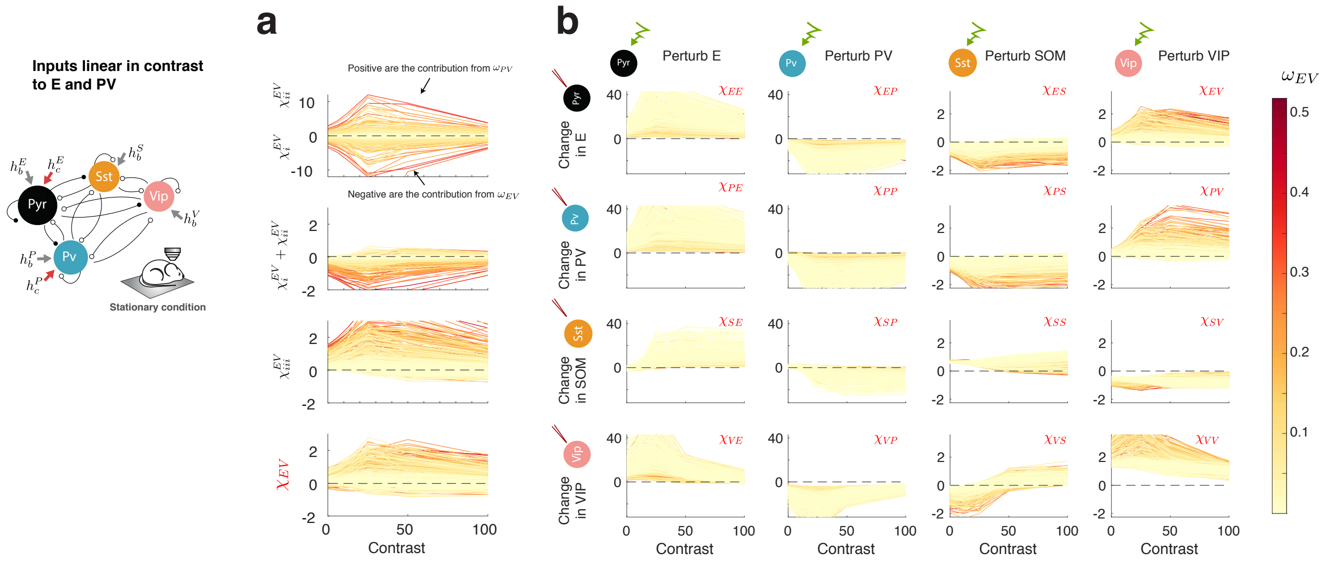

Figure S5 **VIP->E connection does not disrupt disinhibition in these models.** To investigate whether the fraction of models which VIP has an effectively inhibitory effect to E is due to a direct inhibitory connection from VIP to E as suspected in frontal cortex[52], we analyzed the term χ_{EV} in more detail. **a)** Decomposition of the term χ_{EV} into the three terms defined in Eq. (S28). The top panel is the component of χ_{EV} that is due to the existence of a ω_{EV} connection (negative) or to the existence of a ω_{PV} connection (positive). The color code indicates the value of ω_{EV}. The middle top panel is the sum of those two terms. Imperfect cancellation results in negative values for those models that have large ω_{EV} (dark read), while are close to zero for those with small ω_{EV} (yellow). Middle bottom: component of χ_{EV} proportional to ω_{SV}. We notice that for the dark read models, associated with large ω_{EV} this term is large and positive, and often results in a positive cancellation as can be seen in the lower panel. In summary, models with large values of ω_{EV} do not necessarily result on VIP having an inhibitory effect on E, but rather depends on whether the the net effect of the term $\chi_i^{EV} ii$, proportional to ω_{SV}, which quantifies the contribution of the path SOM->E is positive or negative. **b)** Linear response matrix as in Fig. 3 color-coded by the absolute value of ω_{EV}. Given that there are many more models with small ω_{EV}, only a few with large ω_{EV} can be appreciated. We notice that the models with vanishing ω_{EV}, span all the range of responses.

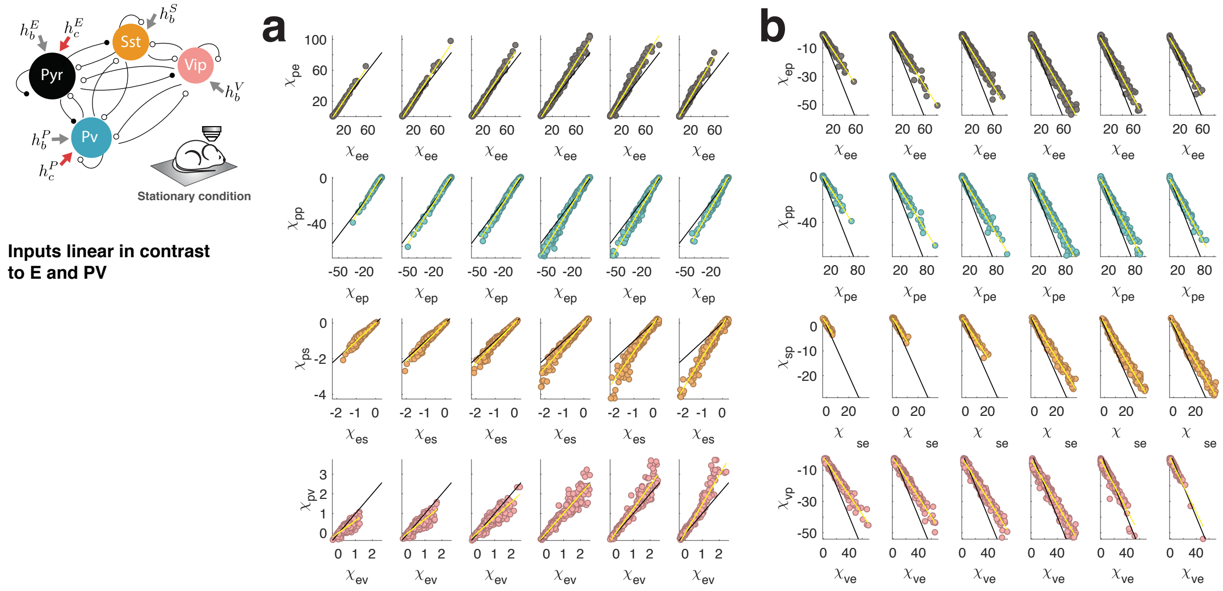

Figure S6 **E-PV symmetries in the response matrix.** **a)** Relationship between the first two rows to the linear response matrix of Fig. 3, highlighting the similarity between the response to perturbation between E and PV. **b)** Relationship between the first two columns to the linear response matrix of Fig. 3. The response of all cell-types to a perturbation to E and to PV have opposite signs and a linear relationship. In the simplified case of figure S7, it is possible to compute the contrast dependent offset and slope of these relations.

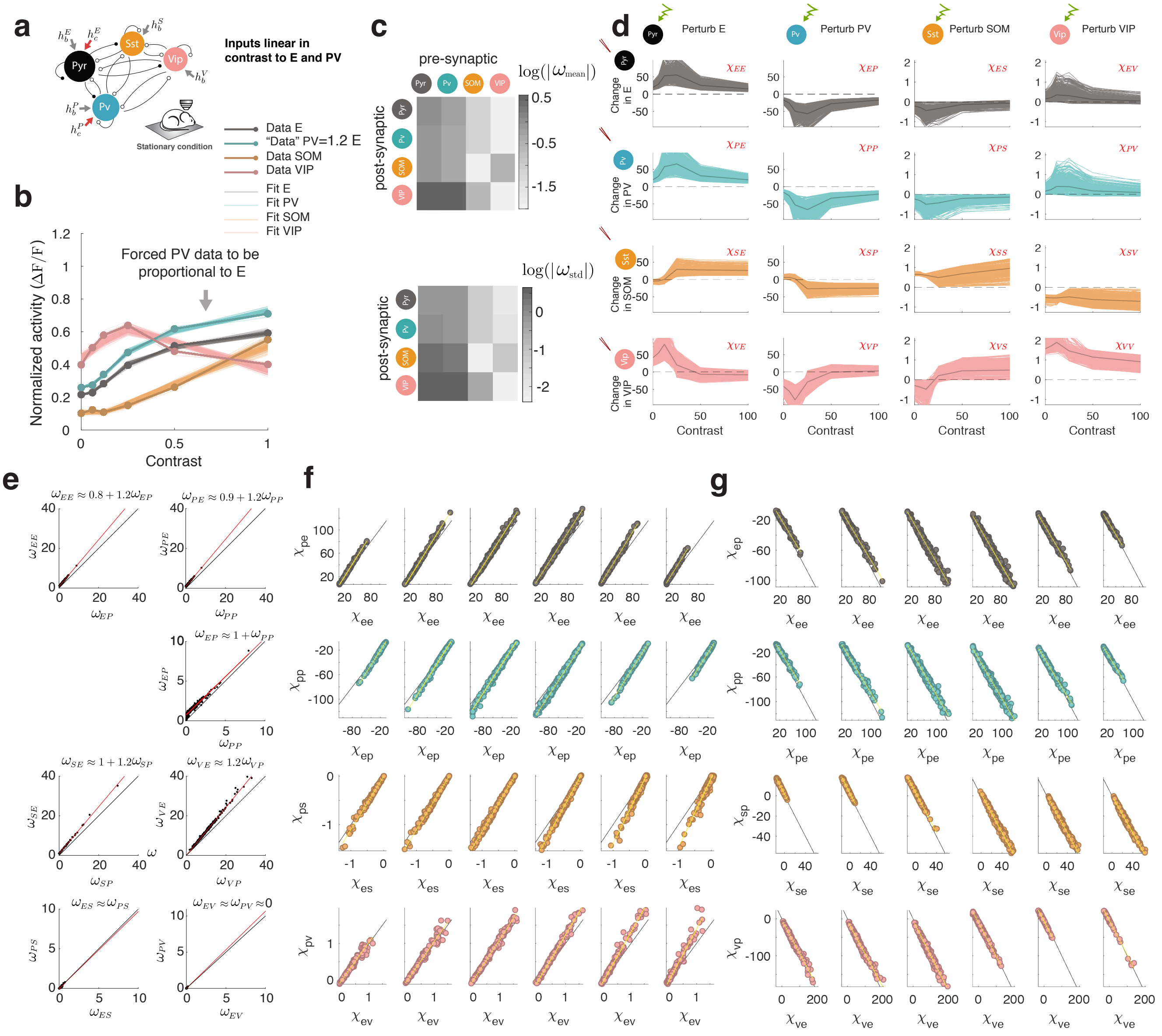

Figure S7 E-PV symmetries in the response matrix are a consequence of parameter correlation in simplified models in which PV is proportional to E **a)** Simplified scenario for the study of the E-PV related symmetries in the linear response matrix. Instead of fitting the measured mean activity, we fit models to data in which PV activity is directly proportional to E (PV= 1.2 E) and keep all other activity untouched. **b)** Model fits for the artifical dataset. **c)** We find that this family of models has even more accentuated structure than that presented in Figs. 3,S1. Fitting models in which projections from VIP to E,PV,VIP were clamped to zero made virtually no difference (not shown).**d)** The linear response matrix in this case is almost identical to that obtained for the original dataset (compare Fig. 3).**e)** In this simplified case, the connections to E and PV and from E and PV have linear dependencies. **f)** When these dependencies are incorporated in the response matrix, it is possible to calculate how each of these quantities are related (see section Matrix symmetries in simplified circuits and Eq. (S33)).

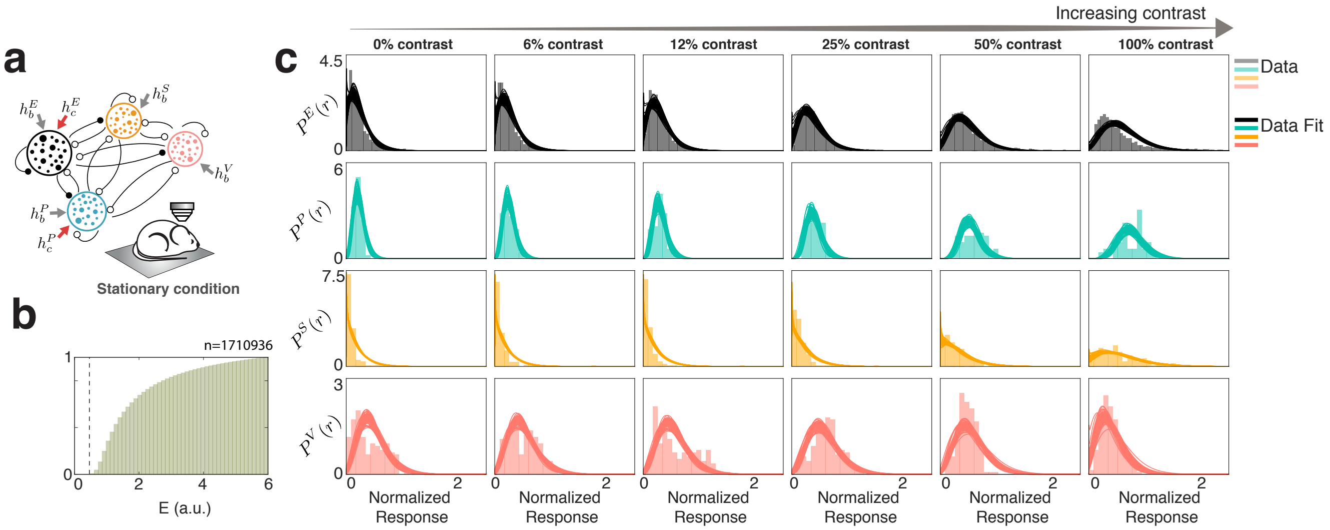

Figure S8 Family of high-D models. **a)** High-dimensional system. The proportion of neurons in each population is as in Fig. 5. **b)** Distribution of KL divergences as defined in Eq. (S64).From the almost 2 million parameter configurations sampled from the prior, we choose a threshold of 0.46 to keep around 100 models.**c)** First hundred models with smallest error. The distributions of VIP are bimodal at low contrast a and have a non monotonic spread respect to contrast that the mean field models fail to capture. In the future it should be investigated what is the reason for this bi-modality and whether its sensical to group those cells under a single category.

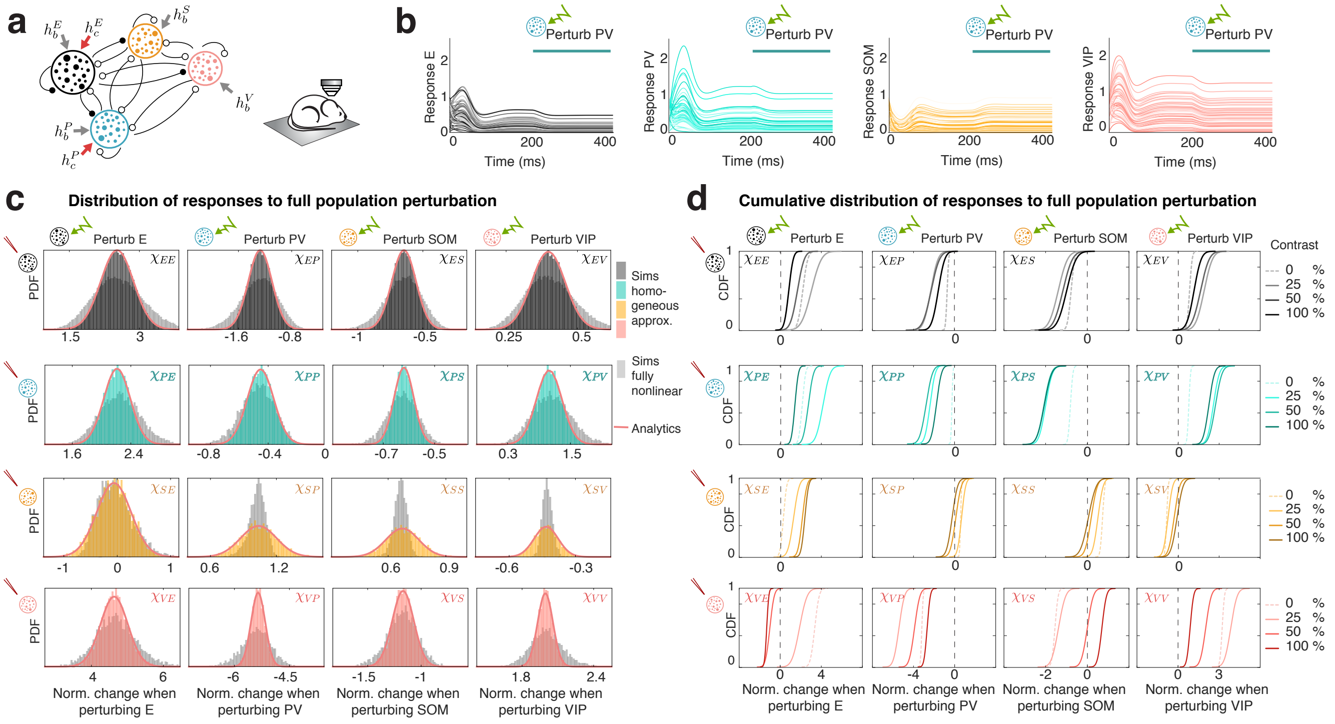

Figure S9 High-D linear response analysis. a) The high-dimensional model used in this figure correspond to the top 5% of models that fit the data as described above. **b)** Numerical example of the response of E, PV, SOM and VIP to a small perturbation. At time zero the network evolves towards its fixed point, and it receives a homogeneous perturbation in the PV population at time 200 ms, reaching a new steady state. The normalized difference between these steady states, divided by the size of the perturbation, gives the normalized change in activity (the linear response distributions). **c)** Distributions of linear responses for a single example network in the absence of visual stimulation. Each column is the response of the network, divided in populations, to a homogeneous stimulation of a population indicated at the top. Gray histograms are the result of the simulation of a fully high-dimensional nonlinear system, while colored histograms are simulations of the network with the HFP approximation (in which the system is linearized around the fixed point of the network without disorder). Full orange lines are the analytics result, only possible with that same approximation. **d)** Cumulative distributions of the responses in (d) for different values of the contrast. We observe that the models keep the overall behavior that the low-dimensional system has. Other model examples in Fig. S5

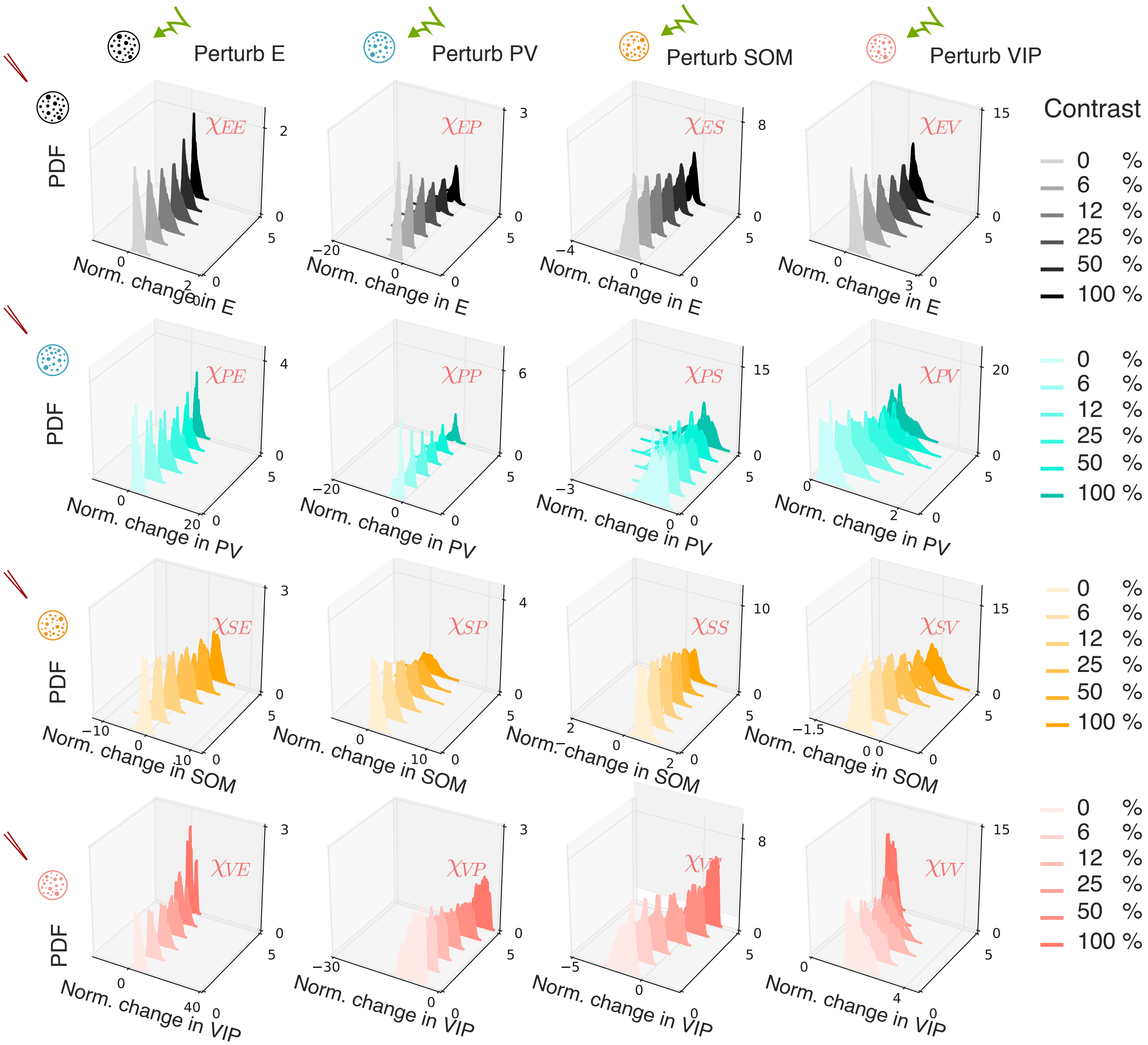

Figure S10 **Contrast modulation of response distributions to full population perturbations in high-D models.** Distribution of linear responses to homogeneous full-population perturbations for different values of the contrast for all those models chosen in Fig. S8. As described in the methods, the seeds for the priors, obtained from the low-D models, were chosen among those in which PV is paradoxical at low contrast. We see that that condition guarantees in this family of models a disinhibitory effect in the activity of E when perturbing VIP.

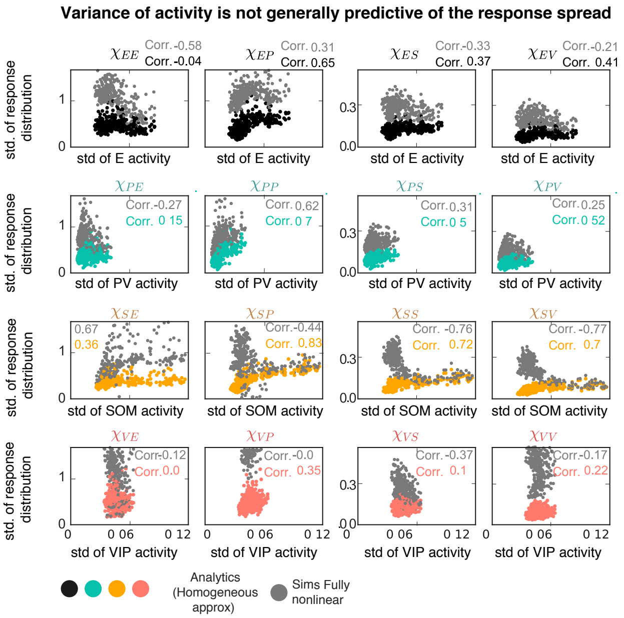

Figure S11 Uncorrelated variability of the distribution of optogenetic responses and the distributions of activity. Standard deviation of the responses of each cell type to perturbations to each cell type as a function of the standard deviation of the activity. We see that neither under the analytical approximation (black E, turquoise PV, orange SOM and pink VIP), or for the fully nonlinear system, there is a clear correlation between the spread of these distributions. The spread of the distribution of activity is therefore not a good predictor of the distribution of responses

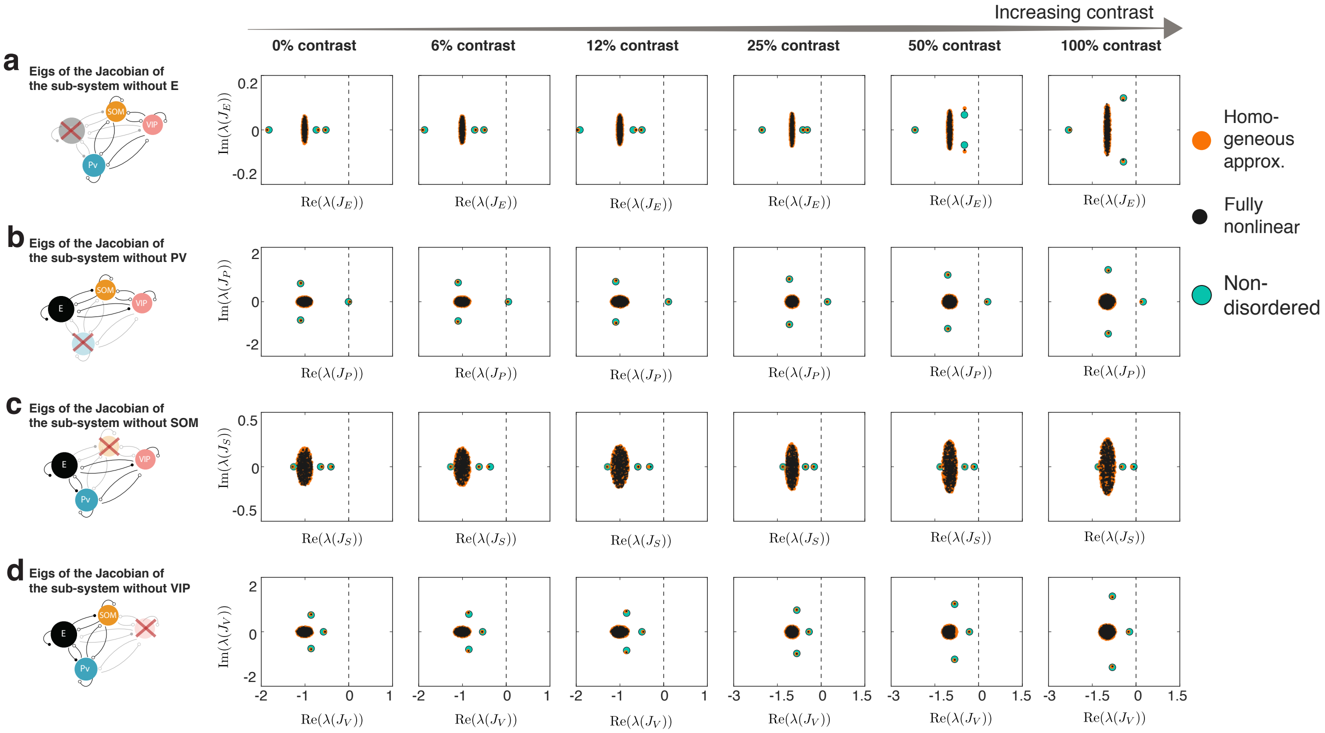

Figure S12 Linking stability and response to perturbations in high-D models. **a)** Eigenvalues of the Jacobian of the sub-circuit without E (J_E) as a function of contrast. The response of pyramidal cells is never paradoxical and the sub-circuit without it its always stable, for the system using the HFP approximation (orange), the non-disordered system (the low-dimensional system), and the fully-non linear system without approximations, for the same simulation shown in Figs. 6 and S11. **b)** Idem a) for the sub-circuit without PV (J_P). The sub-system without PV is always unstable. **c)** Idem a) for the sub-circuit without SOM (J_S), although this sub-circuit approaches instability for large contrast in this particular example it remains non-paradoxical. **d)** Idem a) for the sub-circuit without VIP (J_P).

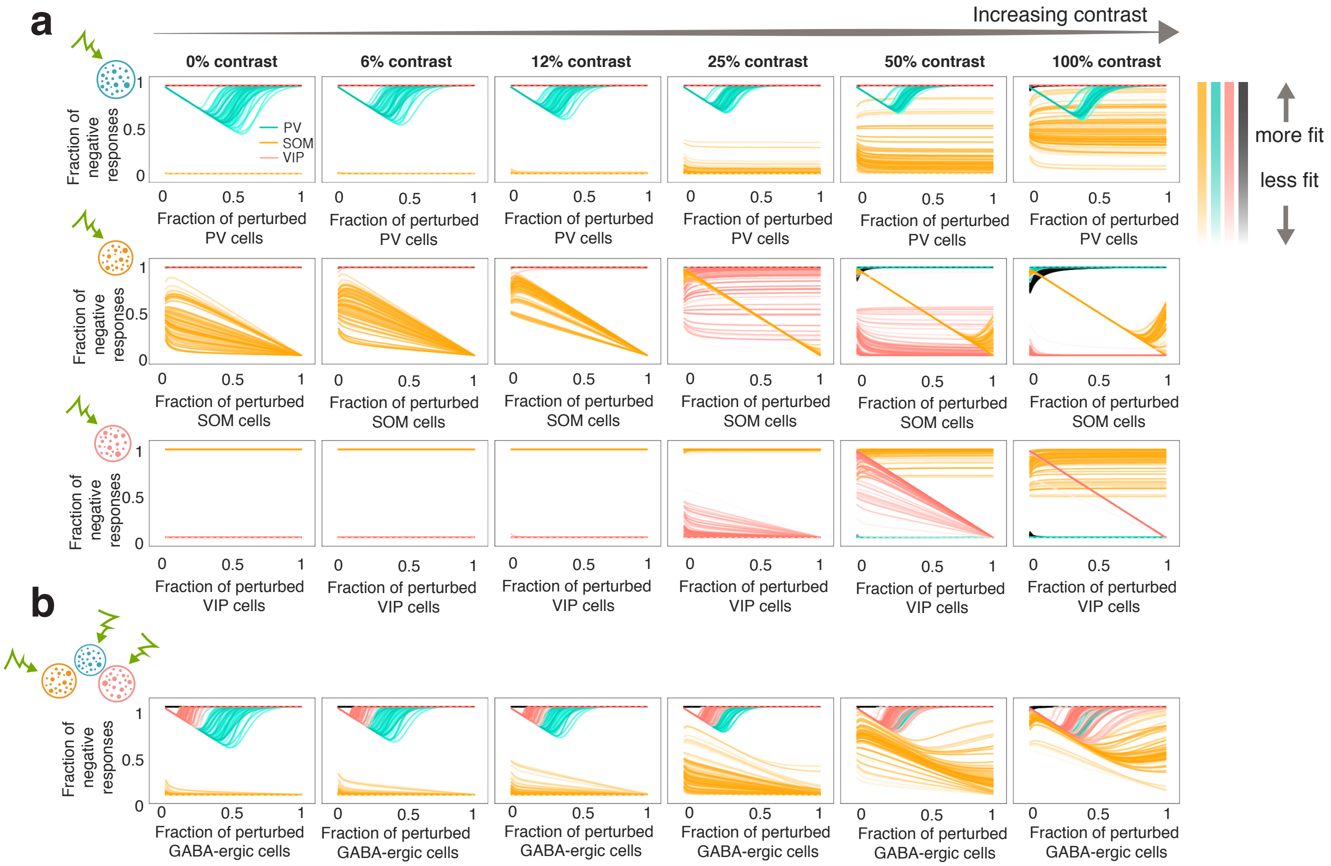

Figure S13 Fractional paradoxical effect in the family of high-D models. Dependence of the fraction of negative responses of the PV, SOM and VIP populations as a function of the fraction of stimulated PV (top), SOM (middle) and VIP (bottom) cells, as in Fig. 8. **b)** Same as a) but while stimulating simultaneously all GABAergic cells. We again emphasize the non-monotonic dependence on this fraction is a prominent feature, which is recovered in all analyzed models. In this case the shaded color indicates fitness. The different models have qualitatively similar behavior but some features seem to be quite sensitive to the parameters of the model. We generally find that the overall behavior described in Fig. 8 holds for this family of models. The fraction of PV cells that respond paradoxically is a nonlinear function of the perturbed cells.

Methods

1 Methods summary

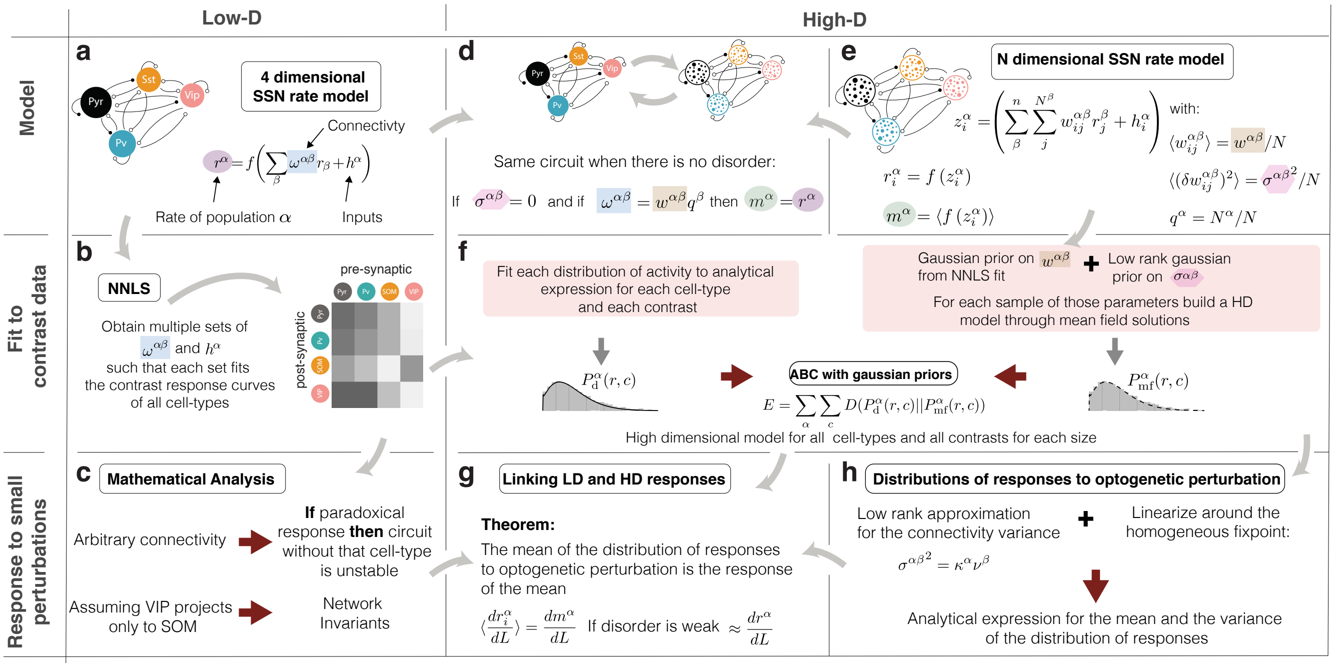
Low-D
High-D
Model
Fit to contrast data
Response to small perturbations
a
Pyr
Sst
Vip
Pv
4 dimensional SSN rate model
Connectivty
$r^\alpha = f\left(\sum_\beta \omega^{\alpha\beta} r_\beta + h^\alpha\right)$
Rate of population α
Inputs
b
NNLS
Obtain multiple sets of $\omega^{\alpha\beta}$ and h^α such that each set fits the contrast response curves of all cell-types
pre-synaptic
post-synaptic
Pyr
Pv
SOM
VIP
c
Mathematical Analysis
Arbitrary connectivity
If paradoxical response then circuit without that cell-type is unstable
Assuming VIP projects only to SOM
Network Invariants
d
Same circuit when there is no disorder:
If $\sigma^{\alpha\beta} = 0$ and if $\omega^{\alpha\beta} = w^{\alpha\beta} q^\beta$ then $m^\alpha = r^\alpha$
e
N dimensional SSN rate model
$z_i^\alpha = \left(\sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta} r_j^\beta + h_i^\alpha\right)$
with:
$\langle w_{ij}^{\alpha\beta} \rangle = w^{\alpha\beta}/N$
$r_i^\alpha = f\left(z_i^\alpha\right)$
$\langle (\delta w_{ij}^{\alpha\beta})^2 \rangle = {\sigma^{\alpha\beta}}^2/N$
$m^\alpha = \langle f\left(z_i^\alpha\right) \rangle$
$q^\alpha = N^\alpha/N$
f
Fit each distribution of activity to analytical expression for each cell-type and each contrast
$P_\mathrm{d}^\alpha(r,c)$
ABC with gaussian priors
$E = \sum_\alpha \sum_c D(P_\mathrm{d}^\alpha(r,c)||P_\mathrm{mf}^\alpha(r,c))$
High dimensional model for all cell-types and all contrasts for each size
Gaussian prior on $w^{\alpha\beta}$ from NNLS fit
Low rank gaussian prior on $\sigma^{\alpha\beta}$
For each sample of those parameters build a HD model through mean field solutions
$P_\mathrm{mf}^\alpha(r,c)$
g
Linking LD and HD responses
Theorem:
The mean of the distribution of responses to optogenetic perturbation is the response of the mean
$\langle \frac{dr_i^\alpha}{dL} \rangle = \frac{dm^\alpha}{dL}$ If disorder is weak $\approx \frac{dr^\alpha}{dL}$
h
Distributions of responses to optogenetic perturbation
Low rank approximation for the connectivity variance
Linearize around the homogeneous fixpoint:
${\sigma^{\alpha\beta}}^2 = \kappa^\alpha \nu^\beta$
Analytical expression for the mean and the variance of the distribution of responses

Figure S14 Outline of this paper. a-b) Low-D circuit: Multi cell-type circuit describing the population activity of E, PV, SOM, and VIP cells when presented with stimuli of different contrasts. By using non-negative least squares (NNLS) we find the parameters to describe the circuit's contrast response. Results in Fig. 2. **c)** Assuming that VIP only projects to SOM and SOM does not project to itself, we find relations between stability and responses to optogenetic perturbations and find hidden structure in the response matrix. These findings are applied to the models that fit the data. Results in Fig. 3 **d-e)** High-dimensional model: When all the cells of one population connect to the cells of the other population with the same strength (no disorder), the high-dimensional circuit describes the same dynamics as the circuit described in (a) given that the parameters are chosen appropriately. Inclusion of disorder changes that mean activity. **f)** We use approximate bayesian inference (ABC) to fit the high-dimensional system. Firstly, given that the models we use have an analytical form for the distribution of activity, we use it to separately fit the distribution of activity of each cell-type and each stimulus condition. Secondly, we build MF models with parameters sampled from a distributions with priors obtained from the NNLS analysis. By minimizing the Kullback-Leibler divergence[53] between these two sets of distributions (the one obtained from the data and the one obtained from the MF family), we find the models that best approximate the distribution of all cell-types at all contrasts with a single parameter set. **g-h)** Analytical expressions for the distribution of responses to optogenetic perturbations are available for linear systems. Through an approximation, we linearize the high-dimensional system around the HFP and use existing mathematical expressions to compute the entire distribution of responses to an arbitrary pattern of optogenetic stimulation.

We develop a three-stage program for the prediction of responses to weak optogenetic perturbations of circuits with multiple inhibitory types (Fig. S14). In a first stage, we use non-negative least squares (NNLS)[20,54] (see Eq. S5)to fit a low-dimensional (low-D) dynamical system to the mean responses observed experimentally in all four cell types (excitatory(E), PV, SOM, VIP) in mouse layer 2/3 to stimulation by a small (5 degree diameter) visual stimulus of varying contrast. These fits make predictions about the mean connection strengths between neurons of any two given cell types, (Fig. S14**b**), which allows a mathematical understanding of the response to perturbations to different cell-types (Fig. S14**c**). In a second stage, we build a family of high-D models, with different numbers of cells per population. For that, we work with a high-D rate model[26] (Fig. S14**d-e**, see Eq.S58). In this model, the distribution of activity has a tractable analytical form[35] (see Eq. S62) that depends on the mean and variance of the input currents to each population. We can obtain that mean and variance for each by fitting that distribution to the data via maximum likelihood, but that is not sufficient to build a model: we need a way to find model parameters (*e.g.*, means and variances of connection strengths) that will generate the mean and variance of the input currents and the firing rates self-consistently for all stimulus conditions and all cell-types. Working from the other direction, given a high-D model and its parameters, we can use MF theory[26] to self-consistently find the activity distributions that result for a given stimulus. Finally, in order to find the parameters of high-D models that fit the experimental data, we use a the distance between the fit to the data distribution and the distribution obtained by the MF solutions of a given model (Fig. S14**f**, and see Eq. S64). By choosing a suitable threshold on this distance (0.45)[55], we find high-D models whose distribution of activity and dependence on stimulus contrast reproduce those observed experimentally. In the final stage, we use theoretical results on random matrices[36] which allow us to analytically compute the distribution of neuronal responses to patterned optogenetic perturbations under a suitable approximation (Fig. S14**h**) and determine its relation to the predictions in the low-D circuit (Fig. S14**g**).

2 Data Collection and Analysis

All the data presented here was collected by Daniel Mossing and forms the subject of another publication. Details on the data collection will be provided elsewhere.

3 Low-dimensional circuit models

We consider a network of 4 units, each describing the activity r^α of a particular cell-type population α, with $\alpha = \{\text{E, PV, SOM, VIP}\}$ in layer 2/3 of the visual cortex of the mouse. The network integrates input currents z^α in the following way

$$\tau^\alpha \dot{r^\alpha} = -r^\alpha + f(z^\alpha) \qquad z^\alpha = \left(\sum_\beta^n \omega^{\alpha\beta} r^\beta + h^\alpha(c) \right) \tag{S3}$$

where τ^α is the relaxation time scale, $\omega^{\alpha\beta}$ is the connectivity matrix, and $f(z) = []_+^\xi$ is the activation function with $\xi = 2$ unless otherwise specified. The inputs $h^\alpha(c)$ are composed of a baseline input h_b, a sensory-related input $h_s(c)$. This last input can be either chosen to be proportional to the contrast c, for which $h_s(c) = h_c c$, with h_c a contrast independent variable to be fitted or it can be chosen as proportional to measured layer 4 pyramidal-cell activity data: $h_s(c) = h_{\text{L4}} r_{\text{L4}}(c)$, with h_{L4} an activity independent variable to be fitted.

$$h^\alpha(c) = h_b^\alpha + h_c^\alpha c \qquad \text{or} \qquad h^\alpha(c) = h_b^\alpha + h_{\text{L4}}^\alpha r_{\text{L4}}(c) \tag{S4}$$

3.1 Data fitting

To simultaneously fit the rates of all four interneurons at all contrast values (six in total $c = \{0, 6, 12, 25, 50, 100\}$), we consider the steady-state equations corresponding to (S3). Since the recorded firing rates are positive and non-vanishing, the inverse is well defined $f^{-1}(z) = \sqrt{r}$ and the nonlinear steady-state equation corresponding to (S3) becomes a linear equation with respect to the connectivity parameters:

$$\sqrt{r^\alpha} = \sum_\beta^n \omega^{\alpha\beta} r^\beta + h^\alpha(c) \tag{S5}$$

Eq. (S5) represents a system of linear equations $Ax = y$, where x is an unknown vector containing the flattened connectivity matrix entries $\omega^{\alpha\beta}$ and the input constants h_b^α, h_c^α, and h_{L4}^α. The entries of the matrix A and the vector y are the functions of the recorded firing rates at six contrast values. The matrix A has 24 rows: for each of the six contrast values a set of four rows corresponds to the steady-state equations in (S5). The number of columns of the matrix A is equal to the number of unknown connectivity and input constants. In the most general case, when each four populations receive background and sensory related input, there are 24 unknowns and the matrix A has 24 columns. This case in which the number of equations (rows of A) and the number of parameters (all chosen weights and inputs) are equal, the system $Ax = y$ can be solved exactly. To be concrete, taking as an example the case presented in the main text in which sensory inputs are linear in c and target only E and PV cells, we will have:

To the solve the system in this case, values of parameters that approximately solve the Eq. (S5) be found by computing the non-negative least squares (NNLS)[56] solution. In Figure S1, we explored four configurations, modeling the inputs either via a linear function of the contrast (fiting h^{α}_{c}) or a linear function of layer 4 data (fitting h^{α}_{L4}) for either $\alpha = \{\mathrm{E,PV}\}$ or $\alpha = \{\mathrm{E,PV,SOM}\}$. The NNLS solution of Eq. (S5) constructed from mean firing rates, gives *one* set of connectivity and input parameters x. To obtain distributions of connectivity and input parameters instead, we created surrogate contrast responses sets by sampling from a multivariate Gaussian distribution with mean $r^{\alpha}_{c_i}$ and standard error of the mean $s^{\alpha}_{c_i}$. For each input configuration, we sampled 2500000 seeds to create these surrogate contrast response curves. For each sample contrast response k, NNLS gave one connectivity and input parameter set. Using each parameter set and the steady-state equations in (S5) we computed the fit $\hat{r}^{\alpha}(k)$ of the kth sample contrast response. Keeping the stable solutions (negative eigenvalues, all time constants were chosen to be equal to 1), the likelihood of that parameter set k

$$\mathscr{L}_k = \prod_{c_i,\alpha} \frac{1}{\sqrt{2\pi} s^{\alpha}_{c_i}} \exp\left\{ -\frac{(\hat{r}^{\alpha}_{c_i}(k) - r^{\alpha}_{c_i})^2}{2 s^{\alpha\,2}_{c_i}} \right\} \tag{S6}$$

defined a hierarchy for the contrast response samples. For Figs. 2,3,S1,S3,S4,S5 and S6, the top 0.1% parameter sets were used to plot fits of the firing rates. In this way we obtained approximately 2350 parameter sets for each of four input configurations.

3.2 Linear response and paradoxical effects

The linear response matrix is defined as the steady state change in rate of a population α given by a change in the input current h to population β

$$\chi_{\alpha\beta} = \frac{dr_{\alpha}}{dh_{\beta}} = (\mathbf{f}'^{-1} - \omega)^{-1}_{\alpha\beta} \qquad \mathbf{f}' = \delta_{\alpha\beta} f'_{\alpha} \tag{S7}$$

Where f'_{α} is the gain of population α at the considered steady state, $\mathbf{f}'$ is the $n = 4$ diagonal matrix with elements f'_{α}, $\delta_{\alpha\beta}$ is a Kroenecker delta which is 1 only if $\alpha = \beta$. Defining the diagonal matrix of time constants $T_{\alpha}\beta = \delta_{\alpha\beta}\tau_{\alpha}$, Eq. (S7) can be written as a function of the the Jacobian $J = T^{-1}(-I + \mathbf{f}'\omega)$

$$\chi \mathbf{f}'^{-1} T = -J^{-1} \qquad \rightarrow \qquad \frac{\tau_{\beta}}{f'_{\beta}} \chi_{\alpha\beta} = \frac{-1}{\det J} (-1)^{\alpha+\beta} M_{\alpha\beta} \tag{S8}$$

where $M_{\alpha\beta}$ is the corresponding minor of the Jacobian. In particular, the diagonal entries of χ are

$$\frac{\tau_{\alpha}}{f'_{\alpha}} \chi_{\alpha\alpha} = \frac{-1}{\det J} M_{\alpha\alpha} \tag{S9}$$

Given that $M_{\alpha\alpha}$ corresponds to the determinant of the Jacobian of the sub-circuit without the cell-type α, which we call J_{α}, we find that:

$$\frac{\tau_\alpha}{f'_\alpha}\chi_{\alpha\alpha} = \frac{-1}{\det J}\det J_\alpha \tag{S10}$$

For a system with n populations, stability of the full system requires that $\text{sign}(\det J) = (-1)^n$. Stability of the sub-circuit without α requires that $\text{sign}(\det J_\alpha) = (-1)^{n-1}$. Given that the gain f'_α is always positive, if both the entire circuit and the subcircuit are stable, then $\chi_{\alpha\alpha} > 0$. Alternatively, if $\chi_{\alpha\alpha} < 0$, and the cell-type α has a paradoxical response, then the sub-circuit without it will be unstable. This does not depend on the dimension of the system.

3.3 EI netwotks

Evaluating Eq. (S10) in the EI case we obtain the result from[2]

$$\chi_{II} \propto 1 - f'_E\omega_{EE} \tag{S11}$$

which makes the parameter independent prediction that when recurrent excitation strong, the response of inhibition is paradoxical, $\chi_{II} < 0$.

3.4 E-PV stability and SOM paradoxical response when VIP projects only to SOM

In the particular case in which VIP projects only to SOM, the Eq. S10 reduces to

$$\frac{\tau_S}{f'_S}\chi_{SS} = \frac{1}{\det J}\det J_{EP} \tag{S12}$$

Given that in a 2D system, the conditions for stability are the trace to be positive and the determinant to be positive, and that the trace can be generally made positive by choosing a suitable large excitatory time constant, we say not only that measuring the paradoxical response of SOM translates in E-PV being unstable, but that observing a non-paradoxical response of SOM means that E-PV is stable given a suitable time constant.

3.5 Hidden response symmetries (VIP projects only to SOM)

The values $\chi_{\alpha\beta}$ for the particular case in which the connections from the VIP population to the rest is exactly zero can be found to satisfy the following relations, *Hidden response symmetries*.

$$\chi_{\text{EV}} = -f'_{\text{V}}\omega^{\text{SV}}\chi_{\text{ES}} \tag{S13}$$
$$\chi_{\text{PV}} = -f'_{\text{V}}\omega^{\text{SV}}\chi_{\text{PS}} \tag{S14}$$
$$\chi_{\text{SV}} = -f'_{\text{V}}\omega^{\text{SV}}\chi_{\text{SS}} \tag{S15}$$
$$\chi_{\text{VV}} = f'_{\text{V}} - f'_{\text{V}}\omega^{\text{SV}} * \chi_{\text{V}S} \tag{S16}$$

$$\chi_{\mathrm{VS}} = f'_{\mathrm{V}}(\omega^{\mathrm{VE}}\chi_{\mathrm{ES}} - \omega^{\mathrm{VP}}\chi_{\mathrm{PS}} - \omega^{\mathrm{VS}}\chi_{\mathrm{SS}}) \tag{S17}$$

$$\tag{S18}$$

This can be easily seen by explicitly writing the response matrix as

$$\chi_{\alpha\beta} = \frac{1}{D}k^{\alpha\beta} \qquad \frac{1}{D} = \frac{\det(T^{-1})}{\det(-J)} \tag{S19}$$

Where $\det(-J)$ is the determinant of the negative Jacobian of the full system, defined above Eq. (S8). Given that the eigenvalues of J have to be negative for linear stability, $\det(-J)$ is always positive, and the above relations can be instead written as a function of $k^{\alpha\beta}$ with

$$k^{\mathrm{ES}} = -f'_{\mathrm{E}}f'_{\mathrm{S}}(\omega^{\mathrm{ES}}(1+f'_{\mathrm{P}}\omega^{\mathrm{PP}}) - f'_{\mathrm{P}}\omega^{\mathrm{EP}}\omega^{\mathrm{PS}}) \tag{S20}$$

$$k^{\mathrm{EV}} = f'_{\mathrm{E}}f'_{\mathrm{S}}f'_{\mathrm{V}}\omega^{\mathrm{SV}}(\omega^{\mathrm{ES}}(1+f'_{\mathrm{P}}\omega^{\mathrm{PP}}) - f'_{\mathrm{P}}\omega^{\mathrm{EP}}\omega^{\mathrm{PS}}) \tag{S21}$$

$$k^{\mathrm{PS}} = -f'_{\mathrm{P}}f'_{\mathrm{S}}(\omega^{\mathrm{PS}}(1-f'_{\mathrm{E}}\omega^{\mathrm{EE}}) + f'_{\mathrm{E}}\omega^{\mathrm{ES}}\omega^{\mathrm{PE}}) \tag{S22}$$

$$k^{\mathrm{PV}} = f'_{\mathrm{P}}f'_{\mathrm{S}}f'_{\mathrm{V}}\omega^{\mathrm{SV}}(\omega^{\mathrm{PS}}(1-f'_{\mathrm{E}}\omega^{\mathrm{EE}}) + f'_{\mathrm{E}}\omega^{\mathrm{ES}}\omega^{\mathrm{PE}}) \tag{S23}$$

$$k^{\mathrm{SS}} = f'_{\mathrm{S}}((1-f'_{\mathrm{E}}\omega^{\mathrm{EE}})(1+f'_{\mathrm{P}}\omega^{\mathrm{PP}}) + f'_{\mathrm{E}}f'_{\mathrm{P}}\omega^{\mathrm{EP}}\omega^{\mathrm{PE}}) \tag{S24}$$

$$k^{\mathrm{SV}} = -\omega^{\mathrm{SV}}f'_{\mathrm{V}}f'_{\mathrm{S}}((1-f'_{\mathrm{E}}\omega^{\mathrm{EE}})(1+f'_{\mathrm{P}}\omega^{\mathrm{PP}}) + f'_{\mathrm{E}}f'_{\mathrm{P}}\omega^{\mathrm{EP}}\omega^{\mathrm{PE}}) \tag{S25}$$

$$k^{\mathrm{VS}} = -f'_{\mathrm{S}}f'_{\mathrm{V}}\left(f'_{\mathrm{E}}\left(\omega^{\mathrm{ES}}\omega^{\mathrm{VE}} - \frac{f'_{\mathrm{P}}|\omega^0|}{\omega^{\mathrm{SV}}}\right) + \omega^{\mathrm{VS}}(1-f'_{\mathrm{E}}\omega^{\mathrm{EE}}) + f'_{\mathrm{P}}(\omega^{\mathrm{PP}}\omega^{\mathrm{VS}} - \omega^{\mathrm{PS}}\omega^{\mathrm{VP}})\right) \tag{S26}$$

$$k^{\mathrm{VV}} = f'_{\mathrm{V}}(f'_{\mathrm{E}}f'_{\mathrm{P}}f'_{\mathrm{S}}\omega^{\mathrm{SE}}(\omega^{\mathrm{ES}}\omega^{\mathrm{PP}} - \omega^{\mathrm{EP}}\omega^{\mathrm{PS}}) + (1-f'_{\mathrm{E}}\omega^{\mathrm{EE}})(1+f'_{\mathrm{P}}\omega^{\mathrm{PP}}) + f'_{\mathrm{E}}f'_{\mathrm{P}}\omega^{\mathrm{EP}}\omega^{\mathrm{PE}} + f'_{\mathrm{E}}f'_{\mathrm{S}}\omega^{\mathrm{ES}}\omega^{\mathrm{SE}}) \tag{S27}$$

3.6 Disinhibition in the fully connected circuit

If we instead assume that $\omega^{\mathrm{VV}} = \omega^{\mathrm{SS}} = 0$ but all other connections exist, the response of pyramidal cells to a perturbation to VIP will have two extra terms:

$$\chi^{EV} = \chi_i^{EV} + \chi_{ii}^{EV} + \chi_{iii}^{EV} = \frac{1}{D}(k_i^{\mathrm{EV}} + k_{ii}^{\mathrm{EV}} + k_{iii}^{\mathrm{EV}}) \tag{S28}$$

$$k_i^{\mathrm{EV}} = \omega^{\mathrm{EV}}(-1 - \omega^{\mathrm{PP}}f'_P + \omega^{\mathrm{PS}}\omega^{\mathrm{SP}}f'_Pf'_Sf'_Ef'_V \tag{S29}$$

$$k_{ii}^{\mathrm{EV}} = \omega^{\mathrm{PV}}f'_P(\omega^{\mathrm{EP}} - \omega^{\mathrm{ES}}\omega^{\mathrm{SP}}f'_S)f'_Ef'_V \tag{S30}$$

$$k_{iiii}^{\mathrm{EV}} = \omega^{\mathrm{SV}}(\omega^{\mathrm{ES}}(1 + \omega^{\mathrm{PP}}f'_P) - \omega^{\mathrm{EP}}\omega^{\mathrm{PS}}f'_P)f'_Sf'_Ef'_V \tag{S31}$$

3.7 Matrix symmetries in simplified circuit

In Figure 3 and Figure S6 we find that beyond the *Hidden response symmetries* there are further symmetries in the linear response matrix. First, the response of E and PV to perturbations is highly correlated. Second, the response of all cell-types in response to an E perturbation are mirror imaged to those elicited by a PV perturbation. In summary, the first two rows and the first two columns of the response matrix $\chi_{\alpha\beta}$ appear highly correlated. To investigate the origin of this correlations, we study a simplified system, in which E and PV are not only similary but multiple of one another. We replace the measured PV data by $r_{\mathrm{PV}} = s r_{\mathrm{E}}$ with $s = 1.2$, and repeat the fitting procedure. In those cases, and as shown in Figure S7c, the connectivity matrix has the following structure:

$$\begin{bmatrix} \omega^{\mathrm{EE}} = a_1 + s\omega^{\mathrm{EP}} & \omega^{\mathrm{EP}} = 1 + \omega^{\mathrm{PP}} & \omega^{\mathrm{ES}} = \omega^{\mathrm{PS}} & \omega^{\mathrm{EV}} = \omega^{\mathrm{PV}} \approx 0 \\ \omega^{\mathrm{PE}} = a_2 + s\omega^{\mathrm{PP}} & \omega^{\mathrm{PP}} & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \approx 0 \\ \omega^{\mathrm{SE}} = a_3 + s\omega^{\mathrm{SP}} & \omega^{\mathrm{SP}} & 0 & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} = s\omega^{\mathrm{VP}} & \omega^{\mathrm{VP}} & \omega^{\mathrm{VS}} & 0 \end{bmatrix} \tag{S32}$$

Using the definition in Eq. (S19), this boils down to understanding the relationship between $k^{\mathrm{E}\alpha}$ and $k^{\mathrm{P}\alpha}$ on the one hand and the relationship between $k^{\alpha\mathrm{E}}$ and $k^{\alpha\,\mathrm{P}}$ on the other.

If we further take $a_1 = a_2 = 1 = a_3 = 1$ (from the fits these are $a_1 = 0.85$, $a_2 = 0.9$ and $a_3 = 1$), we find that

$$\begin{aligned} k^{PE} &= f'_E f'_P + \omega^{\mathrm{PP}} s f'_E f'_P - \omega^{\mathrm{PS}} f'_E f'_P f'_S - \omega^{\mathrm{PS}}\omega^{\mathrm{SP}} s f'_E f'_P f'_S + \\ &\qquad \left(-\omega^{\mathrm{SV}}\omega^{\mathrm{VS}} + \omega^{\mathrm{PS}}\omega^{\mathrm{SV}}\omega^{\mathrm{VP}} s - \omega^{\mathrm{PP}}\omega^{\mathrm{SV}}\omega^{\mathrm{VS}} s\right) f'_E f'_P f'_S f'_V \end{aligned} \tag{S33}$$

$$\begin{aligned} k^{PP} &= -f'_E f'_P - \omega^{\mathrm{PP}} s f'_E f'_P + \omega^{\mathrm{PS}} f'_E f'_P f'_S + \omega^{\mathrm{PS}}\omega^{\mathrm{SP}} s f'_E f'_P f'_S - (1 - s f'_E)(\omega^{\mathrm{SV}}\omega^{\mathrm{VS}} f'_S f'_V - 1) f'_P - \\ &\qquad \left(-\omega^{\mathrm{SV}}\omega^{\mathrm{VS}} + \omega^{\mathrm{PS}}\omega^{\mathrm{SV}}\omega^{\mathrm{VP}} s - \omega^{\mathrm{PP}}\omega^{\mathrm{SV}}\omega^{\mathrm{VS}} s\right) f'_E f'_P f'_S f'_V \end{aligned} \tag{S34}$$

$$k^{EE} = f'_E (1 + \omega^{\mathrm{PP}} f'_P) - \omega^{\mathrm{PS}}\omega^{\mathrm{SP}} f'_E f'_P f'_S + \left(\omega^{\mathrm{PS}}\omega^{\mathrm{SV}}\omega^{\mathrm{VP}} - \omega^{\mathrm{PP}}\omega^{\mathrm{SV}}\omega^{\mathrm{VS}}\right) f'_E f'_P f'_S f'_V - \omega^{\mathrm{SV}}\omega^{\mathrm{VS}} f'_E f'_S f'_V \tag{S35}$$

$$k^{EP} = -f'_E (1 + \omega^{\mathrm{PP}}) f'_P + \omega^{\mathrm{PS}}\omega^{\mathrm{SP}} f'_E f'_P f'_S - \left(\omega^{\mathrm{PS}}\omega^{\mathrm{SV}}\omega^{\mathrm{VP}} + \omega^{\mathrm{PP}}\omega^{\mathrm{SV}}\omega^{\mathrm{VS}}\right) f'_E f'_P f'_S f'_V + \omega^{\mathrm{SV}}\omega^{\mathrm{VS}} f'_E f'_P f'_S f'_V \tag{S36}$$

$$k^{ES} = \omega^{\mathrm{PS}} f'_E f'_S (-1 + f'_P) \tag{S37}$$

$$k^{PS} = \omega^{\mathrm{PS}} f'_P f'_S (-1 + s f'_E) \tag{S38}$$

$$k^{EV} = \omega^{\mathrm{PS}}\omega^{\mathrm{SV}} f'_E f'_S f'_V (1 - f'_P) \tag{S39}$$

$$k^{PV} = \omega^{\mathrm{PS}}\omega^{\mathrm{SV}} f'_P f'_S f'_V (1 - s f'_E) \tag{S40}$$

$$\tag{S41}$$

$$k^{SE} = (1 + \omega^{\mathrm{SP}} s) f'_E f'_S + \omega^{\mathrm{PP}} f'_E f'_P f'_S - \omega^{\mathrm{SP}} f'_E f'_P f'_S + (f'_P - s)\omega^{\mathrm{SV}}\omega^{\mathrm{VP}} f'_E f'_S f'_V \tag{S42}$$

$$k^{SP} = -(f'_E + \omega^{\mathrm{SP}}) f'_P f'_S - \omega^{\mathrm{PP}} f'_E f'_P f'_S + \omega^{\mathrm{SP}} f'_E f'_P f'_S - (f'_E - 1)\omega^{\mathrm{SV}}\omega^{\mathrm{VP}} f'_P f'_S f'_V \tag{S43}$$

$$k^{VE} = \omega^{\mathrm{VP}} s f'_E f'_V - \omega^{\mathrm{VP}} f'_E f'_P f'_V - \omega^{\mathrm{VS}} f'_E f'_S f'_V - \omega^{\mathrm{SP}}\omega^{\mathrm{VS}} s f'_E f'_S f'_V + \left(\omega^{\mathrm{PS}}\omega^{\mathrm{VP}} - \omega^{\mathrm{PP}}\omega^{\mathrm{VS}} + \omega^{\mathrm{SP}}\omega^{\mathrm{VS}}\right) f'_E f'_P f'_S f'_V \tag{S44}$$

$$k^{VP} = -\omega^{\mathrm{VP}} f'_P f'_V + \omega^{\mathrm{VP}} f'_E f'_P f'_V + \omega^{\mathrm{VS}} f'_E f'_P f'_S f'_V + \omega^{\mathrm{SP}}\omega^{\mathrm{VS}} f'_P f'_S f'_V - \left(\omega^{\mathrm{PS}}\omega^{\mathrm{VP}} - \omega^{\mathrm{PP}}\omega^{\mathrm{VS}} + \omega^{\mathrm{SP}}\omega^{\mathrm{VS}}\right) f'_E f'_P f'_S f'_V \tag{S45}$$

These expressions are not identical, but considering that $f'_P = 2\sqrt{r_P} \approx 2\sqrt{s r_E} = \sqrt{s} f'_E$, and that $\sqrt{s} = \sqrt{1.2}$ is close to one, most of these equations will be numerically similar.

3.8 Transformation to firing rate effect on the linear response.

To understand how the conclusions derived here would be modified by considering firing rates instead of deconvolved calcium imaging data we follow[57], where it is reported that calcium activity $\frac{\Delta F}{F}$ and firing rates can be related via a linear relationship. In general, given a power law input-output function $f(z) = []_+^{\xi}$, we can define a class of equivalent models by redefining activity together with weights and inputs

$$r^{\text{new}} = A^{\xi} r \tag{S46}$$

$$W^{\text{new}} = AWA^{-\xi} \tag{S47}$$

$$h^{\text{new}} = Ah \tag{S48}$$

Where A is the diagonal transformation matrix from calcium activity r to firing rates $rtextnew$. The Jacobian and the linear response matrix of this new system are related by:

$$J^{\text{new}} = A^{\xi} J A^{-\xi} \qquad R^{\text{new}} = A^{\xi} R A^{-1} \tag{S49}$$

In particular given that the new and old Jacobian are related by a similarity transformation, this change of variables (or the equivalence class) will not change the stability. The the linear response can have re-scaled values but will preserve sign, and the *Hidden response symmetries* equations will be re-scaled.

4 High-dimensional circuit models

In this section we describe the high-dimensional models. These are rate networks with power-law non-linearity $f(z) = []_+^{\xi}$ as before with $\xi = 2$. The network has $n = 4$ populations with have different number of neurons N^{α} for $\alpha = \{E, PV, S, V\}$, which we take to be a fraction $q = [0.8, 0.1, 0.05, 0.05]$ of the total number of cells N. The activity of the unit i in the population α, r_i^{α} in the steady state will be given by:

$$r_i^{\alpha} = f(z_i^{\alpha}) \qquad z_i^{\alpha} = \left(\sum_{\beta}^{n} \sum_{j}^{N^{\beta}} w_{ij}^{\alpha\beta} r_j^{\beta} + h_i^{\alpha} \right) \tag{S50}$$

Where the connectivity elements $w_{ij}^{\alpha\beta}$ are Gaussian distributed with mean and variance defined by:

$$\langle w_{ij}^{\alpha\beta} \rangle = w^{\alpha\beta}/N \qquad \langle (w_{ij}^{\alpha\beta})^2 \rangle - \langle w_{ij}^{\alpha\beta} \rangle^2 = {\sigma^{\alpha\beta}}^2/N \tag{S51}$$

The inputs to each unit h_i^{α} are also Gaussian distributed with mean $\langle h_i^{\alpha} \rangle = h_0^{\alpha}$ and variance $\langle h_i^{\alpha 2} \rangle - \langle h_i^{\alpha} \rangle^2 = (\lambda^{\alpha})^2$. The function f is a chosen nonlinearity and τ_i^{α} is the unit's time constant. The steady state solution of Eq. (S50) can be re written as

$$z_i^{\alpha} = \sum_{\beta}^{n} \sum_{j}^{N^{\beta}} w_{ij}^{\alpha\beta} f(z_i^{\alpha}) + h_i^{\alpha} \qquad r_i^{\alpha} = f(z_i^{\alpha}) \tag{S52}$$

4.1 Set-up and mean field equations

In order to compute, for each set of parameters, the mean and variance of the activity in each population self-consistently, we follow the approach in Kadmon and Sompolinsky[26]. The input to a cell $z_i^\alpha =$ can be described as fluctuations around a mean: $z_i^\alpha = u^\alpha + \delta z_i^\alpha$. Defining

$$m^\alpha = \langle f(z_i^\alpha) \rangle \tag{S53}$$

$$v^\alpha = \langle f(z_i^\alpha)^2 \rangle \tag{S54}$$

$$q^\alpha = N^\alpha / N \tag{S55}$$

By taking the mean and the variance of Eq. (S50) and incorporating the definitions above, we re-obtain the self-consistent equations for the mean and the variance of z, given by u^α and Δ^α

$$u^\alpha = \sum_\beta w^{\alpha\beta} q^\beta m^\beta + h^\alpha \tag{S56}$$

$$\Delta^\alpha = \sum_\beta (\sigma^{\alpha\beta})^2 q^\beta v^\beta + (\lambda^\alpha)^2 \tag{S57}$$

where

$$m^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z) e^{-z^2/2} dz \tag{S58}$$

$$v^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z)^2 e^{-z^2/2} dz \tag{S59}$$

We observe that if there is no disorder, Eqs. (S56) and (S58) reduce to the low-dimensional model from Eq. (S3) with $\omega^{\alpha\beta} = w^{\alpha\beta} q^\beta$ and $m^\alpha = r^\alpha$.

4.2 Mean field perturbation

If we imagine an *homogeneous* perturbation to the entire population α, the change in response of each cell when perturbed by the laser will be given by

$$\frac{dr_i^\alpha}{dL} = f'(u^\alpha + \sqrt{\Delta^\alpha} z_i) \left(\frac{du^\alpha}{dL} + \frac{1}{2\sqrt{\Delta^\alpha}} \frac{d\Delta^\alpha}{dL} z_i \right) \tag{S60}$$

Taking the average and using Eq. (S58), we find that the mean of the response distribution to laser perturbation is given by the change in the mean activity of the population:

$$
\begin{aligned}
\langle \frac{dr_i^\alpha}{dL} \rangle =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(u^\alpha + \sqrt{\Delta^\alpha} z) \left(\frac{du^\alpha}{dL} + \frac{1}{2\sqrt{\Delta^\alpha}} \frac{d\Delta^\alpha}{dL} z \right) e^{-z^2/2} dz \\
=& \frac{d}{dL} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z) e^{-z^2/2} dz \right) \\
=& \frac{dm^\alpha}{dL}
\end{aligned}
\tag{S61}
$$

This equation relates how the mean of the distribution of responses to perturbation relates to the response to the mean activity.

4.3 Data Fitting

To fit the system defined by Eqs. (S56,S58), we used that the distribution of activity of a population α with input output function of the form $r = [z]_+^\xi$ can be written, when assuming the inputs are Gaussian, as a function of the mean input u^α and its variance Δ^α [35]:

$$
P^\alpha(r) = \frac{1}{\sqrt{2\pi} r^{1-1/\xi} \xi \Delta} e^{-\frac{(r^{1/\xi} - u^\alpha)^2}{2\Delta^\alpha}} \Theta(r) + \frac{1}{2}(1 - \text{erf}(\frac{u^\alpha}{\sqrt{2\Delta^\alpha}}))\delta(r) \tag{S62}
$$

$$
P^\alpha(r) = P^+(r)\Theta(r) + P^0\delta(r) \tag{S63}
$$

In order to find the parameters that better approximate the real distribution of the data we proceed as follows (Fig. 5**a**): For each cell-type α and each contrast value c we fit, via maximum likelihood, the analytical form described above (i.e. Eq. S62) with a fixed value $\xi = 2$, to the distribution of activity r obtained in the experiments (we call this $P_\text{d}^\alpha(r, \mu_\text{d}^\alpha(c), \Delta_\text{d}^\alpha(c))$). These are the full lines in Fig. 5**d**. That gives us an estimate of the mean ($\mu_\text{d}^\alpha(c)$) and variance ($\Delta_\text{d}^\alpha(c)$) of the input distributions that come from the data for each cell-type and each value of the contrast. To find which parameters $w^{\alpha\beta}$, $\sigma^{\alpha\beta}$, h^α and λ^α (and assuming that the inputs depend linearly on the contrast $h^\alpha = h_b^\alpha + h_c^\alpha c$) best fit the data we proceed as follows: we do ABC search from prior distributions for the mean and variance of the weights and inputs to this network to build multiple instances of $P_\text{mf}^\alpha(r, \mu_\text{mf}^\alpha(c), \Delta_\text{mf}^\alpha(c))$. The priors for $w^{\alpha\beta}$ and h^α were Gaussian distributions with around the low-dimensional system values (with a 5% std) and a normal prior for $\sigma^{\alpha\beta}$ and λ^α. We define an error that depends uniquely on $\mu_\text{d}^\alpha(c), \Delta_\text{d}^\alpha(c), \mu_\text{mf}^\alpha(c), \Delta_\text{mf}^\alpha(c)$. Specifically, we define the total error as the sum of the squared norm of the matrix of the Kullback-Leibler divergences between these two distributions:

$$
E = \sum_c \sum_\alpha D(P_\text{d}^\alpha(c) || P_\text{mf}^\alpha(c))^2 \tag{S64}
$$

where, and dropping termporarily the dependence on the contrast for ease of notation we have:

$$
D(P_\text{d}^\alpha || P_\text{mf}^\alpha) = \int_{-\infty}^{\infty} P_\text{d}^\alpha(r) \log \frac{P_\text{d}^\alpha(r)}{P_\text{mf}^\alpha(r)} dr = P_\text{d}^{0,\alpha} \log \frac{P_\text{d}^{0,\alpha}}{P_\text{mf}^{0,\alpha}} + \int_{0^+}^{\infty} P_\text{d}^{+,\alpha}(r) \log \frac{P_\text{d}^{+,\alpha}(r)}{P_\text{mf}^{+,\alpha}(r)} dr = I_A^\alpha + I_B^\alpha \tag{S65}
$$

with

$$\mathrm{I}_A^\alpha = \frac{1}{2}\,\mathrm{erf}\,(\frac{\mu_\mathrm{d}^\alpha}{\sqrt{2\Delta_\mathrm{d}^\alpha}})\log\left(\frac{\mathrm{erf}\left(\frac{\mu_\mathrm{d}^\alpha}{\sqrt{2\Delta_\mathrm{d}^\alpha}}\right)}{\mathrm{erf}\left(\frac{\mu_\mathrm{mf}^\alpha}{\sqrt{2\Delta_\mathrm{mf}^\alpha}}\right)}\right) \tag{S66}$$

$$\mathrm{I}_B^\alpha = \frac{1}{2}\left(1+\mathrm{erf}\,(\frac{\mu_\mathrm{d}^\alpha}{\sqrt{2\Delta_\mathrm{d}^\alpha}})\right)\log\frac{\sqrt{\Delta_\mathrm{mf}^\alpha}}{\sqrt{\Delta_\mathrm{d}^\alpha}} + \frac{1}{4\Delta_\mathrm{mf}^\alpha}\left(\left(\Delta_\mathrm{d}^\alpha - \Delta_\mathrm{mf}^\alpha + \left(\mu_\mathrm{d}^\alpha - \mu_\mathrm{mf}^\alpha\right)^2\right)\left(1+\mathrm{erf}\,(\frac{\mu_\mathrm{d}^\alpha}{\sqrt{2\Delta_\mathrm{d}^\alpha}})\right)+\right. \tag{S67}$$

$$\left.\sqrt{\frac{2}{\pi\Delta_\mathrm{d}^\alpha}}\,e^{-\frac{\mu_\mathrm{d}^{\alpha 2}}{2\Delta_\mathrm{d}^\alpha}}\left(\Delta_\mathrm{d}^\alpha(\mu_\mathrm{d}^\alpha - 2\mu_\mathrm{mf}^\alpha) + \Delta_\mathrm{mf}^\alpha\mu_\mathrm{d}^\alpha\right)\right) \tag{S68}$$

Instead of following the gradient to find an optimal solution we keep the solutions that have a sufficiently small error from the random sampling. This defines a family of high-dimensional models (Fig. 5**b,c**, see also Fig. S8) with skewed distributions that are in good agreement with the calcium activity, and capture not only the nonlinear dependence of the mean the activity but also spreading out with increasing contrast.

5 Analytical approach to linear response of disordered networks

5.1 Set up

We call the steady state solution of Eq. (S50) ${}^*r_i^\alpha$ and the steady state input ${}^*z_i^\alpha$. The time evolution of the response to a perturbation δh_i^α, can be described by the dynamics of δr_i^α:

$$\tau_i^\alpha \delta\dot{r}_i^\alpha = -\delta r_i^\alpha + f'^{\alpha}_i \cdot \left(\sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta}\delta r_j^\beta + \delta h_i^\alpha\right) \qquad f'^{\alpha}_i = f'^{\alpha}({}^*z_i^\alpha) = f'\left(\sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta}\,{}^*r_j^\beta + h_i^\alpha\right) \tag{S69}$$

Switching form now onwards to matrix notation, we define: $F_{ij}^{\alpha\beta} = \delta_{ij}\delta_{\alpha\beta}f_i'^\alpha$, the diagonal matrix of derivatives, where δ is the Kroenecker delta, and $f_i'^\alpha$ is the gain of neuron i in population α. The connectivity matrix W has elements $w_{ij}^{\alpha\beta}$. The steady state response to an arbitrary increase in the input given by δh will be:

$$\delta\vec{r} = (F^{-1} - W)^{-1}\vec{\delta}h \qquad \vec{\delta}r = R\vec{\delta}h \tag{S70}$$

Which defines the high-dimensional linear response matrix $R = (F^{-1} - W)^{-1}$. If we constrain the cell-type-specific variance to be low rank, meaning that the block-wise variance of W (defined in Eq. (S51)) is written as $(\sigma^{\alpha\beta})^2/N = \nu^\alpha\kappa^\beta/N$, we can write W as the sum of a homogeneous component W_0, with ${W_0}_{ij}^{\alpha\beta} = w^{\alpha\beta}$ and a random component $\Pi_L J \Pi_R$, where J is a matrix with Gaussian distributed random numbers with zero mean and unit variance , and Π_L and Π_R are non-random diagonal matrices:

$$W = W_0 - \Pi_L J \Pi_R \qquad \text{with} \qquad \Pi_L = \delta_{ij}\sqrt{\kappa_{\alpha_j}} \qquad \Pi_R = \delta_{ij}\sqrt{\nu_{\alpha_j}} \tag{S71}$$

5.2 Linearization around the homogeneous fixed point

The mathematical treatment we are going to outline later is only possible in linear system in which the disorder does not affect the gain of each neuron. All the linear response calculations of the following sections will assume that the linearized system can be written as

$$\tau_i^\alpha \delta \dot{r}_i^\alpha = -\delta r_i^\alpha + f'^\alpha \cdot \left(\sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta} \delta r_j^\beta + \delta h_i^\alpha \right) \qquad f'^\alpha_i = f'^\alpha = f' \left(\sum_\beta^n \sum_j^{N^\beta} w^{\alpha\beta *} r^\beta + h^\alpha \right) \tag{S72}$$

What this means, is that, we solve the non-disordered system to compute f'^α and look at a linear disordered system around the HFP.

5.3 General framework established in Ahmadian et al.[36]

Using results from[36] we find that in the special case of $f'^\alpha_i = f'^\alpha$ described above, the mean linear response matrix over different instantiations of the disorder is the linear response of the non- disordered case:

$$\langle (F^{-1} - W)^{-1} \rangle_J = \langle (F^{-1} - W_0 + \Pi_L J \Pi_R)^{-1} \rangle_J = (F^{-1} - W_0)^{-1} = R^0 \tag{S73}$$

This fundamental relation links the mean of the distribution of responses to the response of the non-disordered system, its general in linear networks and works as a good insightful approximation in this case of study. Generally, in experiments, we will have a perturbation pattern δh describing the proportion of stimulation each neurons receive, and a measuring vector δb, describing which are the neurons contributing (linearly) to the signal that we are going to be monitoring $s = \vec{\delta r}^\intercal \delta b$. We are interested in computing the mean and variance of that signal across different instantiations of the disorder. By defining:

$$\text{the } \textit{measuring} \text{ matrix} \qquad B = \delta b \delta b^\intercal \tag{S74}$$

$$\text{the } \textit{optogenetic perturbation} \text{ matrix} \qquad \Sigma = \delta h \delta h^\intercal \tag{S75}$$

we can write the second moment of that measured signal s[36]:

$$\langle s^2 \rangle = \langle (\vec{\delta r}^\intercal \delta b)^2 \rangle = \langle \vec{\delta r}^\intercal B \vec{\delta r} \rangle = \mathscr{F} + \Delta\mathscr{F} \tag{S76}$$

where

$$\mathscr{F} = Tr(BR^0 \Sigma R^{0\intercal}) \qquad \Delta\mathscr{F} = \frac{1}{N} \frac{Tr\left(BR^0 \Pi_L \Pi_L R^{0\intercal}\right) Tr\left(\Pi_R \Pi_R R^0 \Sigma R^{0\intercal}\right)}{1 - \frac{1}{N} Tr\left(\Pi_R \Pi_R R^0 \Pi_L \Pi_L R^{0\intercal}\right)} \tag{S77}$$

Where we used the definitions in Eq. (S71). We observe that in the absence of disorder, in which $W = W_0$, then $\Delta\mathscr{F} = 0$ and the signal we measure will be given only by $s = \delta b R^0 \delta h^\intercal$

In the case in which we are interested in looking at single neuron statistics, then we have $\delta b = e_i$ with $e_i = \{0, ..., 1, ..., 0\}$

$$\langle \delta r_i \rangle_J = e_i R^0 \vec{\delta h} \qquad \text{where} \qquad e_i = \{0, ..., 1, ..., 0\} \tag{S78}$$

Meaning that for each neuron, the distribution of linear responses over different instantiations of the connectivity will have a mean given by the linear response in the absence of disorder (due to Eq. S73) and the variance Λ_i

$$\Lambda_i^2 = \langle \delta r_i^2 \rangle_J - \langle \delta r_i \rangle_J^2 = \langle \delta r_i^2 \rangle_J - \mathscr{F} = \Delta \mathscr{F} \tag{S79}$$

Equations (S73), (S76) and (S77) are general formulas of how to compute the mean and the variance of the linear response distributions as a function of the optogenetic perturbation Σ and the observation matrix B. In the following sections we will explicitly compute the mean response matrix R^0 for both full and low rank connectivity and the variance in different optogenetic perturbation configurations.

5.4 Computation of the response R^0 matrix without disorder

For computing R^0 (given by Eq. (S73)) we write the block-structured matrix W_0 as a function of the low-dimensional system connectivity $\omega^{\alpha\beta} = w^{\alpha\beta} q^{\beta}$. We choose the matrices U and V with columns given by vectors $u^{\alpha} = \frac{1}{N_\alpha}\delta_{i\in\alpha}$ and $v^{\alpha} = \delta_{i\in\alpha}$, meaning that $u^{(k)} = \frac{1}{N_k}(\underbrace{0,\dots,0}_{\sum_{l=1}^{k-1} N_l}, \underbrace{1,\dots,1}_{N_k}, \underbrace{0,\dots,0}_{N-\sum_{l=1}^{k} N_l})$ and similarly for v.

$$W_0 = V\omega U^{\intercal} \tag{S80}$$

Where $w^{\alpha\beta}$ and q^{α} were introduced in Eqs (S51) and (S55) respectively. To obtain R^0 defined in Eq. S73 we are going to exploit the fact that this is a low rank matrix. Depending on whether ω is also low rank or not, we will need to consider different strategies

1) case of invertible ω

If ω is invertible, we can use the Woodsbury lemma to find a succinct expression for R^0:

$$R^0 = \left(F^{-1} - W_0\right)^{-1} = \left(F^{-1} - V\omega U^{\intercal}\right)^{-1} = F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^{\intercal} F \tag{S81}$$

Introducing the notation α_i as the population to which the neuron i belongs to, the entries of the response function can be written as

$$R^0_{ij} = \delta_{ij} f'_{\alpha_i} - \frac{f'_{\alpha_i} f'_{\alpha_j}}{N_{\alpha_j}} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S82}$$

Where $\mathbf{f}'$ was defined in Eq. (S7). We note that for this expression to be valid, ω needs to be invertible and in particular full rank. We also note that this expression is given by two terms: the first one, private to each neuron, is only non-zero if we are observing the same neuron that we are stimulating, while the second term, which depends on which population the stimulated neuron belongs to and which population the observed neuron belongs to, but is independent on whether the perturbed neuron is the observed one.

We define $S_{\alpha_i\alpha_j}$, the sum of the linear response of a single neuron in population α_i to a homogeneous input to the neurons in population α_j

$$S_{\alpha_i\alpha_j} = \delta_{\alpha_i\alpha_j} f'_{\alpha_j} - f'_{\alpha_i} f'_{\alpha_j} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S83}$$

Replacing into Eq. (S83) into (S82), we obtain an expression for the linear response which will be useful in later sections:

$$R^0_{ij} = \delta_{ij} f'_{\alpha_i} + \frac{S_{\alpha_i\alpha_j}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j} f'_{\alpha_j}}{N_{\alpha_j}} \tag{S84}$$

We point out that the Eq. (S83) is independent of N, and is finite in the limit of large N.

2) case of rank-one ω

In the case in which ω is rank-one, the operation in Eq. (S81) cannot be performed. Instead we can write:

$$W_0 = \frac{vu^\intercal}{N} \qquad v = \{1, ..., 1, ..., 1\} \tag{S85}$$

$$u = \{\underbrace{w_1, \cdots, w_1}_{N_1}, \underbrace{w_2, \cdots, w_2}_{N_2}, \cdots, \underbrace{w_n, \cdots, w_n}_{N_n}\} \tag{S86}$$

By means of the Sherman–Morrison formula we find that

$$R^0 = (F^{-1} - W_0)^{-1} = F + \frac{1}{N} \frac{Fvu^T F}{1 - \frac{u^T Fv}{N}} \tag{S87}$$

Where the denominator is always positive given that $D = \left(1 - \frac{u^T Fv}{N}\right) = \det\left(F^{-1} - \frac{vu^T}{N}\right) \det(F) = \det(I - FW_0)$.

$$R^0_{ij} = \delta_{ij} f'_{\alpha_j} + \frac{1}{N} \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} = \delta_{ij} f'_{\alpha_j} + \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S88}$$

Again defining $S_{\alpha_i\alpha_j}$ as the sum of the linear response of a single neuron in population α_i to a homogeneous input to the neurons in population α_j

$$S_{\alpha_i\alpha_j} = \delta_{\alpha_i\alpha_j} f'_{\alpha_j} + N_{\alpha_j} \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S89}$$

Were we point out that the above expression is also finite in the large N limit. We can write the linear response identically as in Eq. (S84) also in this case.

5.5 Response distribution to partial (homogeneous) perturbations: Mean term

We want the sum of the linear response of a single neuron in population α_i to a perturbation to a fraction γ_j to the neurons in population α_j, for arbitrary populations α_j. Within each perturbed population, We call the set of perturbed neurons $\mathscr{P}_{\alpha_j}$. If we perturb γ_{α_j} neurons in population α_j then we find that the response of the neurons in population in population α_j that were stimulated have a mean response that depends on whether they were directly stimulated or not.

1) case of invertible ω

In the case in which ω is full rank (R_{ij} is given by (S82)), if we perturb γ_η neurons in population η, for an arbitrary amount of populations, the total perturbation vector will then be $\delta h = \{\delta h_1, \delta h_2, \cdots, \delta h_n\}$, where $h_\eta = \{0,\cdots,0,\underbrace{1,\cdots,1}_{\gamma_\eta N_\eta},0,\cdots,0\}$, then we find that the response of the neurons is given by

$$\mu_i = \sum_j R_{ij}\delta h_j = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} - f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} f'_{\alpha_j}\left[\left(\mathbf{f'}-\omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S90}$$

which simply amounts to two different means for the perturbed and non perturbed cells. One for neurons that were stimulated given by

$$\mu^{IN}_{i\in\mathscr{P}_{\alpha_i}} = f'_{\alpha_i} - f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} f'_{\alpha_j}\left[\left(\mathbf{f'}-\omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S91}$$

whereas the neurons in α_j that were *not* stimulated and the neurons from other populations follow the equation:

$$\mu^{OUT}_{i\notin\mathscr{P}_{\alpha_i}} = -f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} f'_{\alpha_j}\left[\left(\mathbf{f'}-\omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S92}$$

We note that in this statements about the critical fraction depend critically on the sign of $\left(\mathbf{f'}-\omega^{-1}\right)^{-1}$. We can define a matrix

$$\begin{aligned}\chi^{\boldsymbol{\gamma}} =& \mathbf{f'}\delta_p - \mathbf{f'}\left(\mathbf{f'}-\omega^{-1}\right)^{-1}\mathbf{f'}\boldsymbol{\gamma} && \text{(S93)}\\ =& \mathbf{f'}\delta_p + \mathbf{f'}\omega\left(\mathbf{f'}^{-1}-\omega\right)^{-1}\boldsymbol{\gamma} && \text{(S94)}\\ =& \mathbf{f'}\delta_p + \mathbf{f'}\omega\chi\boldsymbol{\gamma} && \text{(S95)}\end{aligned}$$

where $\delta_p = 0$ if we are describing the mean of the non-perturbed population and $\delta_p = 1$ otherwise, that summarized the above.

In the case in which we want to study the paradoxical response, meaning that we are stimulating and measuring only one and the same population we find that using that $\sum_{\alpha_j}(\mathbf{f'}^{-1}-\omega)_{\alpha_i\alpha_j}\chi_{\alpha_j\alpha_i} = 1$ (matrix times its inverse is the identity) we have that $[\omega\chi]_{\alpha_i\alpha_i} = \frac{\chi_{\alpha_i\alpha_i}}{f'_{\alpha_i}} - 1$ We can write (S90) as

$$\mu_i = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} + f'_{\alpha_i}\gamma_{\alpha_i}[\omega\chi]_{\alpha_i\alpha_i} = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} + \gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i}) \tag{S96}$$

If the response is paradoxical in the low-dimensional system ($\chi_{\alpha_i\alpha_i} < 0$), the distribution of responses of non stimulated neurons will have a negative mean, and will become increasingly negative the more neurons are perturbed. Those neurons that are being stimulated, for an arbitrarily small fraction of perturbed cells, the above term is positive and therefore there will be a critical fraction of perturbed cells for the mean to change sign. This will be

$$0 < \frac{f'_{\alpha_i}}{f'_{\alpha_i} - \chi_{\alpha_i\alpha_i}} < \gamma^C_{\alpha_i} < 1 \tag{S97}$$

In Fig. 8, the *fractional paradoxical* effect occurs while the perturbed cells are not responding paradoxically. Nevertheless, before the mean of the distribution of perturbed cells changes sign, the distribution itself shifts left and therefore this critical fraction is different from the critical fraction for which the system is exhibiting a *fractional paradoxical* effect.

2) case of rank-one ω

$$R_{ij} = \delta_{ij} f'_{\alpha_j} - \frac{1}{N}\frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{1-\sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} = \delta_{ij} f'_{\alpha_j} + \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S98}$$

Identically as above, neurons that are directly stimulated will have a response given by

$$\mu_i^{IN} = f'_{\alpha_i} + f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} q_{\alpha_j}\frac{f'_{\alpha_j} w_{\alpha_j}}{1-\sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} \tag{S99}$$

whereas the neurons in α_j that were *not* stimulated and the neurons from other populations follow the equation:

$$\mu_i^{OUT} = f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} q_{\alpha_j}\frac{f'_{\alpha_j} w_{\alpha_j}}{1-\sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} \tag{S100}$$

Critical fraction

In the case in which we only have and EI circuit, and we stimulate only the inhibitory population, we can see that for inhibitory neurons in which w_j is negative, the response of the neurons that were not stimulated is always paradoxical (meaning that Eq. (S100) is always negative), but the response of those neurons that were stimulated will only be paradoxical when $\mu_i^{IN} < 0$

$$\frac{1-\sum\limits_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{q_{\alpha_j} f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S101}$$

First lets consider the case in which we have a fixed amount of neurons but we have an increasing amount of populations n. Given that $N = \sum_{\alpha_k} N_{\alpha_k}$ if we take $N_{\alpha_k} = N/n$ then $q_{\alpha_k} = N_{\alpha_k}/N = 1/n$. We find that the critical fraction in (S101) is now

$$\frac{n-\sum\limits_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S102}$$

We find that given a fixed sum of w_i and fixed N, the fraction of stimulated neurons γ_k needs to increase linearly in n to have a paradoxical response.

Comparison with Sadeh et al.[25]

If now we would normalize the weights in Eq. (S85) as in[25], meaning that $w_j/N \to w_j/(N/n)$ (See Eq. 2 of their paper), the above equation would be

$$\frac{1-\sum_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{f'_{\alpha_j}|w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S103}$$

Taking $f'_{\alpha_k} = 1$ (linear network) we recover their result below Eq. 12 of their paper.

5.6 Response distribution to partial (homogeneous) perturbations: Variance term

From Eq. (S77) we know that the variance of the response is going to depend on the response of the system without disorder R^0. The goal of this section is writing R^0 in the form expressed in (S84). We will first find the general expression and then evaluate for particular cases: For that we write (S77) as

$$\Lambda^2 = \frac{1}{N}\frac{Tr\left(BR^0\Pi_L\Pi_L R^{0\intercal}\right)Tr\left(\Pi_R\Pi_R R^0\Sigma R^{0\intercal}\right)}{1-\frac{1}{N}Tr\left(\Pi_R\Pi_R R^0\Pi_L\Pi_L R^{0\intercal}\right)} = \frac{M.O}{1-D} \tag{S104}$$

Where the *optogenetic targeting* matrix $\Sigma = \delta h\delta h^\intercal$. If we write $\delta h = \{\delta h_1, \delta h_2, \cdots, \delta h_n\}$, where δh_η is the perturbation vector for the population η, then for each δh_η we can write $h_\eta = \{0,\cdots,0,\underbrace{1,\cdots,1}_{\gamma_\eta N_\eta},0,\cdots,0\}$, meaning that given n populations, there will be a vector with entries γ_η that tells us which is the fraction of neurons of each population that we are stimulating. Each element of the *optogenetic targeting* matrix will then be:

$$\Sigma_{jk} = \sum_\eta\sum_{\eta'}\delta_{\alpha_j\eta}\delta_{\alpha_k\eta'}\delta_{k\in\mathscr{P}_{\alpha_k}}\delta_{j\in\mathscr{P}_{\alpha_j}} \tag{S105}$$

Observation: The *optogenetic targeting* matrix has entries in the off diagonal terms.

We write down here the final expression for the variance of the response of a single neuron in population α_l while perturbing a fraction γ_η of the population η ($q_\eta = N_\eta/N$):

$$\Lambda^2_{\alpha_l} = \frac{\left(f'^2_{\alpha_l}\kappa_{\alpha_l} + \frac{1}{N}\left(\sum_\eta \frac{\kappa_\eta}{q_\eta}S^2_{\alpha_l\eta} - \frac{f'^2_{\alpha_l}\kappa_{\alpha_l}}{q_{\alpha_l}}\right)\right)\left(\sum_\eta \nu_\eta f'^2_\eta\gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta}\gamma_\eta\right)^2 q_{\eta'}\right)}{1-\sum_\eta \nu_\eta\left(q_\eta f'^2_\eta\kappa_\eta + \frac{1}{N}\left(q_\eta\sum_{\eta'}\frac{\kappa_{\eta'}}{q_{\eta'}}S^2_{\eta\eta'} - f'^2_\eta\kappa_\eta\right)\right)} \tag{S106}$$

We observe that in the large N limit the expression reduces to:

$$\Lambda^2_{\alpha_l} = \frac{f'^2_{\alpha_l}\kappa_{\alpha_l}\left(\sum_\eta \nu_\eta f'^2_\eta\gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta}\gamma_\eta\right)^2 q_{\eta'}\right)}{1-\sum_\eta \nu_\eta q_\eta f'^2_\eta\kappa_\eta} \tag{S107}$$

Which is independent of N iff γ_η is a finite fraction of the population. In the case in which a finite amount of neurons k are stimulated , $\gamma_\eta = k/N_\eta$ and the variance will vanish in the large N limt.

An interesting prediction is a nonlinear dependence of the variance of the populations with increasing fraction of stimulated neurons. The expression Eq. (S107) has a nonlinear term in the fraction of stimulated neurons in each population. When more than a single population is stimulated, there is also a term that nonlinearly mixes the fraction of interacting neurons. This results in non trivial dependences of the variance with the fraction of stimulated cells. Depending of the fraction of stimulated cells, the effect of increasing fraction of one-cell-type stimulation can be to narrow down the distributions or to broaden them. We name this a *second-order paradoxical effect*.

* Simplification: Non-structured variance

In the particular case in which the degree of disorder on the connectivity does not depend on the pre and the postsynaptic cell-type, i.e. when $\kappa_\alpha = \nu_\alpha = \sigma$ we obtain a simpler expression for the variance of the populations:

$$\Lambda^2_{\alpha_l} = \sigma^2 \frac{\left(f'^2_{\alpha_l} + \sum_\eta \frac{S^2_{\alpha\eta}}{N_\eta} - \frac{f'^2_{\alpha_l}}{N_{\alpha_l}} \right) \left(\sum_\eta f'_\eta{}^2 \gamma_\eta (1-\gamma_\eta) q_\eta + \sum_{\eta'} \left(\sum_\eta S_{\eta',\eta} \gamma_\eta \right)^2 q_{\eta'} \right)}{1 - \sigma^2 \sum_\eta \left(q_\eta f'_\eta{}^2 + q_\eta \sum_{\eta'} S^2_{\eta,\eta'} \frac{1}{N_{\eta'}} - \frac{1}{N} f'_\eta{}^2 \right)} \tag{S108}$$

5.7 Response distribution to partial (homogeneous) perturbations: Full Distribution

So far we computed the mean and the variance of the distribution of neurons to partial stimulation, and found that in the case in which γ is neither zero or one, i.e. in the case of partial stimulation, we will have a total distribution that is a mixture of Gaussians with means

$$\rho^{\mathrm{IN}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda_{\alpha_i}} \exp\left\{ -\frac{(x-\mu^{\mathrm{IN}}_{\alpha_i})^2}{2\Lambda^2_{\alpha_i}} \right\} \tag{S109}$$

$$\rho^{\mathrm{OUT}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda_{\alpha_i}} \exp\left\{ -\frac{(x-\mu^{\mathrm{OUT}}_{\alpha_i})^2}{2\Lambda^2_{\alpha_i}} \right\} \tag{S110}$$

So the total distribution of responses is

$$\rho_{\alpha_i} = \gamma_{\alpha_i} \rho^{\mathrm{IN}}_{\alpha_i} + (1-\gamma_{\alpha_i}) \rho^{\mathrm{OUT}}_{\alpha_i} \tag{S111}$$

Where $\mu^{\mathrm{IN}} = R^{IN}_{j \in \mathscr{P}_{\alpha_j}}$ and $\mu^{\mathrm{OUT}} = R^{OUT}_{j \in \mathscr{P}_{\alpha_j}}$ given by Eqs (S99, S100) for low rank ω or by (S91,S92) for invertible ω, and a variance given by Eq. (S106)

5.8 Simple description of the fractional paradoxical effect

The fractional paradoxical effect can be intuitively understood in the system without disorder (the EI, low-rank case of the non-disordered case was studied by[25]). In this case, the distribution of responses will be bimodal, with two delta functions at the values given by Eq. (S96). The density then will be given by the limit of vanishing variance of . (S111)

$$\rho_{\alpha_i}(x) = (1-\gamma_{\alpha_i})\delta(x - \gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i})) + \gamma_{\alpha_i}\delta(x - f'_{\alpha_i} - \gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i})) \tag{S112}$$

If the unit α_i is paradoxical in the low-dimensional system, then $\chi_{\alpha_i\alpha_i} < 0$. The left peak will always be negative, and for sufficiently small γ_{α_i} the peak of the perturbed cells will be positive. As computed in Eq. (S97), for values of γ_{α_i} smaller than $\gamma^C_{\alpha_i}$, the mean of the perturbed population will remain positive. In this range, increasing the fraction of perturbed cells, will result in a decrease of the mass of negative responses $^{0}_{-\infty}\rho_{\alpha_i}(x)dx$ like $(1-\delta)$. In the non-disordered case, as soon as $\gamma_{\alpha_i} > \gamma^C_{\alpha_i}$, the mass of negative responses is unity. Given that when working with the homogeneous approximation, the response of the non -disordered system is the mean of the response of the disordered system, the intuitions here apply to the mean of the disordered case.

5.9 Fractional paradoxical effect and link to a 5D low-dimensional system.

Here we show that the mean response of the perturbed population can be mapped to the response of a system with 5 dimensions, in which the α_i population, that here for simplicity we take to be PV, is split in a perturbed and non-perturbed population. We know that mapping a high-dimensional non-disordered network to a low D system can be done by rescaling the weights according to the fraction of cells in that population. That manipulation will not change the activity of either cell-type given that they receive the exact same input currents. The linear response of that system in consideration is , χ^5 is given by

$$\chi^5 = [f_5'^{-1} - \omega_5]^{-1} = -I \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}\gamma_P & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}\gamma_P - 1/f_P' & \omega^{\mathrm{PP}}(1-\gamma_P) & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}\gamma_P & \omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P' & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}\gamma_P & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}\gamma_P & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix}^{-1} \tag{S113}$$

$$\chi^5_{pp} = [f_5'^{-1} - \omega_5]^{-1}_{pp} = \frac{1}{\det\left[\mathbf{f'}_5^{-1} - \omega_5\right]} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P' & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S114}$$

$$\chi^5_{PP} = \frac{1}{\det\left[\mathbf{f'}_5^{-1} - \omega_5\right]} \Big(- \omega^{\mathrm{PE}} \det \begin{bmatrix} \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S115}$$

$$+ (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P') \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S116}$$

$$- \omega^{\mathrm{PS}} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S117}$$

$$+ \omega^{\mathrm{PV}} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} \end{bmatrix} \tag{S118}$$

Each 3D determinant is *minus* the minor of the original 4D matrix $(\mathbf{f'}^{-1} - \omega)$. Using that

$$\chi^5_{PP} = \frac{1}{\det\left[\mathbf{f'}_5^{-1} - \omega_5\right]} \left(\omega^{\mathrm{PE}} M_{PE}(1-\gamma_P) - (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P') M_{PP} + \omega^{\mathrm{PS}} M_{PS}(1-\gamma_P) - \omega^{\mathrm{PV}} M_{PV}(1-\gamma_P)\right) \tag{S119}$$

Where $M_{\alpha\beta}$ are the minors of the original 4D matrix $(\mathbf{f'}^{-1}-\omega)$. Using that $\chi_{\alpha\beta} = \frac{1}{\det\left(\mathbf{f'}^{-1}-\omega\right)}(-1)^{\alpha\beta}M_{\beta\alpha}$

$$\chi^5_{PP} = -\frac{\det\left(\mathbf{f'}^{-1}-\omega\right)}{\det\left[\mathbf{f'}_5^{-1}-\omega_5\right]}\left(\omega^{\mathrm{PE}}\chi_{EP}(1-\gamma_P)+(\omega^{\mathrm{PP}}(1-\gamma_P)-1/f'_P)\chi_{PP}+\omega^{\mathrm{PS}}\chi_{SP}(1-\gamma_P)+\omega^{\mathrm{PV}}\chi_{VP}(1-\gamma_P)\right) \tag{S120}$$

$$-\frac{\det\left(\mathbf{f'}^{-1}-\omega\right)}{\det\left[\mathbf{f'}_5^{-1}-\omega_5\right]}\left([\omega\chi]_{PP}(1-\gamma_P)-1/f'_P\chi_{PP}\right) \tag{S121}$$

Using again the trick that $[\omega\chi]_{\alpha_i\alpha_i} = \frac{\chi_{\alpha_i\alpha_i}}{f'_{\alpha_i}} - 1$

$$\chi^5_{PP} = \frac{\det\left(\mathbf{f'}^{-1}-\omega\right)}{\det\left[\mathbf{f'}_5^{-1}-\omega_5\right]}\left(\gamma_P[\omega\chi]_{PP}+1\right) \tag{S122}$$

Given that the mean response of the perturbed population in a high-dimensional system, given by Eq. (S90) (and also Eq. S143) is $\chi^{\boldsymbol{\gamma}}_{PP} = f'_P(\gamma_P[\omega\chi]_{PP}+1)$, we obtain that

$$f'_P\chi^5_{PP} = \frac{\det\left(\mathbf{f'}^{-1}-\omega\right)}{\det\left[\mathbf{f'}_5^{-1}-\omega_5\right]}\chi^{\boldsymbol{\gamma}}_{PP} \tag{S123}$$

And as both determinants are positive because of linear stability, this two things have the same sign. This calculation, together with Eq. (1), tells us that whenever the mean of the perturbed population is positive, then the sub-circuit without them will be unstable.

5.10 Response distribution to partial and non-homogeneous perturbations.

We now consider the case in which each population can not only received a perturbation that is partial, but this perturbation is different for each neuron mimicking disorder in the ChR2 expression. More specifically we need to recompute the expressions in equations (S76 , S77) for the case in which we have a perturbation vector $\delta h = \{\delta h_1, \delta h_2, \cdots, \delta h_n\}$ instead of having entries like $h_\eta = \{0,\cdots,0,\underbrace{1,\cdots,1}_{\gamma_\eta N_\eta},0,\cdots,0\}$, has entries given by $h_\eta = \{0,\cdots,0,\underbrace{D^\eta_1,\cdots,D^\eta_{\gamma_\eta N_\eta}}_{\gamma_\eta N_\eta},0,\cdots,0\}$, where $D^\eta_i \sim \mathscr{N}(d_\eta, g^2_\eta)$.

The optogenetic targeting matrix Σ, instead of being given by Eq. (S105), will be in this case:

$$\Sigma_{jk} = \sum_\eta\sum_{\eta'}\delta_{\alpha_j\eta}\,\delta_{\alpha_k\eta'}\,\delta_{k\in\mathscr{P}_{\alpha_k}}\,\delta_{j\in\mathscr{P}_{\alpha_j}}\,(D^{\alpha_i}_i)(D^{\alpha_j}_j) \tag{S124}$$

The expression for the perturbation to cell i will then be written as a mean given by the response that the network would have in the absence of disorder in the connectivity and a variance computed via Eqs (S76 , S77). Specifically:

$$\delta r_i = \sum_j R^0_{ij}\delta h_j + \Lambda_{\alpha_i}\xi_i \tag{S125}$$

$$= f'_{\alpha_i} D_i - f'_{\alpha_i}\sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\eta} d_\eta + \Lambda_{\alpha_i}\xi_i \tag{S126}$$

$$\tag{S127}$$

where $\xi_i \sim \mathcal{N}(0,1)$ and Λ_{α_i} is the generalization of Eq. (S106) to disordered perturbations, obtained by replacing Eq. (S124) into (S104) (we note that the only term that needs to be re-computed is the term M).

$$\Lambda^2_{\alpha_i} = \frac{\left(f'^2_{\alpha_i}\kappa_{\alpha_i} + \frac{1}{N}\left(\sum_\eta \frac{\kappa_\eta}{q_\eta} S^2_{\alpha_i\eta} - \frac{f'^2_{\alpha_i}\kappa_{\alpha_i}}{q_{\alpha_i}}\right)\right)\left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta\left((1-\gamma_\eta)d^2_\eta + g^2_\eta\right)q_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta} d_\eta \gamma_\eta\right)^2 q_{\eta'}\right)}{1 - \sum_\eta \nu_\eta\left(q_\eta f'^2_\eta \kappa_\eta + \frac{1}{N}\left(q_\eta \sum_{\eta'}\frac{\kappa_{\eta'}}{q_{\eta'}} S^2_{\eta\eta'} - f'^2_\eta \kappa_\eta\right)\right)} \tag{S128}$$

In the large N limit, this equation reduces to

$$\Lambda^2_{\alpha_i} = \frac{f'^2_{\alpha_i}\kappa_{\alpha_i}\left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta\left((1-\gamma_\eta)d^2_\eta + g^2_\eta\right)q_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta} d_\eta \gamma_\eta\right)^2 q_{\eta'}\right)}{1 - \sum_\eta \nu_\eta q_\eta f'^2_\eta \kappa_\eta} \tag{S129}$$

Which in the end means that the response of a neuron that belongs to the population α_i will respond to the optogenetic perturbation with a mean and a variance given by

$$\mu^{IN}_{\alpha_i} = f'_{\alpha_i} d_{\alpha_i} - f'_{\alpha_i}\sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\eta} d_\eta \tag{S130}$$

$$\mu^{OUT}_{\alpha_i} = -f'_{\alpha_i}\sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\eta} d_\eta \tag{S131}$$

$$\Lambda^{2,IN}_{\alpha_i} = f'^{\,2}_{\alpha_i} g^2_{\alpha_i} + \Lambda^2_{\alpha_i} \tag{S132}$$

$$\Lambda^{2,OUT}_{\alpha_i} = \Lambda^2_{\alpha_i} \tag{S133}$$

Analogously as before, we obtain a distribution of responses for the perturbed cells given by

$$\rho^{\mathrm{IN}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda^{\mathrm{IN}} a_{\alpha_i}} \exp\left\{-\frac{(x-\mu^{\mathrm{IN}}_{\alpha_i})^2}{2\Lambda^{\mathrm{IN}^2}_{\alpha_i}}\right\} \tag{S134}$$

$$\rho^{\mathrm{OUT}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda^{\mathrm{OUT}}_{\alpha_i}} \exp\left\{-\frac{(x-\mu^{\mathrm{OUT}}_{\alpha_i})^2}{2\Lambda^{\mathrm{OUT}^2}_{\alpha_i}}\right\} \tag{S135}$$

So the total distribution of responses is

$$\rho_{\alpha_i} = \gamma_{\alpha_i}\rho_{\alpha_i}^{\mathrm{IN}} + (1-\gamma_{\alpha_i})\rho_{\alpha_i}^{\mathrm{OUT}} \tag{S136}$$

5.11 Link to the low-dimensional system linear response

The activity of the low-dimensional system is equivalent to the mean of the non-disordered high-dimensional system. Perturbing all the neurons in a population α_j and then measuring the mean activity in the population α_i should be equivalent to computing the linear response in the low-dimensional system. To show this, we need to show that i) the measuring vector $\delta b = \frac{1}{N\alpha_i}\delta_{i\in\alpha_i}$ and δh is the optogenetic perturbation to all neurons in a given population, then

$$\chi_{\alpha_i,\alpha_j} = \frac{1}{N_{\alpha_i}}\sum_{i\in\alpha_i}\Big(\sum_{j\in\alpha_j} {R^0}_{ij}\Big) = U_{\alpha_i}^{\intercal}R^0V_{\alpha_j} \tag{S137}$$

Inserting (S81) into the above expression we obtain:

$$\chi = U^{\intercal}RV = U^{\intercal}(F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1}U^{\intercal}F)V \tag{S138}$$

$$= \mathbf{f}' - \mathbf{f}'\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\mathbf{f}' \tag{S139}$$

$$= \left(\mathbf{f}'^{-1} - \omega\right)^{-1} \tag{S140}$$

Which is the definition of χ as in Eq. (S7). We also need to show that the variance vanishes for large N. Writing $B_{ij} = \frac{1}{N_{\eta'}^2}\delta_{\alpha_i\alpha_j}\delta_{\alpha_i\eta'}$, and inserting it and Eq. (S98) into Eq. (S77) we find that :

$$\Lambda_{\eta\eta'} = \frac{\sigma^2}{N}\frac{\left(\sum\limits_{\alpha_i} S^2_{\alpha_i\eta'}N_{\alpha_i}\right)\left(\sum\limits_{\alpha_i}\frac{S^2_{\alpha_i\eta'}}{N_{\alpha_i}}\right)}{1-\frac{\sigma^2}{N}\left(\sum_{\alpha_i}N_{\alpha_i}{f'_{\alpha_i}}^2 + \sum_{\alpha_i,\alpha_j}S^2_{\alpha_i,\alpha_j}\frac{N_{\alpha_i}}{N_{\alpha_j}} - \sum_{\alpha_i}{f'_{\alpha_i}}^2\right)} \tag{S141}$$

This variance vanishes for large N, making the usage of the small circuit as a limit of the average behavior of the large one rigorous for linear networks.

Low-dimensional representation of the linear response when perturbing a fraction gamma

If we now do the average but instead of perturbing all cells in α_j, we compute the mean response over those that are perturbed, meaning $\gamma_{\alpha_j} * N_{\alpha_j}$.

We choose the matrices $\tilde{U}$ (like U above but instead of all ones for a population only has γ_{α_k}) and $\tilde{V}$ with columns given by vectors $\tilde{u}^{\alpha} = \frac{1}{N_\alpha}\delta_{i\in\mathscr{P}_\alpha}$ and $\tilde{v}^{\alpha} = \delta_{i\in\mathscr{P}_\alpha}$, meaning that $\tilde{u}^{(k)} = \frac{1}{N_k}(\underbrace{0,\dots,0}_{\sum_{l=1}^{k-1}N_l},\underbrace{1,\dots,1}_{\alpha_k N_k},\underbrace{0,\dots,0}_{N-\sum_{l=1}^{k}N_l})$ and similarly for $\tilde{v}$.

$$\chi^{\boldsymbol{\gamma}}_{\alpha_i,\alpha_j} = \frac{1}{\gamma_{\alpha_i}N_{\alpha_i}}\sum_{i\in\mathscr{P}_{\alpha_i}}\Big(\sum_{j\in\mathscr{P}_{\alpha_j}} {R^0}_{ij}\Big) = \frac{\tilde{U}_{\alpha_i}^{\intercal}}{\gamma_{\alpha_i}}R^0\tilde{V}_{\alpha_j} = \tilde{U}_{\alpha_i}^{\intercal}R^0\tilde{V}_{\alpha_j} \tag{S142}$$

Before we had $U^\intercal F V = \mathbf{f}'$. Now, we define $\boldsymbol{\gamma}$ which is a diagonal matrix with entries γ_{α_k} we have $\tilde{U}^\intercal F \tilde{V} = \mathbf{f}'$. Its worth noting that $\tilde{U}^\intercal F V = \mathbf{f}'$ but $U^\intercal F \tilde{V} = \boldsymbol{\gamma}\mathbf{f}'$. Using that

$$\chi^{\boldsymbol{\gamma}} = \tilde{U}^\intercal R \tilde{V} = \tilde{U}^\intercal (F - FV \left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^\intercal F)\tilde{V} \tag{S143}$$

$$= \mathbf{f}' - \mathbf{f}' \left(\mathbf{f}' - \omega^{-1}\right)^{-1} \mathbf{f}' \boldsymbol{\gamma} \tag{S144}$$

$$= \mathbf{f}' + \mathbf{f}' \omega \left(\mathbf{f}'^{-1} - \omega\right)^{-1} \boldsymbol{\gamma} \tag{S145}$$

$$= \mathbf{f}' + \mathbf{f}' \omega \chi \boldsymbol{\gamma} \tag{S146}$$

$$\tag{S147}$$

We notice that the *PP* element of this is

$$\chi^{\boldsymbol{\gamma}}_{\alpha_i\alpha_i} = f'_{\alpha_i} + f'_{\alpha_i}[\omega\chi]_{\alpha_i\alpha_i}\gamma_{\alpha_i} \tag{S148}$$

$$= f'_{\alpha_i} + f'_{\alpha_i}\left(\frac{\chi_{\alpha_i\alpha_i}}{f'_{\alpha_i}} - 1\right)\gamma_{\alpha_i} \tag{S149}$$

$$= f'_{\alpha_i} + \gamma_{\alpha_i}\chi_{\alpha_i\alpha_i} - f'_{\alpha_i}\gamma_{\alpha_i} \tag{S150}$$

$$\tag{S151}$$

where we again used that $[\omega\chi]_{\alpha_i\alpha_i} = \frac{\chi_{\alpha_i\alpha_i}}{f'_{\alpha_i}} - 1$ this is the exact same expression that Eq. (S91) as we wanted but I cannot bring this to any form that relates to what is below

Nomenclature

α_i	Short for population to which cell i belongs
χ	Linear response matrix of the low-dimensional circuit
Δ^α	Variance of the input to population α
κ and ν	low rank vectors that compose σ
Λ_α^2	Variance in the population α
ω	Low-dimensional connectivity matrix
Π_L	Diagonal matrix with entries κ
Π_R	Diagonal matrix with entries ν
Σ	Optogenetic targeting matrix
$\sigma^{\alpha\beta}$	matrix of the standard deviations of the weight matrix W
τ	Time constant
ξ	Power in a threshold power law input-output function
A	Diagonal matrix with factors to transform calcium to rates
B	Measuring matrix
c	Contrast value, usually normalized to 1
E	Error function
F	Diagonal matrix with the derivatives of f at the fixed point of the high-dimensional circuit
f	Input-output function /nonlinearity
f'	Derivative of f
h	External inputs to the network
J	Jacobian
k	Normalized entries of the low-dimensional linear response matrix χ
m^α	Mean firing rate in population α for high-D model
N	Number of neurons in the high-D system
n	Number of populations (different cell-types) in the network
N^α	Number of neurons in population α
P^α	Distribution of activity over population α
q^α	Fraction of cells in population α : N^α/N
R	Linear response of the high-D system
r	Activity, r^α is the activity in population α
R^0	Linear response of the high-D system in the absence of disorder
T	Diagonal matrix of time constants

u^{α}	Mean input to population α
v^{α}	Second moment of the activity distributions in population α
W	Weight matrix of the high-dimensional model
$w^{\alpha\beta}$	Mean connection strength form population β to population α
$w_{ij}^{\alpha\beta}$	Weight connecting neuron j in population β to neuron i in population α
W_0	matrix of entries $w^{\alpha\beta}$
z	Input current
$\mathbf{f}'$	Diagonal matrix with the derivatives of f at the fixed point of the low-dimensional circuit
HFP	Homogeneous fixed point
high-D	High-dimensional (i.e. N dimensional) model, with 4 populations
low-D	Low-dimensional (i.e. 4-dimensional) model

1. Hirofumi Ozeki, Ian M. Finn, Evan S. Schaffer, Kenneth D. Miller, and David Ferster. Inhibitory Stabilization of the Cortical Network Underlies Visual Surround Suppression. *Neuron*, 62(4):578–592, 2009.
2. Misha Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L Mcnaughton. Paradoxical effects of inhibitory interneurons. *J. Neurosci.*, 17(11):4382–4388, 1997.
3. Alexandra K. Moore, Aldis P. Weible, Timothy S. Balmer, Laurence O. Trussell, and Michael Wehr. Rapid Rebalancing of Excitation and Inhibition by Cortical Circuitry. *Neuron*, 97(6):1341–1355.e6, 2018.
4. Hiroyuki K. Kato, Samuel K. Asinof, and Jeffry S. Isaacson. Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron*, 95(2):412–423.e4, 2017.
5. Alessandro Sanzeni, Bradley Akitake, Hannah C. Goldbach, Caitlin E. Leedy, Nicolas Brunel, and Mark H. Histed. Inhibition stabilization is a widespread property of cortical networks. *Elife*, 9:1–39, 2020.
6. Hillel Adesnik. Synaptic Mechanisms of Feature Coding in the Visual Cortex of Awake Mice. *Neuron*, 95(5): 1147–1159.e4, 2017.
7. Yashar Ahmadian and Kenneth D. Miller. What is the dynamical regime of cerebral cortex? (March), 2019.
8. Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016.
9. Nathan R Wilson, Caroline a Runyan, Forea L Wang, and Mriganka Sur. Division and subtraction by distinct cortical inhibitory networks in vivo. *Nature*, 488(7411):343–8, aug 2012.
10. Bassam V. Atallah, William Bruns, Matteo Carandini, and Massimo Scanziani. Parvalbumin-Expressing Interneurons Linearly Transform Cortical Responses to Visual Stimuli. *Neuron*, 73(1):159–170, 2012.
11. Seung-Hee Lee, Alex C Kwan, Siyu Zhang, Victoria Phoumthipphavong, John G Flannery, Sotiris C Masmanidis, Hiroki Taniguchi, Z Josh Huang, Feng Zhang, Edward S Boyden, Karl Deisseroth, and Yang Dan. Activation of specific interneurons improves V1 feature selectivity and visual perception. *Nature*, 488(7411):379–83, aug 2012.
12. Aaron M. Kerlin, Mark L. Andermann, Vladimir K. Berezovskii, and R. Clay Reid. Broadly Tuned Response Properties of Diverse Inhibitory Neuron Subtypes in Mouse Visual Cortex. *Neuron*, 67(5):858–871, 2010.
13. Ulf Schnabel, Lisa Kirchberger, Enny van Beest, Sreedeep Mukherjee, Areg Barsegyan, Jeannette Lorteije, Chris van der Togt, Matthew Self, and Pieter Roelfsema. Feedforward and feedback processing during figure-ground perception in mice. *bioRxiv*, (October):456459, 2018.
14. Hillel Adesnik, William Bruns, Hiroki Taniguchi, Z. Josh Huang, and Massimo Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–230, 2012.
15. Julia Veit, Richard Hakim, Monika P. Jadi, Terrence J. Sejnowski, and Hillel Adesnik. Cortical gamma band synchronization through somatostatin interneurons. *Nat. Neurosci.*, 20(7):951–959, 2017.
16. Andreas J. Keller, Morgane M. Roth, and Massimo Scanziani. Feedback generates a second receptive field in neurons of the visual cortex. *Nature*, 582(7813):545–549, 2020.
17. Andreas Keller, Morgane Roth, Matthew Caudil, Mario Dipoppa, Kenneth Miller, and Massimo Scanziani. A Disinhibitory Circuit for Contextual Modulation in Primary Visual Cortex. pages 1–23, 2020.
18. Cristopher M. Niell and Michael P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, 65(4):472–479, feb 2010.
19. Asli Ayaz, Aman B. Saleem, Marieke L. Schölvinck, and Matteo Carandini. Locomotion controls spatial integration in mouse visual cortex. *Curr. Biol.*, 23(10):890–894, 2013.
20. Mario Dipoppa, Adam Ranson, Michael Krumin, Marius Pachitariu, Matteo Carandini, and Kenneth D. Harris. Vision and Locomotion Shape the Interactions between Neuron Types in Mouse Visual Cortex. *Neuron*, 98(3):602–615.e8, 2018.

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

22. Carsten K Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nat. Neurosci.*, 16(8):1068–76, 2013.

23. Mahesh M M. Karnani, Jesse Jackson, Inbal Ayzenshtat, Jason Tucciarone, Kasra Manoocheri, William G G. Snider, and Rafael Yuste. Cooperative Subnetworks of Molecularly Similar Interneurons in Mouse Neocortex. *Neuron*, 90(1): 86–100, 2016.

24. Allen Mouse Brain Connectivity Atlas. Available from https://portal.brain-map.org/explore/connectivity/synaptic-physiology. 2011-2020.

25. Sadra Sadeh, R. Angus Silver, Thomas D. Mrsic-Flogel, and Dylan Richard Muir. Assessing the Role of Inhibition in Stabilizing Neocortical Networks Requires Large-Scale Perturbation of the Inhibitory Population. *J. Neurosci.*, 37(49): 12050–12067, 2017.

26. Jonathan Kadmon and Haim Sompolinsky. Transition to chaos in random neuronal networks. *Phys. Rev. X*, 5(4):1–28, 2015.

27. Tanguy Cabana and Jonathan Touboul. Large Deviations, Dynamics and Phase Transitions in Large Stochastic and Disordered Neural Networks. *J. Stat. Phys.*, 153(2):211–269, 2013.

28. Diego Adrian Gutnisky, Jianing Yu, Samuel Andrew Hires, Minh Son To, Michael Ross Bale, Karel Svoboda, and David Golomb. *Mechanisms underlying a thalamocortical transformation during active tactile sensation*, volume 13. 2017.

29. Yazan N. Billeh, Binghuang Cai, Sergey L. Gratiy, Kael Dai, Ramakrishnan Iyer, Nathan W. Gouwens, Reza Abbasi-Asl, Xiaoxuan Jia, Joshua H. Siegle, Shawn R. Olsen, Christof Koch, Stefan Mihalas, and Anton Arkhipov. Systematic Integration of Structural and Functional Data into Multi-scale Models of Mouse Primary Visual Cortex. *Neuron*, 106(3): 388–403.e18, 2020.

30. Binghuang Cai, Yazan N. Billeh, Selmaan N Chettih, Christopher Harvey, Christof Koch, Anton Arkhipov, and Stefan Mihalas. Modeling robust and efficient coding in the mouse primary visual cortex using computational perturbations. *bioRxiv*, page 2020.04.21.051268, 2020.

31. Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *J. Neurophysiol.*, page jn.00732.2015, 2016.

32. Alexandre Mahrach, Guang Chen, Nuo Li, Carl van Vreeswijk, and David Hansel. Mechanisms underlying the response of mouse cortical networks to optogenetic manipulation. *Elife*, 9:1–37, 2020.

33. Hannah Bos, Anne-Marie Oswald, and Brent Doiron. Untangling stability and gain modulation in cortical circuits with multiple interneuron classes. *bioRxiv*, page 2020.06.15.148114, 2020.

34. Luis Carlos Garcia Del Molino, Guangyu Robert Yang, Jorge F. Mejias, and Xiao Jing Wang. Paradoxical response reversal of top- down modulation in cortical circuits with three interneuron types. *Elife*, 6:1–15, 2017.

35. A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk. On the Distribution of Firing Rates in Networks of Cortical Neurons. *J. Neurosci.*, 31(45):16217–16226, 2011.

36. Yashar Ahmadian, Francesco Fumarola, and Kenneth D. Miller. Properties of networks with partially structured and partially random connectivity. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 91(1):1–36, 2015.

37. Daniel B. Rubin, Stephen D. VanHooser, and Kenneth D. Miller. The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2015.

38. Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *J. Neurophysiol.*, 115(3):1399–1409, 2016.

39. Yashar Ahmadian, Daniel B. Rubin, and Kenneth D Miller. Analysis of the stabilized supralinear network. *Neural Comput.*, 25(8):1994–2037, 2013.

40. Christopher M. Bishop. *Pattern Recognition and Machine Learning*, volume 27. 2006.

41. Hyun Jae Pi, Balázs Hangya, Duda Kvitsiani, Joshua I. Sanders, Z. Josh Huang, and Adam Kepecs. Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477):521–524, 2013.

42. Leena E. Williams and Anthony Holtmaat. Higher-Order Thalamocortical Inputs Gate Synaptic Long-Term Potentiation via Disinhibition. *Neuron*, 101(1):91–102.e4, 2019.

43. Robert Rosenbaum and Brent Doiron. Balanced networks of spiking neurons with spatially dependent recurrent connections. *Phys. Rev. X*, 4(2):021039, may 2014.

44. Gabriel Daniel J. Millman, Gabriel Koch Ocker, Shiella Caldejon, India Kato, Josh D. Larkin, Eric Kenji Lee, Jennifer Luviano, Chelsea Nayan, Thuyanh V. Nguyen, Kat North, Sam Seid, Cassandra White, Jerome A. Lecoq, R. Clay Reid, Michael A. Buice, and Saskia E.J. de Vries. Title: VIP interneurons selectively enhance weak but behaviorally-relevant stimuli. Technical report, 2019.

45. Manuel Beiran, Alexis Dubreuil, Adrian Valente, Francesca Mastrogiuseppe, and Srdjan Ostojic. Shaping dynamics with multiple populations in low-rank recurrent networks. pages 1–29, 2020.

46. Sean Bittner, Agostina Palmigiano, Alex Piet, Chunyu Duan, Carlos Brody, Kenneth Miller, and John Cunningham. Interrogating theoretical models of neural computation with deep inference. 2019.

47. Pedro J. Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F. Podlaski, Sara A. Haddad, Tim P. Vogels, David S. Greenberg, and Jakob H. Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *bioRxiv*, page 838383, 2019.

48. Sen Song, Per Jesper Sjöström, Markus Reigl, Sacha Nelson, and Dmitri B. Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.*, 3(3):0507–0519, 2005.

49. Guillaume Hennequin, Yashar Ahmadian, Daniel B. Rubin, Máté Lengyel, and Kenneth D. Miller. The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron*, 98(4):846–860.e5, 2018.

50. Supratim Ray and John H R Maunsell. Differences in Gamma Frequencies across Visual Cortex Restrict Their Possible Use in Computation. *Neuron*, 67(5):885–896, sep 2010.

51. Aman B. Saleem, Anthony D. Lien, Michael Krumin, Bilal Haider, Miroslav Román Rosón, Asli Ayaz, Kimberly Reinhold, Laura Busse, Matteo Carandini, Kenneth D. Harris, and Matteo Carandini. Subcortical Source and Modulation of the Narrowband Gamma Oscillation in Mouse Visual Cortex. *Neuron*, 93(2):315–322, 2017.

52. Pablo Garcia-Junco-Clemente, Taruna Ikrar, Elaine Tring, Xiangmin Xu, Dario L. Ringach, and Joshua T. Trachtenberg. An inhibitory pull-push circuit in frontal cortex. *Nat. Neurosci.*, 20(3):389–392, 2017.

53. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2005.

54. Simon Renner, Nataliya Kraynyukova, Yannik Bauer, Gregory Born, Ann Hossam Kotkat, Xinyu Liu, Martin Spacek, Georgi Tushev, Tatjana Tchumatchenko, and Laura Busse. Inference of network connectivity from responses to briefly flashed gratings in mouse V1 using a stabilized supralinear network model (SSN). *Bernstein Conf. Abstr.*, 16(1982): 10–11, 2020.

55. Xavier Didelot, Richard G. Everitt, Adam M. Johansen, and Daniel J. Lawson. Likelihood-free estimation of model evidence. *Bayesian Anal.*, 6(1):49–76, 2011.

56. Donghui Chen and Robert J. Plemmons. Nonnegativity constraints in numerical analysis. *Birth Numer. Anal.*, pages 109–140, 2009.

57. Adil G Khan, Jasper Poort, Angus Chadwick, Antonin Blot, Maneesh Sahani, Thomas D. Mrsic-Flogel, and Sonja B Hofer. Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. *Nat. Neurosci.*, 21(6):851–859, jun 2018.