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

Predicting the response of the cortical microcircuit to perturbations is a prerequisite to determine the mechanisms that mediate its response to stimulus; yet, an encompassing perspective that describes the full ensemble of the network's response in models that accurately recapitulate recorded data 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 inferred parameters recover signatures of the connectivity structure found in mouse visual cortex. Analysis of this structure subsequently reveals 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. Finally applying the theory to inferring feedback to V1 during locomotion, we find that it is predominantly mediated by both SOM and VIP modulation.

Introduction

A defining feature of the operating regime of cortex is strong recurrent excitation that is stabilized and loosely balanced by recurrent inhibition [1–6]. This understanding was achieved through the discovery of a fundamental link between circuit stabilization and the response to specific perturbations, and was established in minimalistic recurrent network models with only two units, one describing the mean excitatory activity and another describing the mean activity of a single inhibitory type [1,7]. In these models, when recurrent excitation is sufficiently strong and stabilized by inhibition, an increase in the input drive to the inhibitory population elicits a simultaneous decrease of the excitatory and, *paradoxically*, of the inhibitory steady-state activity. This link provided a proxy to test inhibition stabilization in *in vivo* cortical circuits and an understanding of counter-intuitive responses to perturbations [1–4]. Nevertheless, and despite successful predictions, our understanding of the implications of the circuit's response to specific perturbations is still at its onset.

First, there is little consensus on how to generalize the fundamental link between stabilization and response to perturbations to the case of multiple inhibitory types [8,9]. 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) [10]. Importantly, these interneurons form a microcircuit characterized by a specific connectivity pattern [11–13], but how the stabilization of strong recurrent excitation is implemented by these interneurons and whether the structure in the synaptic connectivity in any way constrains the circuit's response to perturbations is not understood. Second, viral (but not transgenic) cell-type specific optogenetic perturbation is insufficient to elicit a paradoxical response [4,14], demonstrating that minimalistic models are insufficient to account for the response to concrete optogenetic manipulations and highlighting the need to advance the theoretical understanding of the circuit's response to perturbations in more detailed models of cortical activity in which cell, cell-type, and perturbation diversity play a role. Finally, if new models hope to account for this emerging complexity, they will be rife with parameter degeneracy. Yet, a data-driven framework designed to sub-select from the universe of such models has not been established. 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.

Here we developed a program for inferring high-dimensional cell-type-specific network models from data and a theoretical framework for the quantitative prediction of the circuit's response to patterned optogenetic perturbations. This framework allowed us to i) find a mechanism for network control based on hidden symmetries in the response matrix ii) link stability and response in high-dimensional multi-cell-type circuits, iii) predict an unexpected effect to partial perturbations and iv) infer which are the inputs that would induce changes in the network activity akin to those induced by behavioral modulations. Specifically, we analyzed calcium-imaging recordings of the activity of each interneuronal type in the visual cortex of the awake mouse, in response to stimuli of increasing contrast while the mouse was in a stationary condition. We identified, via a combination of fitting methods and theoretical tools [15–17] a family of mathematically tractable high-dimensional models that exactly describe the distribution of cell-type-specific calcium-imaging activity and its dependence on the stimulus contrast. Using recent results in random matrix theory [18], we defined an approximation that allowed us to obtain explicit expressions for the mean and variance of the distributions of responses to patterned optogenetic perturbations of the high-dimensional models. By linking the mean responses of these distributions to the response to perturbations in simpler, more minimalistic models and by evaluating these expressions with the parameters of the models fit we were able to make quantitative predictions. 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 [11,19]. By studying mathematically the implications of this structure for the response to population-wide 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[20] and theoretical[9] arguments, that PV interneurons play a major role in circuit stabilization. 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, opening a new avenue for experimental inquiry. Finally, we reveal the mechanism by which locomotion affects V1 by inferring the distribution of inputs that each cell-type population would need to receive for the network response to mimic the effect of locomotion.

Results

The analysis of low-dimensional (LD) models, in which there is one unit per population, revealed that the response to controlled perturbations could be interpreted to characterize the operating regime of cortex[1,7]. This was established in models that considered only two populations, excitatory and inhibitory. In these models, the circuit's response to perturbations is linked to its stability (see Eq. (S11)). When recurrent excitation is strong and stabilized by inhibition (an inhibition-stabilized network or ISN), an increase in the external input drive to the inhibitory population results in a *paradoxical* decrease of its steady-state activity. Conversely, a paradoxical response can only be observed in ISNs, and can therefore be utilized as a proxy to experimentally assess the stabilization properties of the cortical circuit[4].

Cortical circuits *in vivo* are composed of multiple inhibitory types and generate broad distributions of activity. In models that account for these features, the paradoxical response of a given inhibitory cell-type is not a predictor of the ISN condition[21] and its implications for circuit stabilization are not understood[8]. Here, we set out to establish a framework (Fig. 1) that enables quantitative, cell-type specific predictions of the response to perturbations in models that that incorporate the diversity of inhibitory cell-types and are high-dimensional (HD), meaning that there are many units per population that may be heterogeneous in connectivity and in other properties.

Mean-field theoretical approach to model high dimensional data

We study the response to visual stimuli of varying contrast in neurons of layer 2/3 of primary visual cortex (V1) of awake, head-fixed mice. Specifically, we study the responses of Pyramidal (E) cells and of Parvalbumin (PV), Somatostatin (SOM) and Vasoactive Intestinal Polypeptide (VIP) -expressing interneurons while the animal is shown square patches of drifting grating stimuli of a small size (5 degrees) at varying contrast.

To describe contrast modulations observed within each cell-type population, we build HD models with different proportions of cells in each population as measured experimentally[10]. To infer the model parameters, we begin by first inferring the parameters of a LD circuit with four units, each representing the mean activity of one cell-type population (Fig. 2**a**, fitting pipeline). Each unit has a power-law input-output function[22]. All four cell-types receive a baseline input to account for the spontaneous activity observed, while feed-forward inputs only target the E and PV populations and are taken to be a linear function of contrast. 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 starting with the measured mean response of each cell type at each contrast, and adding Gaussian noise to each of these data points, with mean zero and standard deviation given by the standard error of the given data point. We fit each LD model by finding the non-negative least squares (NNLS)[23,24] solution to each surrogate data set, and select from these, hundreds of data-compatible model parameters for which the network steady-states provide the best fit. Starting from these *seed* parameters, we search over the parameters of the HD model to find those HD models that match well the experimentally measured distributions of responses of all of the cell types (see below and Fig. 8). In HD models (Fig. 2**b**), each neuron has a

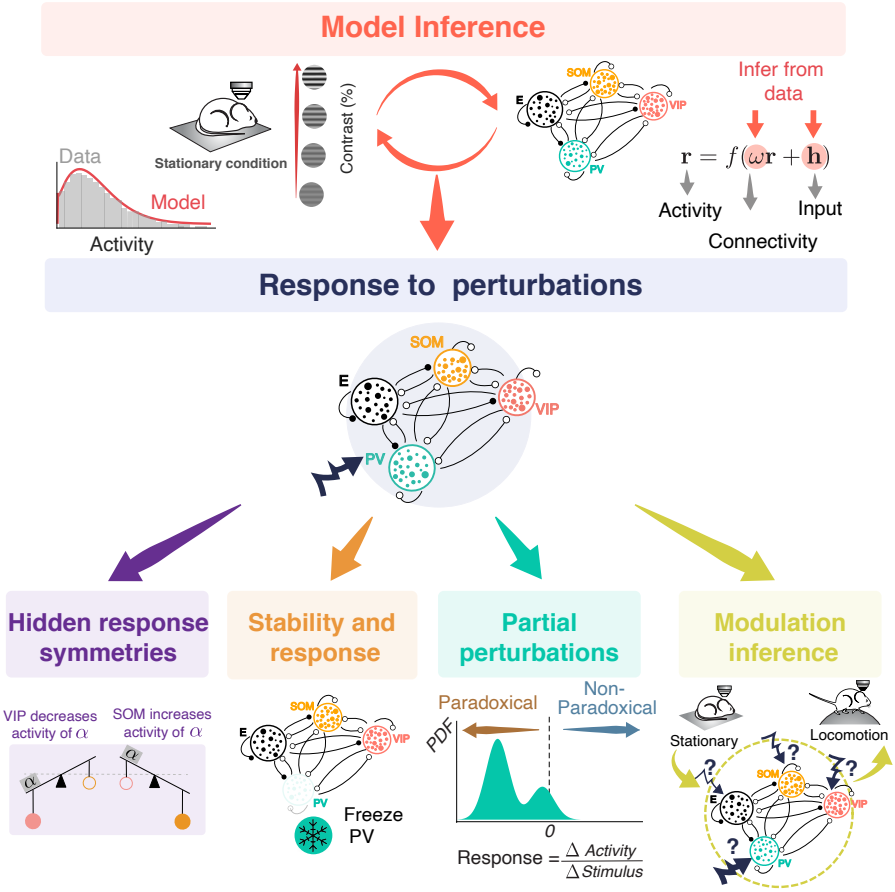

Figure 1 **Workflow.** Top: Model inference stage. In high-dimensional models with multiple cell types, the response of the circuit to perturbations is strongly dependent on parameters. In order to build models with predictive power, we fit the distribution of activity of each cell-type population (E, PV, SOM and VIP) to cell-type specific calcium imaging data of mouse visual cortex in response to stimuli of different contrasts, in a stationary condition (see also Fig. 2). Given a certain accuracy of the fit, we work with a family of data-constrained models. Middle: Response to perturbations stage. We developed a theoretical framework that allows us to derive explicit expressions for the mean and the variance of each cell-type population response to perturbations, under a suitable approximation. This approximation allows us to map the insights obtained in the perturbations analysis in LD models to HD models (see also Fig. 3). Bottom Left: Hidden response symmetries. We find hidden symmetries in the response to perturbations that lead to two mutually-exclusive mechanisms for network control via the manipulation of SOM and VIP activity (see also Fig. 4). Bottom middle left: Stability and response. Building the mapping between LD and HD models, we link the mean response to full-population perturbations with the stability of the network sub-circuit without the perturbed cell-type population, extending results of LD models with a single inhibitory type (see also Fig. 5). Bottom middle right: Partial perturbations. When the perturbations to the circuit are restricted to a subset of neurons, the responses to perturbations are bimodal. If a full population perturbation induces a paradoxical effect, we show that a partial perturbation exhibits a *fractional* paradoxical effect (see also Fig. 6). Bottom right: Modulation inference. Finally, we infer the perturbation pattern that would elicit a model response that matches the activity modulation induced by locomotion (see also Fig. 7).

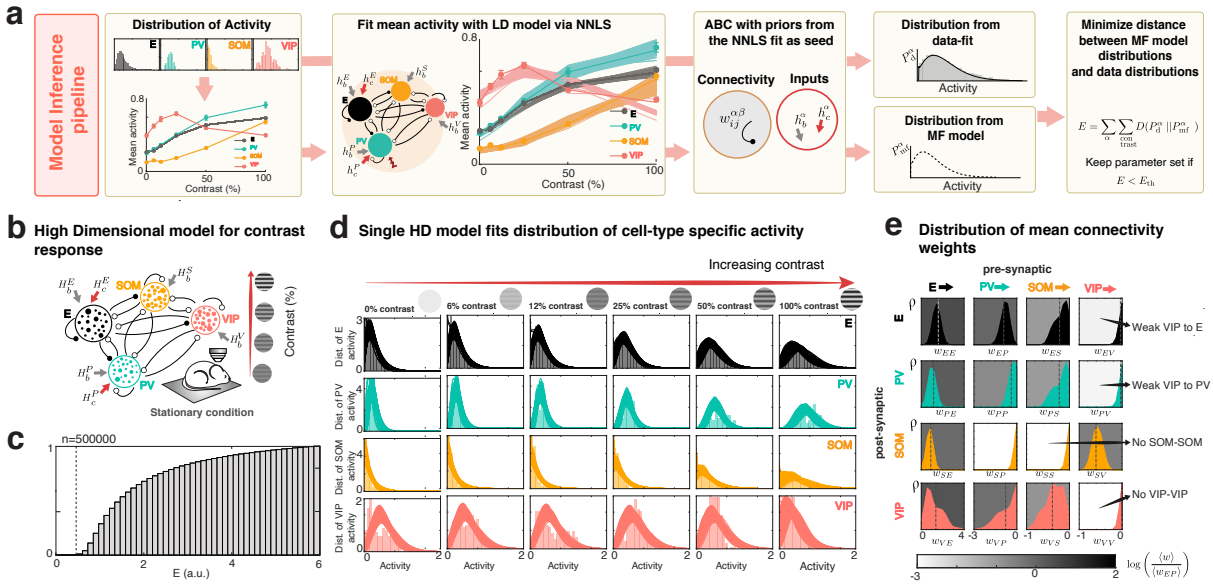

Figure 2 **Fitting the distribution of responses to multiple contrasts with a HD model.** **a)** Model inference pipeline. We firstly fit the mean activity of the pyramidal-cell (E, black), PV (turquoise), SOM (orange) and VIP (pink) populations, as measured with two photon calcium imaging (thick line, $\pm$ s.e.m.), as a function of stimulus contrast with a LD model of four populations. 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 feed-forward inputs from layer 4. Stimulus-related inputs are linear functions of the stimulus contrast. 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. We aim to find a family of HD models that recapitulate not only the means but the entire distributions of activity of all cell-types at all contrast values. Then, we use the inferred LD model parameters as a *seed* to build Gaussian priors for the connectivity mean ($w^{\alpha\beta}$) and the input means (H_b^α and H_c^α), whereas priors for the variance of the connectivity and the inputs are chosen arbitrarily. We generate HD models by sampling from those prior distributions of parameters, and compare the obtained model distributions to the fitted data distributions using an error function. This error, is given by the sum of the Kullback-Leibler divergences of the distributions given by the model ($P_{\rm mf}$) and the data ($P_{\rm c}$) for all cell-types and contrast values, which can be found explicitly. By only accepting models with error less than a threshold of 0.5 (top 0.005%), we build a family of suitable models. **b)** HD model has a distribution of external baseline inputs with mean H_b, and a stimulus related current, H_c, to E and PV, which is a linear function of the stimulus contrast. The variance of the input does not depend on contrast. The model has 34 parameters (24 that account for the 16 mean weights and the 8 low-rank weight variance, 4 mean baseline inputs, 2 mean stimulus-related inputs and 4 input variances, independent of contrast). For more details see Figure 8. **c)** Distribution of KL divergences, indicating the 0.5 threshold. We used models below this threshold for the analysis in the remaining text (see Methods for details). **d)** Example of a parameter configuration within the threshold. Data (histogram, colored bars) and data fits (solid colored line) are in good agreement. **e)** Distribution of mean connectivity weights over all possible models is shown. The gray-scale background of each panel is the logarithm of the mean of each distribution. Notice that, as in experiments (see Fig.9), the models lack recurrent SOM and VIP connections, and the connections from VIP to E and PV are small on average.

power-law input-output function [22,25] and receives heterogeneous baseline inputs and a stimulus-related inputs that have cell-type-specific means and variances (and the means of stimulus-related input depend on the stimulus). The connectivity is heterogeneous with a mean and a variance dependent on both the pre- and post-synaptic cell-type. This class of models reduces to the LD class whenever there is no heterogeneity in the connections or the inputs (homogeneous network).

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 system without heterogeneity. Consequently, it is not sufficient to use the parameters found for the LD model as the mean values of the heterogeneous connectivity and input distributions; the mean and variance of the connections and the inputs have to be found simultaneously for the HD model to fit the data. We expect the HD mean values to be near the LD values, so we focus our search for HD mean values on the vicinity of the LD values.

In order to find HD models, we build on two facts. First, given a power-law input-output function, there is a closed-form expression that maps the distribution of inputs that a given cell-type population receives to the distributions of activity that cell-type population produces [17] (see Eq. S44). Given that this expression is explicit, it allows us to infer, from the distributions of activity for each cell-type and each stimulus contrast, the distributions of inputs to each population. Second, given a HD circuit model (for a fixed set of parameters), the distributions of inputs and activities it will produce can be computed self-consistently through mean-field theory [15,16] (see Eq. S40). These two facts, taken together, allowed us to obtain an explicit error function that quantifies how different the measured distributions of activity of all cell types at all contrast are from those distributions produced by the mean-field equations with a given set of parameters. To generate candidate models, we sample from prior distributions on the parameters. These prior distributions are Gaussian distributions for the mean and variances of the weights and the external inputs. The priors for the means are centered on the LD *seed* parameters. We keep the solutions that have a sufficiently small error, to define a family of HD models that fit the data (Fig. 2**c,d**). This family of models recapitulates the dependence of the distribution of responses of all cell types on contrast, and captures both the spreading out of the distributions with increasing contrast and the heavy tails of the distributions seen in calcium data.

We require that the recurrent excitation is strong, and that the LD system has a paradoxical response in the PV population in the absence of visual stimulation (i.e. at zero contrast), as measured experimentally [4]. Beyond that, this optimization takes as sole input the response data and uses no other prior information on the synaptic structure, hence it is not obvious that any meaningful synaptic structure should be recoverable from such a procedure. Surprisingly, the structure of the inferred connectivity matrices has a striking resemblance to that reported experimentally (Fig. 2**e**), see also 9**b**). In particular, recurrent connections within the SOM population and the VIP population were absent in most models, as observed in mouse V1 [11,13,19] (Fig. 9); and,whenever inputs were chosen to target only E and PV, VIP interneurons had weak or absent connections to all other cell-types except SOM interneurons, also as reported in mouse V1 [11,13,19].

Analytical approach to full and partial perturbations

To develop a theoretical framework for using optogenetic perturbations to probe the circuit, we compute the distribution of responses of the network to perturbations, *e.g.* optogenetic activation or suppression of sets of cells. For each pair of cells, the change in steady-state response of the cell i (belonging to a population α) per small change in the input to a cell j (belonging to a population β) will be given by the element $\chi_{ij}^{\alpha\beta}$ of the response matrix χ. We developed a theoretical framework that allows analytic computation of the mean and the variance of the response over each population to (small) perturbations, under the following approximation. We assume that the gain of the neurons in a given population is the same for each cell (equal to the gain of the homogeneous system). Our system then satisfies the assumptions needed to build on recent work on random matrix theory [18] to compute these response distributions. Under this approximation, which we refer to as the *homogeneous fixed point*

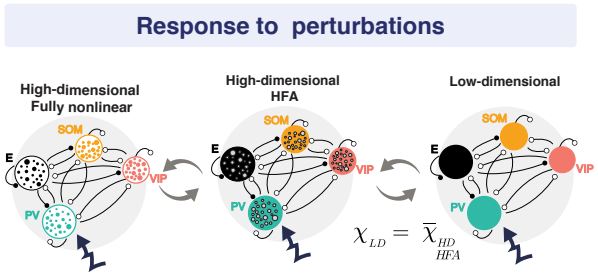

Figure 3 Analytical framework for the study of perturbations. In order to study perturbations in HD models (top left), in which the heterogeneity in the connections and the inputs induces heterogeneity in the response gain of each neuron, we make an approximation. By linearizing around the homogeneous fixed point (i.e. linearizing around the fixed point of the network without heterogeneity, middle panel), we are able to leverage results from random matrix theory to obtain explicit expressions for the mean and the variance of the response of each cell-type population. The analytical approach reveals that, when perturbing all cells in a given population, the mean response of an HD system in the HFA is equal to the response of the LD system (right panel, which is equivalent to the system without heterogeneity).

approximation (hereinafter HFA, see Eq. S54), we are able to obtain analytical expressions for the behavior of the mean and the variance of the distributions of optogenetic responses in each population to either full or partial, and either homogeneous or heterogeneous, perturbations. Importantly, we find that under the HFA, neurons belonging to a specific cell-type's population of the HD heterogeneous system, have a mean response to cell-type-specific perturbations given by the response of the homogeneous system without heterogeneity, equivalent to the LD system (Fig. 3, see also Eq. S121), allowing us to directly link the response of the LD and HD models. In the following, we will distinguish analytics using the HFA, from simulations of the fully nonlinear system, in which different cells of a given cell-type can have different gains at the network's fixed point.

Symmetry principles of optogenetic response

Figure 4 shows the first application of the link between LD and HD systems offered by the HFA. When computing the response distributions to perturbations, we find consistent symmetries in the responses to perturbation of the SOM population vs. a perturbation of the VIP population. In order to understand this, based on our recovery of the structure of the connectivity matrix found in mouse V1 (Fig. 2), we examined the linear response matrix of LD circuits (Eq. S7) for a generic connectivity that satisfies the condition that VIP projects only to SOM, but is otherwise arbitrary. We found that in this case, the linear response matrix has a symmetry between the response of E, PV, SOM and VIP 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 four cell types (Fig. 4**a**), see also Eq. S13). In the case of VIP, there will be an additional shift given by its own gain. 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

$$\chi_{LD}^{\alpha V} = -f'_V \omega_{SV} \chi_{LD}^{\alpha S} + \delta_{\alpha V} f'_V \qquad \alpha = \{E, P, S, V\} \tag{1}$$

We refer to these equalities as *Hidden Response Symmetries (HRS)* (Fig. 4**a-c**). Because the mean response of a population to the perturbation of all neurons in another population under the HFA is given by the response of the LD circuit (see Figure 2), these symmetries also apply to the mean of the distributions in the high dimensional system under the HFA. Figure 4**b** shows the response distributions of an example model from Figure 2, to perturbations of the entire SOM and VIP populations. The distributions obtained under the approximation (colored lines) are in good agreement with the results of simulations of the fully non-linear system (green). 4**c** quantifies to which extent

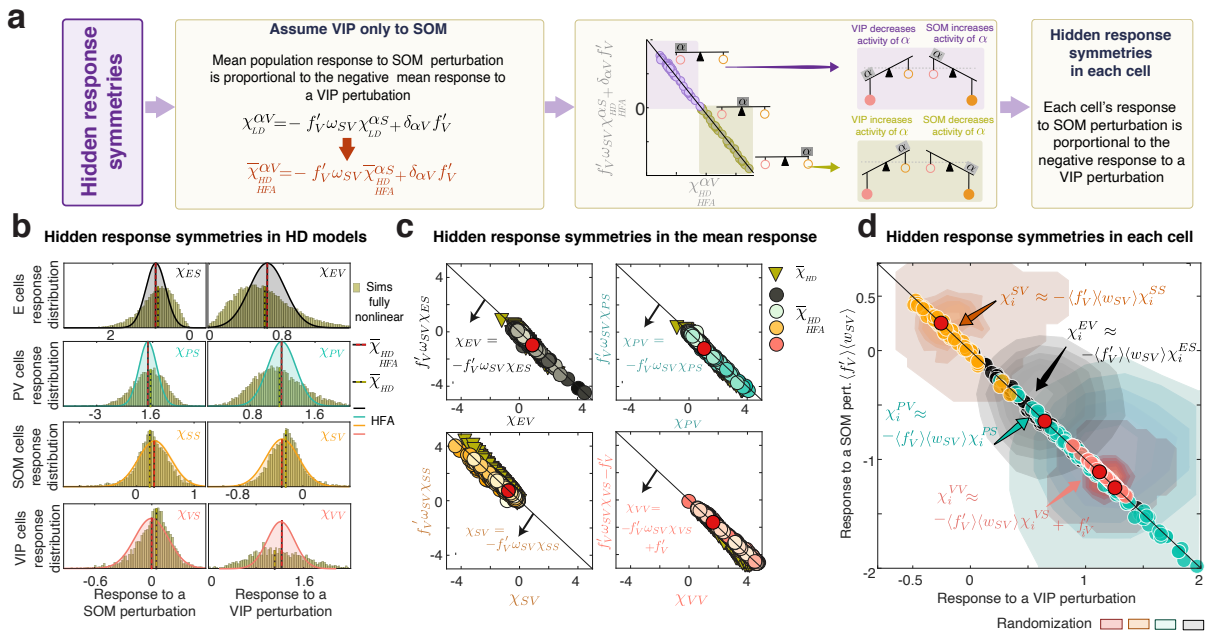

Figure 4 Hidden response symmetries (HRS). a) A parameter-independent relation holds true between the response of pyramidal cells (E), PV interneurons (P), SOM interneurons (S), and VIP interneurons (V) to perturbation of SOM and VIP. These relations or hidden response symmetries (HRS), described by Eq. (1), are derived for LD models under the assumption of VIP projecting only to SOM, and hold for the mean response of HD models under the HFA. The illustration depicts the response of cell-type α to VIP perturbation vs the response to a SOM perturbation multiplied by the coefficient in Eq. (1). Given a perturbation to the VIP population, the constraints imposed by the HRS define the sign and 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, green) and another one in which the opposite is true (upper right, purple). We hypothesize that this relation could approximately hold at the single cell level. **b)** Distribution of responses to full-population SOM (left) and VIP (right) perturbations, for maximum stimulus contrast. Green histograms are the result of the simulation of a fully nonlinear HD system, while colored histograms and corresponding lines are the analytical result, only possible under the HFA. **c)** Opposite and proportional responses to perturbations of SOM and VIP, for E (top left), PV (top right), SOM (bottom left) and VIP (bottom right), for the top 360 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. 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. **d)** Hidden response symmetries at the single neuron level. Responses of single cells of type E (black), PV (turquoise) and SOM (orange) and VIP (pink) cells to a VIP perturbation, vs. the responses of those same cells to a SOM perturbation multiplied by the factor: $\langle f'_V \rangle \langle w_{SV} \rangle$. The response symmetries hold at the single cell level in the fully nonlinear HD 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 perfectly correlated, but the two perturbations still elicit responses with opposite sign.

the HRS hold in the mean response of HD models that fit the data, both in models under the HFA and the fully nonlinear networks. As the data-compatible models naturally exhibit only weak connections from VIP to other interneurons besides SOM, this symmetry in the mean response is revealed in this family of models.

The HRS formalize a clear intuition: Because VIP neurons only project to SOM neurons, a weak perturbation to VIP will only affect the rest of the circuit through SOM, relaying that perturbation with an opposite sign.

The HRS defines two alternative regimes of network configuration: 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. VIP will be inhibitory if the disinhibitory effect of SOM cells on PV cells outweighs the direct inhibitory effect of SOM cells on E cells; otherwise, it is disinhibitory. In our data-compliant models, activation of VIP has a disinhibitory effect on E, as in experiments[26–29], 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 versus SOM perturbations, with the same proportionality constant for all, are conclusive predictions resulting from our our analysis.

Finally we asked to which extent the mathematical understandings offered by the *Hidden Response Symmetries* hold at the single cell level. We reasoned that the effect of the perturbation that each cell receives, will respect the HRS but now with the average values of the connectivity and the gains. Indeed, Figure 4**d** shows, for a single example fully nonlinear network, the response of each cell to a perturbation to the full SOM population vs the response to a perturbation to the full VIP population with the appropriate corrections. These responses are perfectly anticorrelated.

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

We next investigated the relation between paradoxical response of an inhibitory cell-type and the stability of the network sub-circuits. A multi-cell-type circuit is an inhibition-stabilized network (ISN) if and only if an increase in the input drive to any or all of the inhibitory populations paradoxically results, in the new steady state, in a change in the same direction – both increasing, or both decreasing – in both the inhibitory input to the excitatory population and of the excitatory activity.[25,30]. 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, thereby guaranteeing that the net inhibition received by excitatory cells decreases, and also decreases the excitatory activity, then the circuit is 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 decrease its activity.

To systematically investigate the response of each cell-type to its own stimulation, we start by focusing on the diagonal of the LD linear response matrix (see Eq. S10) $\chi_{\alpha\alpha}$, found by linearizing the dynamics in the vicinity of some stable fixed point of activity. These elements, can be written as a function of the Jacobian J of the entire circuit (which drives the linearized dynamics) and the Jacobian J_α of the sub-circuit without cell-type α:

$$\chi_{\alpha\alpha} \propto \frac{1}{\det(-J)} \det(-J_\alpha) \qquad \alpha = \{E,P,S,V\} \tag{2}$$

At a stable fixed point, the determinant of the negative Jacobian is positive (because all eigenvalues of the Jacobian have negative real part). As a result, $\det(-J) > 0$, so $\chi_{\alpha\alpha}$ has the same sign as $\det(-J_\alpha)$. Thus, if the response of cell-type α at a given fixed point is paradoxical ($\chi_{\alpha\alpha} < 0$), then the sub-circuit without that cell-type is unstable ($\det(-J_\alpha) < 0$, see Eq. (S10)). 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, and links cell-type-specific paradoxical response to sub-circuit stability in a more general setting (Fig. 5**a**)). In particular, we furthermore find that when VIP projects only to SOM, 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).

To link the LD insights to the HD models, we notice that if the connectivity is dominated by its random component (see Eq. S33), the eigenvalues of the Jacobian of the HFA will follow a circular law, except for a set of outliers corresponding to the eigenvalues of the LD system (as proven in[31] for the case of an i.i.d. random matrix, see also

Methods; this seems to well describe our results, but a more precise treatment of our case, in which the variances of different cell types are different, is in [18]). Therefore, in the HD systems, whenever the mean of the LD system is paradoxical, then for sufficiently large variance of the connectivity, the system without that population will, under the HFA, retain the unstable eigenvalue of the LD system and thus be unstable (Fig. 5**a**)). This phenomenon is illustrated for an example model at 100 % contrast in Figure 5**b-c**. Notice that the mean response is only paradoxical for PV cells, and that therefore the eigenvalue distribution of the system without PV, has a positive outlier (top left panel). For comparison, simulations of the fully nonlinear system are also shown. Although there is no theoretical guarantee that the outlier eigenvalues of the fully nonlinear system will be organized as the ones in the HFA, we observe good agreement.

To fit the HD models, we required that the LD *seed* used for the model priors (see Fig. 2 and Methods) had a paradoxical response in PV in the absence of visual stimulus, to match experiment [4], but we did not apply any constraints to the response of the HD system, which therefore could have lacked a paradoxical response in PV. Nevertheless we observe that the mean response of PV to its own stimulation is paradoxical in almost all HD models that fit the data (Fig. 5**e-f**)), and that the outlier eigenvalue of the sub-circuit without PV is positive, suggesting a fundamental role of PV in circuit stabilization in our family of models (Fig. 5**e-f**), top left panel). Furthermore, we find that no other interneuron has a mean paradoxical response, and that the real parts of the eigenvalues of the sub-circuits without them are always negative (Fig. 5**e-f**)).

In summary, and consistent with previous work showing that strong perturbations to PV destabilize the dynamics in V1 [20], 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. 4**b,c**).

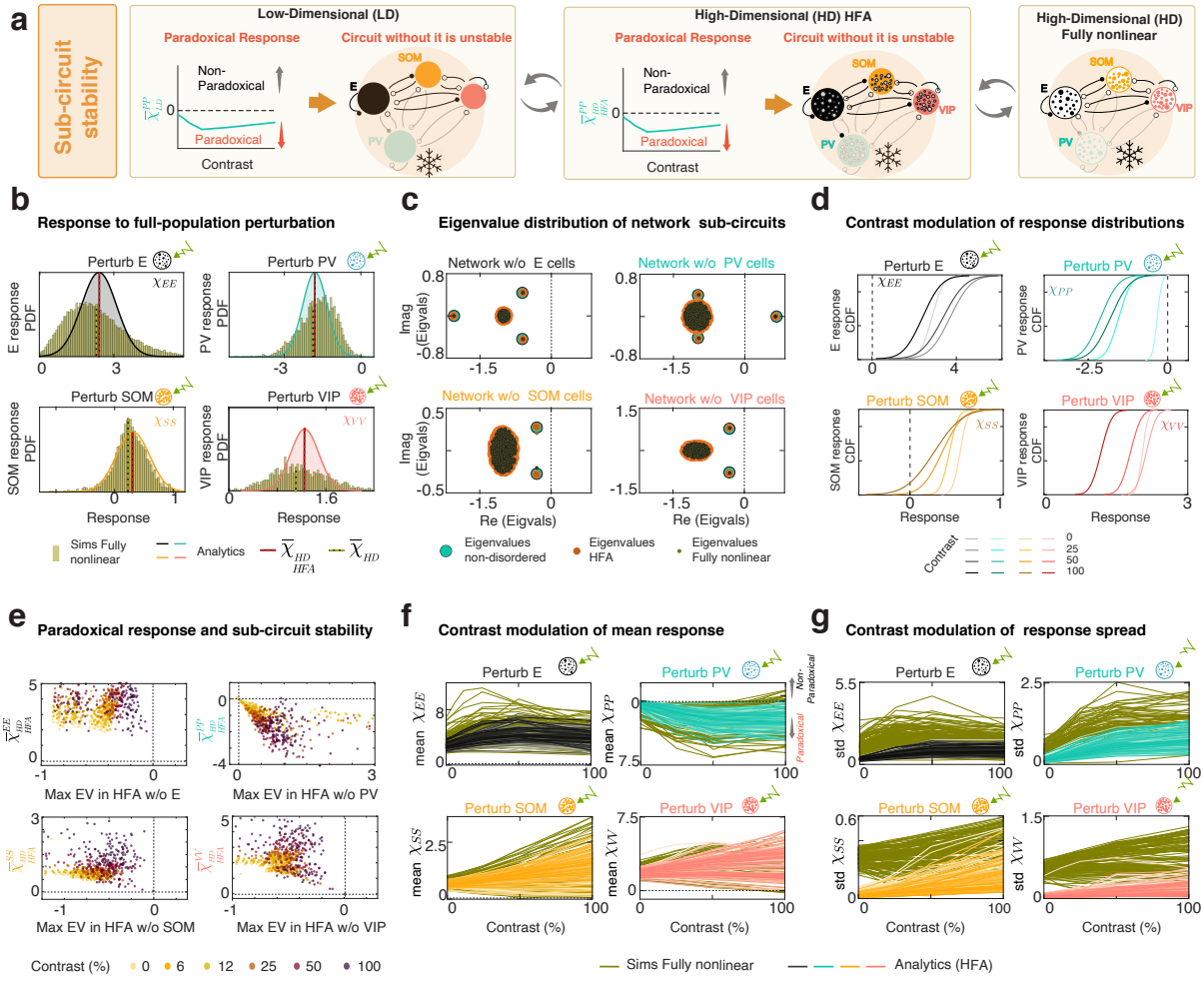

Figure 5 **Paradoxical response and circuit stabilization in data compatible models. a)** Graphic summary of the relation between stability and paradoxical responses. The response of a cell-type α in the LD case, which is the change in activity normalized by the size of the perturbation, is shown as a function of contrast. When the response of the cell-type which is being perturbed is negative, the response is paradoxical. A paradoxical response of a cell-type in an LD model in turn implies that the circuit without that cell-type is unstable (see Eq. 2). This relation holds in the HD system under the HFA if the variance of the weight distribution is sufficiently small. **b)** Distribution of responses to full-population perturbations for a stimulus of 100% contrast. Green histograms are the result of the simulation of a fully HD nonlinear system (dashed green line is its mean), while colored histograms and colored lines are the analytical curves obtained under the HFP approximation (dashed red line is its mean, corresponding to the LD system response). The responses of E, SOM and VIP cells are not paradoxical, while all cells in the PV population respond paradoxically to PV stimulation. **c)** Eigenvalue distribution of the Jacobian of the sub-system without the E (top left), without the PV (top right), without the SOM (bottom left) and without the VIP (bottom right) populations. As the outliers of the eigenvalue spectrum of the Jacobian under the HFA are defined by the LD system for sufficiently small variance of the weight distribution, and because $\bar{\chi}^{\alpha\beta}_{HFA} = \chi^{\alpha\beta}_{LD}$, a mean negative response in the HFA approximation indicated that the sub-circuit without that population is unstable. In the special case in which VIP projects only to SOM, the lack of a paradoxical response in SOM indicated that the E-PV circuit is stable. **d)** Cumulative distribution of responses (HFA) to full-population perturbation in the presence of a visual stimulus for varying stimulus contrast. **e)** Mean response of each cell type to a perturbation to that same cell type, vs the real part of the maximum eigenvalue of the sub-circuit without that cell type, for all values of the contrast. **f)** Mean response of each cell type to a perturbation to that same cell type as a function of contrast, for models in the HFA (E in black, PV in turquoise, SOM in orange, VIP in pink) and for the fully nonlinear network (green). **g)** Same as f) for the standard deviation of the response. 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.

Fractional paradoxical effect

Optogenetic perturbations of cortical circuits do not affect all cells equally. 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. Optogenetic activation in this case will result in a *partial* perturbation. Within the perturbed population, diversity in the opsin expression affects the responsiveness of each cell to light differently and introduces another source of heterogeneity, which we model as a *heterogeneous* perturbation.

Figure 6**a)** shows the distribution of PV responses to perturbing 25%, 75% and 100% of the PV population. We find mathematically, that under the HFA, the distribution of responses of the entire population is bi-modal, given by a mixture of Gaussians (turquoise) composed of a Gaussian distribution corresponding to the perturbed cells (red dashed line) and another one corresponding to the unperturbed population (green dashed line, see Eq. S94). The distributions of responses under the HFA are in good agreement with simulations of the fully nonlinear system (gray). When the number of perturbed PV cells is small, the mean of the Gaussian response distribution corresponding to the perturbed cells is positive (see also Fig. 6**b)**) and all the eigenvalues of the Jacobian of the sub-circuit without those perturbed cells have negative real part (Fig. 6**a)**, bottom left). As the fraction of perturbed PV cells increases, the mean response of the perturbed population moves towards negative values, ultimately changing sign, as does the maximum eigenvalue of the sub-circuit without the perturbed cells (Fig. 6**a)**, bottom right). The negative movement of the responses of the perturbed population mean gives rise to a curious phenomenon: with increasing fraction of PV cells perturbed, the fraction of PV cells responding negatively (paradoxically) can show non-monotonic behavior (Fig. 6**b)**, top right). Over some range, increasing the fraction of stimulated PV cells decreases the probability that we will measure a PV cell showing negative response, because it adds more cells to the perturbed population, which still shows positive responses. With further increase in the fraction perturbed, the responses of an increasing fraction of the perturbed population become negative, ultimately increasing the probability that a PV cell has a negative response. When 100% of PV cells are stimulated, all show a negative response. We name this the *fractional paradoxical effect*. This result extends the concept of critical fraction developed in Ref.[32] to the case in which the neurons have heterogeneous connectivity.

The lower panels of Figure 6**b)** show the dependence of the fraction of PV negative responses on the fraction of perturbed PV cells for different values of the stimulus contrast in the models obtained in Figure 2. Intriguingly, in the models that fit the data, PV has a fractional paradoxical response at all contrasts. Recent experiments (Ref.[4]) have revealed that an optogenetic perturbation of PV interneurons with transgenic opsin expression (affecting essentially all PV cells) elicits a paradoxical effect in most cells, whereas if the expression is viral (and therefore affecting only a fraction of PV cells), a much smaller portion (about 50%) of cells show negative responses. Our models are consistent with that observation, and predict that that property is independent of the stimulus contrast.

To understand the relationship between fractional paradoxical response and stability, we built a LD, 5-dimensional (5D) network (Fig. 6**c)**, top right), with two PV populations, a perturbed one (red) and an unperturbed one (green). The connectivity of this network is chosen such that its response to perturbations is mathematically equivalent to the mean population response to a partial perturbation in the HD system under the HFA. As predicted by Eq. 2, whenever the response of the perturbed PV population in the 5D system becomes negative (paradoxical), the sub-system composed of all of the non-perturbed populations loses stability.

On the one hand, this tailored 5D network links response to partial perturbations in a high dimensional system with response to perturbations in a LD system. On the other hand, by similar arguments than those given in Figure 5, the eigenvalues of the Jacobian of the non-perturbed HFA HD system will have outlier eigenvalues close to those given by the non-perturbed populations in the 5D system. These two facts taken together, imply that when the mean response of the perturbed PV population in the HFA becomes negative, the sub-system without those perturbed neurons will become unstable. The top panel of figure 6**d)**, illustrates this fact by showing the mean response of the perturbed PV population (both in the HFA and the fully nonlinear system) as a function of the outlier eigenvalue for different fractions of perturbed PV neurons. Also shown is the response of the equivalent

5D system to a perturbation of the red PV population. For this example, perturbing more than 60% of the PV population will make the circuit without the perturbed population unstable. 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. We observe that, at high contrast, there are networks for which the sub-system loses stability but for which the mean perturbed population does not change sign. The link between perturbation and stability is not bi-directional; the system can lose stability without changing the sign of the determinant of the Jacobian (see [21] for a full clarification).

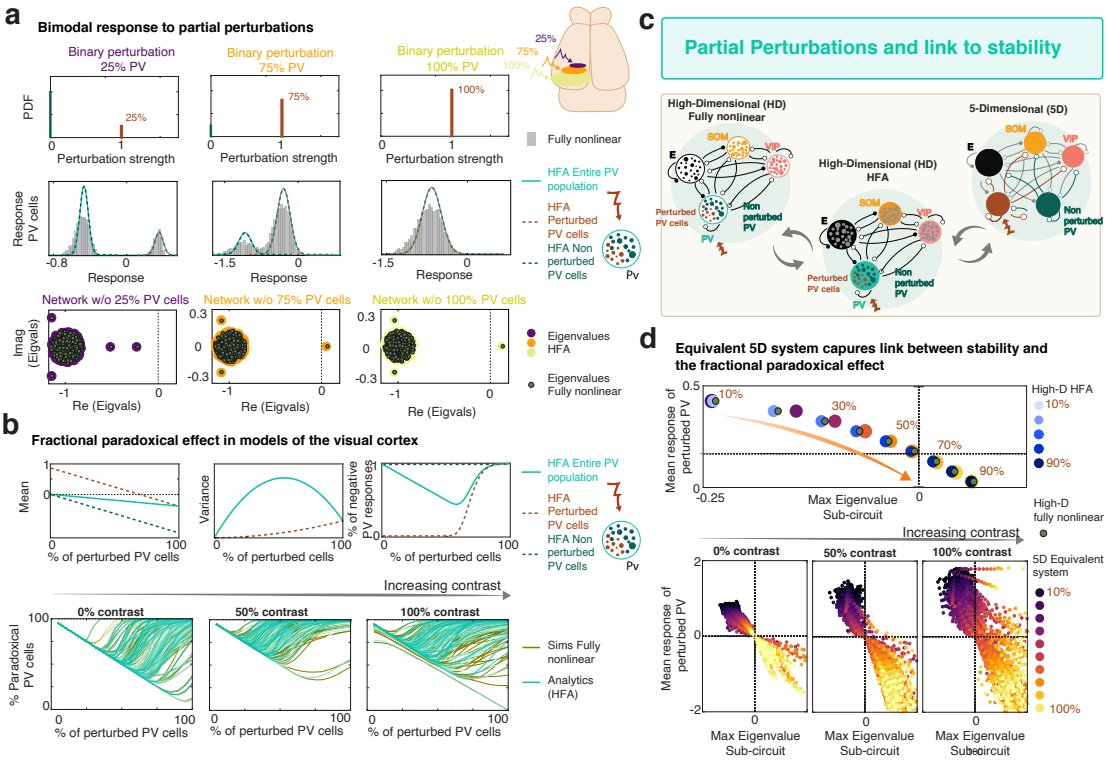

Figure 6 **Fractional paradoxical effect and link to sub-circuit stability: a)** Top: distribution of perturbation strengths, when perturbing 25% (left), 75% (middle) and a 100% (right) of the PV population. This can be understood as having an increasingly larger radius of an optogenetic stimulus, as indicated in the top right scheme of a mouse brain . Middle: Partial perturbations result in a bimodal distribution of responses in the HFA, given by a mixture of two Gaussians (turquoise). The rightmost peak (dashed red) corresponds to the response of the sub-population of stimulated PV cells, while the leftmost peak (dashed green) corresponds to the response of non-stimulated PV cells. The distribution of responses is the HFA is in good agreement with simulations of the fully nonlinear system (gray histograms), for this example model at lowest contrast. Note that the mean response of the perturbed population changes sign with increasing number of perturbed PV cells. Bottom: Eigenvalue spectrum of the Jacobian of the non-perturbed sub-circuit for the HFA approximation (purple, orange and yellow) and the fully nonlinear system (green). The maximum eigenvalue of the network subsystem changes sign with increasing number of perturbed PV cells. **b)** Top left: Mean of the entire (bimodal) distribution of PV cell responses (turquoise), the mean of the perturbed PV cell responses (dashed red) and the non-perturbed PV cell (dashed green) responses as a function of the fraction of PV cells perturbed. Top middle: While all three means monotonically decrease with the fraction of stimulated cells, the variance of both perturbed (dashed red) and non-perturbed (dashed green) monotonically increase, resulting in a non-monotonic variance of the full distribution. Top right: 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. Bottom: The fractional paradoxical effect is a signature of models that fit the data, and occurs for all values of the contrast. Simulations of the fully nonlinear system (green) are in good agreement with calculations from the HFA (turquoise). **c)** Linking response and stability across models. A fully nonlinear system can be linked to a HD system of lower complexity via the HFA. The mean response to a partial perturbation in the HFA can be mapped to the response of a PV sub-population in a LD system with two PV populations, a perturbed one (red) and an unperturbed one (green) see, Eq. (S106). **d)** Top: Mean response of the perturbed PV population as a function of the value of the outlier eigenvalue for different fractions of perturbed PV cells for the fully nonlinear system (green) HFA (blue colors) and the equivalent 5D system (purple orange palette). The mean responses become negative when the maximum eigenvalue crosses zero, indicating instability of the non-perturbed sub-circuit. Bottom: Mean response of the perturbed PV population as a function of the value of the outlier eigenvalue in the equivalent 5D model obtained from different models that fit the data, for different values of the contrast.

Inferring circuit modulations

We derived explicit expressions for the mean and the variance of the response to heterogeneous perturbations, in which each cell is perturbed differently (see Eq. S119). This expression, which implicitly depends on the contrast via the population's gain, allows us to mathematically map the parameters (mean and variance) of the perturbations to the mean and variance of the response distributions, under Gaussian assumption, which can be measured experimentally. We then asked, if we assume that locomotion is an heterogeneous perturbation that affects each cell-type population differently, can we infer the nature of this perturbation from data? In order to do so, we computed the difference between each cell's activity in the locomotion and the stationary condition (Fig. 7 **a**), and found the best Gaussian fit for each case (dashed line). Next, we used the derived expressions to fit the distributions of locomotion modulations, and infer the perturbations which would result in activity changes that mimic the effect of locomotion (Fig. 7 **b**). Specifically, assuming that the effect of locomotion was a cell-type specific, Gaussian-distributed perturbation whose mean and variance depends linearly on the stimulus contrast, we fit the mean and variance of locomotion-induced modulations with the explicit expressions (Fig. 7 **c**, left panels). This fit allowed us to infer, for each model in the family of models that fit the stationary data (see Fig. 2) which is the cell-type specific mean and variance of the inputs that would mimic the effect that locomotion has in the activity (Fig. 7 **c**, right panels).

We found that, consistent with previous findings [28], in the absence of visual stimulation, the mean change in activity is only significantly positive for VIP and PV cells (stars in Fig. 7 **a** and **c**) whereas that of E and SOM cells is not. Interestingly, we find that the perturbations that would account for the observed locomotion effects have a large mean and variance in VIP and surprisingly, also in SOM, but less so in E and PV. This method allows us to infer modulations to the population's activity that are not apparent form the data and that would be unattainable without explicit expressions.

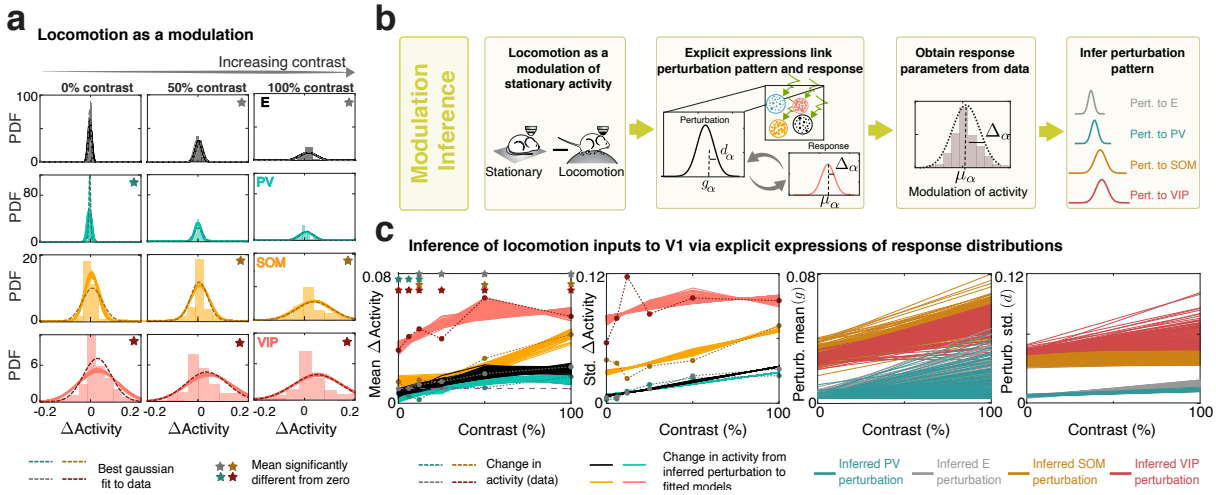

Figure 7 **Inferring inputs that lead to activity modulations that mimic the effect of locomotion.** **a)** Distribution of Δ activity (difference between each cell's activity in the locomotion and the stationary condition) for each cell type and different values of the stimulus contrast. Stars in the top right corner indicate when mean is significantly different from zero (p<0.0001, t-test). Dashed lines indicate best Gaussian fits. Solid lines are fits from the explicit expressions (see Eqs. S113 and S111) **b)** Scheme of how to infer the cell-type specific perturbations (Gaussians of mean g_α and standard deviation g_α for $\alpha = \{\text{E,PV,SOM,VIP}\}$) that give rise to the distribution of Δ activity (with mean μ_α and standard deviation Δ_α). By fitting these last expressions to data, the inputs can be inferred. **c)** Mean of Δ activity and standard deviation of Δ activity as a function of contrast. Dashed lines are the data (E in black, PV in blue, SOM in orange and VIP in dark red), with stars in matching colors when the mean is significantly different from zero (p<0.0001, t-test). Full lines indicate fits as described in **b**, for the family of models that fit the stationary data. The mean (g) and the standard deviation (d) of the inferred perturbation are shown in the left panels.

Discussion

Contemporary optogenetic perturbation protocols allow for precise manipulations of cell-type specific neuronal activity down to the single neuronal level, but it remains an open problem how best to read out circuit properties from such experiments.

In order to inform future perturbation experiments, we developed a framework that allows us to accurately describe the activity as a function of the stimulus, make experimentally testable predictions, and shed light on mechanisms underlying the control of neuronal activity and the influence of behavioral modulations. Specifically, we built a family of mathematically tractable high-dimensional models that can reproduce the distributions of activity of each cell-type's population in response to multiple stimulus contrasts. Building on recent developments on random matrix theory, we devised a theoretical approach that allowed us to derive closed expressions for the mean and variance of the distributions of responses to heterogeneous and partial optogenetic perturbations that are evaluated with the parameters inferred from the data.

We report four main findings. First, we found that there are hidden symmetries in the response matrix which enforce the responses to a SOM and a VIP perturbation to be of opposite sign and proportional, with the same proportionality constant across cell types. Second, we showed that a paradoxical response of any-given cell-type – its negative steady-state response to positive stimulation, or vice versa – implies that the circuit would be unstable without that cell-type, *i.e.* if that cell-type's activity were frozen. In the low-dimensional case, this finding generalizes the well-established concept of inhibition stabilized networks, and extends it to high-dimensional (HD) models. When VIP interneurons project only to SOM neurons, as appears approximately true empirically [11,13,19], we found that a paradoxical response of SOM interneurons implies instability of the E-PV sub-circuit. Given that in all our models the only cell-type that shows a paradoxical response is PV, we conclude that our family of models is PV-stabilized. Thirdly, we found that responses to partial perturbations are described by mixtures of Gaussian distributions whose mean and variance we were able to compute exactly. When the models have a paradoxical response to a full population perturbation, then these models will exhibit a *fractional paradoxical effect* to partial perturbations; namely, the fraction of PV cells showing a paradoxical response will be a non-monotonic function of the fraction of perturbed PV cells. We predict that all models that fit the data display a fractional paradoxical effect of PV for all values of stimulus contrast. , and we predict that the effect can be detected through holographic optogenetic experiments. We find furthermore that whenever the mean value of the perturbed population's response becomes negative, the sub-circuit without the perturbed cells loses stability. Finally our theoretical framework allowed us to compute the inputs to V1 that would elicit a response akin to that generated by locomotion. We predict that, intriguingly, strong inputs to both SOM and VIP but not PV mediate locomotion-dependent changes in V1 activity.

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 approach depends on two things. First, the use of recurrent neuronal models [15,16,33] for which mean-field equations allow us to compute, for a given set of network parameters, the mean and variance of the activities and the mean and variance of the inputs (Eqs. S38, S40). Second, an explicit expression for the distribution of activities in these models [17] that can be fit to the data, allowing an explicit expression for the goodness of fit of the model to the data activity distributions. With suitable simplifications, analogous methods could be used to fit models of multi-cell-type spiking networks, or to extend the model 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 [19].

By fitting the activity of each interneuron type in response to contrast manipulations, we uncovered key features of the synaptic connectivity observed in mouse V1 [11,13,19] (Figs. 2, compare Fig. 9): the lack of recurrence within the VIP and SOM populations, and the small values of the projections from VIP to E and PV. We found that when recurrent excitation is sufficiently strong these features are independent of all other fitting choices, and thus demonstrate that features of the dynamics implicitly carry information about the connectivity. We focused on small

stimulus sizes in order to avoid the treatment of longer-range circuits evoked by larger stimulus sizes[5,20,34], which would presumably require models with spatial structure[35]. Such models could in principle 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 HD models in which VIP only projects to SOM, using the homogeneous fixed-point approximation (HFA), the mean response of all cell-types to small perturbations to SOM or to VIP are perfectly anti-correlated, independent of stimulus configuration or parameter choice. The mean 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. 1,S13). This mathematical prediction of *Hidden Response Symmetries* therefore held, using the HFA, in the mean responses of the models that fit the data with remarkable fidelity (Fig. 4), and we found it to hold approximately for fully nonlinear systems (without the HFA). Furthermore, we conjectured and confirmed that given the nature of the circuit, these symmetries would hold at the single cell level in HD models (Fig. 4**d**), so we would expect them to hold 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[26,27,36], and establishes that tailored, simultaneous perturbations to SOM and VIP could largely cancel external inputs.

Inhibition stabilization is well-defined in circuits with multiple interneuron types[25,30], but how each interneuron type contributes to circuit stabilization, and the link between stabilization and response to perturbations, has not been entirely understood[8] (see also[9]). In this work, we offer a perspective that links the response of a perturbed population to the stability of the sub-circuit without that population, generalizing the notion of inhibition stabilization.In particular, if the sub-circuit without any given population is stable, then that population will not respond paradoxically to a perturbation. Conversely, if the population's response is paradoxical, the sub-circuit without it is unstable. Because the distribution of eigenvalues of the Jacobian of the HD network in the HFA has outliers given by the LD system (as expected theoretically[31]), we can generalize any theoretical finding of the LD system to the mean of the HD system under the HFA.

In our family of models, we find evidence in support of PV being the main circuit stabilizer (Fig. 5): its mean shows a paradoxical response (as in experiments,[4,8]), indicating that the circuit without the PV population is unstable. This instability is consistent with experimental observations[20] and theoretical considerations[9]. 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[5,20,34], it remains to be investigated how stabilization is performed across cortical space.

We find that an inhibitory cell type for which most or all cells respond negatively to a full perturbation (a paradoxical response) will show a fractional paradoxical effect in its responses to partial perturbations: With increasing fraction of stimulated cells of the given type, the fraction of cells of that type that respond negatively changes non-monotonically, first decreasing and then increasing. This is a very robust effect, independent of model details and evident in the many thousands of HD models that fit the data. It depends only on the facts that, when only a small fraction of cells are stimulated, the stimulated cells respond positively and the unstimulated negatively, so that most of the cells respond negatively; as the fraction stimulated increases, more cells become the positive-responding stimulated cells, causing the fraction responding negatively to shrink; but also, the responses of stimulated cells decrease and ultimately become largely or entirely negative, causing the fraction responding negatively to increase again.

Finally, we investigate the effect of locomotion on V1 activity. We consider the change in activity induced by locomotion, and regard those distributions as the response to an unknown perturbation (Fig. 7). Because we can access explicit expressions for the response to perturbations, we are able to fit these distributions and infer the

inputs to the network that would mimic this behavioral change in activity. Surprisingly, we find that the effect of locomotion is not only mediated by VIP[28] but that equally strong and equally wide inputs to SOM are needed to account for this effect. Remarkably, in the absence of visual stimulation, the inputs to E and PV are small, meaning that the inputs to SOM and VIP have canceling effects in pyramidal cells, whose mean change in activity with locomotion is non-significant in the absence of visual stimulation (Fig. 7**a** and[37]).

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. S119), 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. S111), 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 dependencies on visual stimulation[38,39]. Similar methods to the ones utilized here may be useful to explore temporal fluctuations around the fixed points. This work has thus laid foundations upon which a number of wider issues may be addressed, such as the reproducibility of contrast modulations of the population's spectral signatures found in the monkey[40] and the mouse[41] visual cortex and the corresponding predictions for cell-type-specific temporal and spectral responses.

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 and performed the numerical simulations and the analytical calculations with recommendations of F.F. and supervision from 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

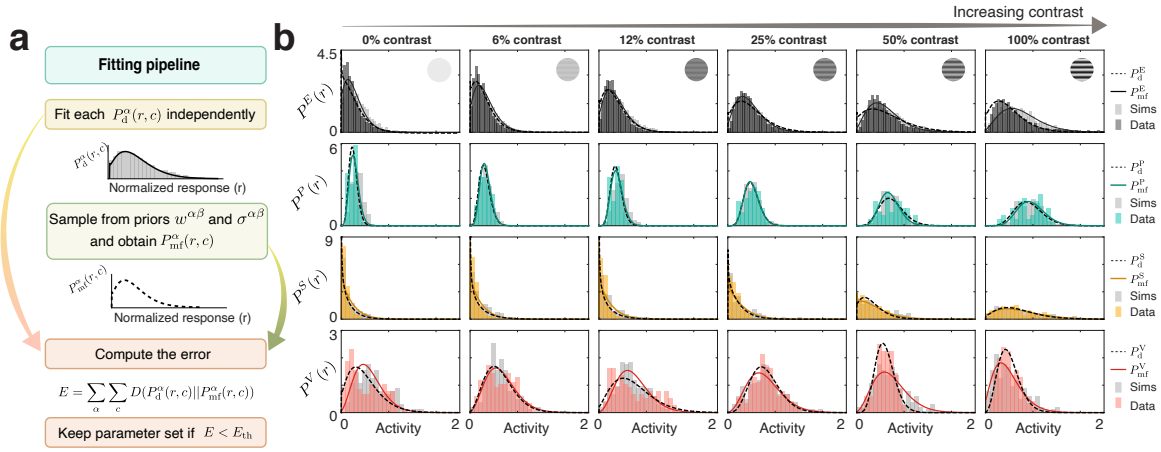

Figure 8 Fitting the distribution of responses to multiple contrasts with a mean-field model. **a)** Fitting pipeline: The distribution of activity generated by a model with a power-law non-linearity and recurrent and feed-forward inputs that are Gaussian distributed has an explicit mathematical form[17] (Eq. (S44)). 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 given by the LD model fit (see Fig. 2), 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 ($P_{\rm mf}^{\alpha}$) and the data ($P_{\rm c}^{\alpha}$) for all cell-types and contrast values, which can be found explicitly. By only accepting models with error less than a threshold of 0.5 (top 0.005%), we build a family of suitable models. **b)** 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).

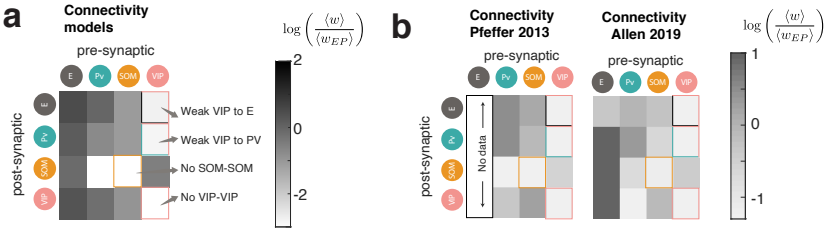

Figure 9 Direct comparison with experimentally reported synaptic connectivity. **a)** Mean of the distributions of weights shown in normalized by the synaptic weight from PV to E, for comparison with the available experimental data (background grayscale in panel 2**e**)). **b)** Left: Synaptic weight connectivity as obained in[11]. 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[11]. These two matrices are shown here for comparison. There is currently no agreement on the strength of the connection from PV to VIP.

Methods

1 Methods summary

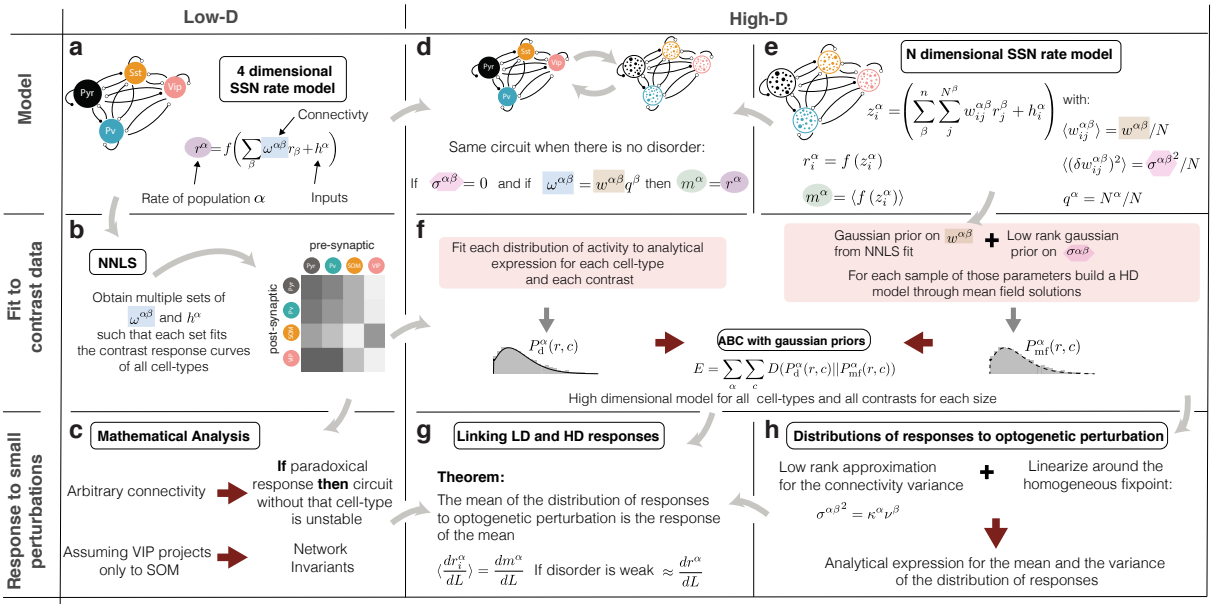

Figure S1 **Outline of this paper.** **a-b)** LD circuit: Multi cell-type circuit describing the population activity of E, PV, SOM, and VIP cells when presented with stimuli of different contrasts. By using non-negative least squares (NNLS) we find the parameters to describe the circuit's contrast response. 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. 4 **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 expression 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[42] 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. S1). In a first stage, we use non-negative least squares (NNLS)[23,26,43,44] (see Eq. S5)to fit a Low-dimensional (LD) 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. S1**b**), which allows a mathematical understanding of the response to perturbations to different cell-types (Fig. S1**c**). In a second stage, we build a family of HD models, with different numbers of cells per population. For that, we work with a HD rate model[15] (Fig. S1**d-e**, see Eq.S40). In this model, the distribution of activity has a tractable analytical form[17] (see Eq. S44) 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 HD model and its parameters, we can use MF theory [15] to self-consistently find the activity distributions that result for a given stimulus. Finally, in order to find the parameters of HD 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. S1**f**, and see Eq. S46). By choosing a suitable threshold on this distance (0.45) [45], we find HD 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 [18] which allow us to analytically compute the distribution of neuronal responses to patterned optogenetic perturbations under a suitable approximation (Fig. S1**h**) and determine its relation to the predictions in the LD circuit (Fig. S1**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 is 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

$$h^\alpha(c) = h_b^\alpha + h_c^\alpha c \tag{S4}$$

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) can be found by computing the non-negative least squares (NNLS)[46] solution.

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 $r_{c_i}^\alpha$ and standard error of the mean $s_{c_i}^\alpha$. 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

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

defined a hierarchy for the contrast response samples. From the family of LD models that fit the data, we only considered those that were ISN, and had a paradoxical response in PV interneurons. We did not enforce any connectivity weights to be zero. Some of our models had also absent connections from SOM to VIP, we disregarded those. Models shown in 2 **a** are the top 200 of the 700 models that later were used as prior *seeds*.

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 networks

Evaluating Eq. (S10) in the EI case we obtain the result from[7]

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

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*.

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

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

$$\tag{S18}$$

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 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[47], 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

$$r^{\text{new}} = A^{\xi} r \tag{S28}$$

$$W^{\text{new}} = AWA^{-\xi} \tag{S29}$$

$$h^{\text{new}} = Ah \tag{S30}$$

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:

$$J^{\text{new}} = A^{\xi} J A^{-\xi} \qquad R^{\text{new}} = A^{\xi} R A^{-1} \tag{S31}$$

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 network models. The network has $n = 4$ populations with N^{α} neurons in each population $\alpha = \{E, PV, S, V\}$. We denote the fraction of neurons in each population by $q^{\alpha} = N^{\alpha}/N$, where N is the total amount of neurons in the network. We took this fraction to be $q = [0.8, 0.1, 0.05, 0.05]$ as is approximately in biology[10]. The steady-state activity r_i^{α} of the unit i in the population α is given by:

$$r_i^{\alpha} = f(z_i^{\alpha}) \qquad z_i^{\alpha} = \left(\sum_{\beta}^{n} \sum_{j}^{N^{\beta}} w_{ij}^{\alpha\beta} r_j^{\beta} + h_i^{\alpha} \right) \tag{S32}$$

Whereby $f(z) = []_+^{\xi}$ with $\xi = 2$ represents the transfer function of the neuronal populations. The connectivity elements $w_{ij}^{\alpha\beta}$ are Gaussian distributed with mean and variance defined by:

$$\langle w_{ij}^{\alpha\beta} \rangle = w^{\alpha\beta}/N \qquad \langle (w_{ij}^{\alpha\beta})^2 \rangle - \langle w_{ij}^{\alpha\beta} \rangle^2 = \sigma^{\alpha\beta\,2}/N \tag{S33}$$

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 steady state Eq. (S32) can be re written as a function of the input to each cell :

$$z_i^\alpha = \sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta} f\left(z_j^\beta\right) + h_i^\alpha \tag{S34}$$

4.1 Set-up and mean field equations

In order to compute the mean and variance of the activity in each population self-consistently, we follow the approach in Kadmon and Sompolinsky [15]. The input z_i^α to a cell can be described as fluctuations around a mean: $z_i^\alpha = u^\alpha + \delta z_i^\alpha$. We define:

$$m^\alpha = \langle f(z_i^\alpha)\rangle \tag{S35}$$
$$v^\alpha = \langle f(z_i^\alpha)^2\rangle \tag{S36}$$
$$q^\alpha = N^\alpha/N \tag{S37}$$

By taking the mean and the variance of Eq. (S34) and incorporating the definitions above, we re-obtain the self-consistent equations for the mean and the variance of z_i^α, given by u^α and Δ^α

$$u^\alpha = \sum_\beta w^{\alpha\beta} q^\beta m^\beta + h_0^\alpha \tag{S38}$$
$$\Delta^\alpha = \sum_\beta (\sigma^{\alpha\beta})^2 q^\beta v^\beta + (\lambda^\alpha)^2 \tag{S39}$$

where

$$m^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z) e^{-z^2/2} dz \tag{S40}$$
$$v^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z)^2 e^{-z^2/2} dz \tag{S41}$$

We observe that if there is no disorder, Eqs. (S38) and (S40) 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 L is a *homogeneous optogenetic* perturbation to the entire population α, the change in response of each cell is given by

$$\frac{dr_i^\alpha}{dL} = f'(u^\alpha + \sqrt{\Delta^\alpha} z_i) \left(\frac{du^\alpha}{dL} + \frac{1}{2\sqrt{\Delta^\alpha}} \frac{d\Delta^\alpha}{dL} z_i \right) \tag{S42}$$

Taking the average and using Eq. (S40), we find that the mean of the response distribution to laser perturbation is given by the change in the mean activity of the population:

$$\begin{aligned}\langle \frac{dr_i^\alpha}{dL} \rangle =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(u^\alpha + \sqrt{\Delta^\alpha} z) \left(\frac{du^\alpha}{dL} + \frac{1}{2\sqrt{\Delta^\alpha}} \frac{d\Delta^\alpha}{dL} z \right) e^{-z^2/2} dz \\ =& \frac{d}{dL} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z) e^{-z^2/2} dz \right) \\ =& \frac{dm^\alpha}{dL}\end{aligned} \tag{S43}$$

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

4.3 Data Fitting

To fit the system defined by Eqs. (S38,S40), we used that the distribution of activity of a population α with the transfer function of the form $f(z) = [z]_+^\xi$ can be written (when assuming that inputs are Gaussian distributed) as a function of the mean total input u^α and its variance Δ^α [17]:

$$P^\alpha(r) = \frac{1}{\sqrt{2\pi} r^{1-1/\xi} \xi \sqrt{\Delta^\alpha}} e^{-\frac{(r^{1/\xi} - u^\alpha)^2}{2\Delta^\alpha}} \Theta(r) + \frac{1}{2} \left(1 - \mathrm{erf}\left(\frac{u^\alpha}{\sqrt{2\Delta^\alpha}} \right) \right) \delta(r) \tag{S44}$$

$$P^\alpha(r) = P^+(r)\Theta(r) + P^0 \delta(r) \tag{S45}$$

Here, Θ and δ denote the Heaviside and delta functions, respectively.

To find the parameters that approximate the distribution of the experimentally recorded activity we use Eq. S44 with $\xi = 2$ and proceed as follows (Fig. 2**a**, see also Fig. 8): For each cell-type α and each contrast c, we fit the analytical distribution of rates from Eq. (S44) to the distribution of experimentally recorded activity. We denote the fit distribution by $P_{\mathrm{d}}^\alpha(r, \mu_{\mathrm{d}}^\alpha(c), \Delta_{\mathrm{d}}^\alpha(c))$ (dashed lines in Fig. 8). The fitted distribution P_{d}^α provides us with an estimate of the mean ($\mu_{\mathrm{d}}^\alpha(c)$) and variance ($\Delta_{\mathrm{d}}^\alpha(c)$) of the total input to each cell-type α and each contrast c. We assume that the external input to the population α has the form $h^\alpha = h_b^\alpha + h_c^\alpha c$. To find which parameters $w^{\alpha\beta}$, $\sigma^{\alpha\beta}$, h^α and λ^α 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 $P_{\mathrm{mf}}^\alpha(r, \mu_{\mathrm{mf}}^\alpha(c), \Delta_{\mathrm{mf}}^\alpha(c))$. The priors for $w^{\alpha\beta}$ and h_b^α and h_c^α were Gaussian distributions with mean given by the parameters of the LD fits and a 5% std. The priors for $\sigma^{\alpha\beta}$ and λ^α were chosen arbitrarily. The only dependence on contrast is through the mean activity, the variance in the inputs was independent of contrast. We define an error that depends uniquely on $\mu_{\mathrm{d}}^\alpha(c), \Delta_{\mathrm{d}}^\alpha(c), \mu_{\mathrm{mf}}^\alpha(c), \Delta_{\mathrm{mf}}^\alpha(c)$. 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:

$$E = \sum_c \sum_\alpha D(P_{\mathrm{d}}^\alpha(c) || P_{\mathrm{mf}}^\alpha(c))^2 \tag{S46}$$

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

$$D(P_{\mathrm{d}}^{\alpha}||P_{\mathrm{mf}}^{\alpha}) = \int_{-\infty}^{\infty} P_{\mathrm{d}}^{\alpha}(r)\log\frac{P_{\mathrm{d}}^{\alpha}(r)}{P_{\mathrm{mf}}^{\alpha}(r)}dr = P_{\mathrm{d}}^{0,\alpha}\log\frac{P_{\mathrm{d}}^{0,\alpha}}{P_{\mathrm{mf}}^{0,\alpha}} + \int_{0^+}^{\infty} P_{\mathrm{d}}^{+,\alpha}(r)\log\frac{P_{\mathrm{d}}^{+,\alpha}(r)}{P_{\mathrm{mf}}^{+,\alpha}(r)}dr = I_A^{\alpha} + I_B^{\alpha} \tag{S47}$$

with

$$\mathrm{I}_A^{\alpha} = \frac{1}{2}\,\mathrm{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)\log\left(\frac{\mathrm{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)}{\mathrm{erf}\left(\frac{\mu_{\mathrm{mf}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{mf}}^{\alpha}}}\right)}\right) \tag{S48}$$

$$\mathrm{I}_B^{\alpha} = \frac{1}{2}\left(1+\mathrm{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)\right)\log\frac{\sqrt{\Delta_{\mathrm{mf}}^{\alpha}}}{\sqrt{\Delta_{\mathrm{d}}^{\alpha}}} + \frac{1}{4\Delta_{\mathrm{mf}}^{\alpha}}\left(\left(\Delta_{\mathrm{d}}^{\alpha}-\Delta_{\mathrm{mf}}^{\alpha}+\left(\mu_{\mathrm{d}}^{\alpha}-\mu_{\mathrm{mf}}^{\alpha}\right)^2\right)\left(1+\mathrm{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)\right)+\right. \tag{S49}$$

$$\left.\sqrt{\frac{2}{\pi\Delta_{\mathrm{d}}^{\alpha}}}e^{-\frac{\mu_{\mathrm{d}}^{\alpha 2}}{2\Delta_{\mathrm{d}}^{\alpha}}}\left(\Delta_{\mathrm{d}}^{\alpha}(\mu_{\mathrm{d}}^{\alpha}-2\mu_{\mathrm{mf}}^{\alpha})+\Delta_{\mathrm{mf}}^{\alpha}\mu_{\mathrm{d}}^{\alpha}\right)\right) \tag{S50}$$

Instead of following the gradient to find an optimal solution we keep the solutions that have a sufficiently small error from the random sampling. Randomly sampling from these priors we obtained 500000 models whose total KL divergence was 0.7. From those, we take the first 300 for most figures. This defines a family of high-dimensional models (Fig. 2) with skewed distributions that are in good agreement with the calcium activity, and capture not only the nonlinear dependence of the activity mean 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. (S32) ${}^*r_i^{\alpha}$ 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}:

$$\tau_i^{\alpha}\delta\dot{r}_i^{\alpha} = -\delta r_i^{\alpha} + f'^{\alpha}_i\cdot\left(\sum_{\beta}^{n}\sum_{j}^{N^{\beta}} w_{ij}^{\alpha\beta}\delta r_j^{\beta} + \delta h_i^{\alpha}\right) \qquad f'^{\alpha}_i = f'^{\alpha}({}^*z_i^{\alpha}) = f'\left(\sum_{\beta}^{n}\sum_{j}^{N^{\beta}} w_{ij}^{\alpha\beta}{}^*r_j^{\beta} + h_i^{\alpha}\right) \tag{S51}$$

Switching from 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 Kronecker 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:

$$\delta\vec{r} = (F^{-1}-W)^{-1}\vec{\delta h} \qquad \vec{\delta r} = R\vec{\delta h} \tag{S52}$$

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. (S33)) 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 $W_{0_{ij}}^{\alpha\beta} = \frac{1}{N} w^{\alpha\beta}$ 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:

$$W = W_0 - \Pi_L J \Pi_R \qquad \text{with} \qquad \Pi_L = \delta_{ij}\sqrt{\kappa_{\alpha_j}} \qquad \Pi_R = \delta_{ij}\sqrt{\nu_j} \tag{S53}$$

5.2 Homogeneous fixed point approximation (HFA)

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

$$\tau_i^\alpha \delta \dot{r}_i^\alpha = -\delta r_i^\alpha + f'^\alpha \cdot \left(\sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta} \delta r_j^\beta + \delta h_i^\alpha \right) \qquad f'^\alpha_i = f'^\alpha = f' \left(\sum_\beta^n \sum_j^{N^\beta} w^{\alpha\beta *} r^\beta + h^\alpha \right) \tag{S54}$$

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 Eigenvalue Spectrum of the Jacobian in the HFA

In[31], it is shown that given a matrix with iid entries to which is added a low rank matrix, under conditions that are satisfied in our models, the distribution of eigenvalues of the Jacobian in the HFA is going to follow the circular law, except for a set of outlier eigenvalues. These eigenvalues are placed asymptotically in the same location as the eigenvalues of the low rank added matrix. In our case, the Jacobian of the HD in the HFA will be

$$J = -I + F W_0 - F \Pi_L J \Pi_R \tag{S55}$$

Where in this case $F_{ij}^{\alpha\beta} = \delta_{ij}\delta_{\alpha\beta} f'^\alpha$, f'^α is the gain of the LD circuit, and $W_{0_{ij}}^{\alpha\beta} = \frac{1}{N} w^{\alpha\beta}$ as before. The non-trivial eigenvalues of the low rank component of the Jacobian $J_{LR} = -I + F W_0$, are exactly the same as the eigenvalues of the LD circuit Jacobian. This can be seen when considering the basis composed of n eigenvectors of the form $u^{(k)} = \frac{1}{\sqrt{N_k}} (\underbrace{0,\ldots,0}_{\sum_{l=1}^{k-1} N_l}, \underbrace{1,\ldots,1}_{N_k}, \underbrace{0,\ldots,0}_{N-\sum_{l=1}^{k} N_l})$ and $N-n$ orthogonal eigenvectors. In this basis, the non-zero entries of the J_{LR} are given by the Jacobian of the LD system.

5.4 General framework to compute the linear response in networks in the HFA

Using results from[18] we find that in the special case of the HFA, $f'^\alpha_i = f'^\alpha$ described above, the mean linear response matrix over different instantiations of the disorder is the linear response of the non- disordered case:

$$\langle (F_{HFA}^{-1} - W)^{-1} \rangle_J = \langle (F_{HFA}^{-1} - W_0 + \Pi_L J \Pi_R)^{-1} \rangle_J = (F_{HFA}^{-1} - W_0)^{-1} = R^0 \tag{S56}$$

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 useful approximation in this case of study. Generally, in experiments, we have a perturbation pattern δh describing the proportion of stimulation each neuron receive, and a measuring vector δb, describing which are the neurons contributing (linearly) to the signal that we are going to be monitoring $s = \vec{\delta r}^{\,\mathsf{T}} \delta b$. We compute the mean and variance of the signal s across different instantiations of the disorder. By defining:

$$\text{the } \textit{measuring} \text{ matrix} \qquad B = \delta b \delta b^{\mathsf{T}} \tag{S57}$$

$$\text{the } \textit{optogenetic perturbation} \text{ matrix} \qquad \Sigma = \delta h \delta h^{\mathsf{T}} \tag{S58}$$

we can write the second moment of that measured signal s [18]:

$$\langle s^2 \rangle = \langle (\vec{\delta r}^{\,\mathsf{T}} \delta b)^2 \rangle = \langle \vec{\delta r}^{\,\mathsf{T}} B \vec{\delta r} \rangle = \mathscr{F} + \Delta\mathscr{F} \tag{S59}$$

where

$$\mathscr{F} = Tr(BR^0 \Sigma R^{0\mathsf{T}}) \qquad \Delta\mathscr{F} = \frac{1}{N} \frac{Tr\big(BR^0 \Pi_L \Pi_L R^{0\mathsf{T}}\big) Tr\big(\Pi_R \Pi_R R^0 \Sigma R^{0\mathsf{T}}\big)}{1 - \frac{1}{N} Tr\big(\Pi_R \Pi_R R^0 \Pi_L \Pi_L R^{0\mathsf{T}}\big)} \tag{S60}$$

Where we used the definitions in Eq. (S53). We observe that in the absence of disorder, in which $W = W_0$, $\Delta\mathscr{F} = 0$ and the recorded signal is given only by $s = \delta b R^0 \delta h^{\mathsf{T}}$.

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

$$\langle \delta r_i \rangle_J = e_i R^0 \vec{\delta h} \qquad \text{where} \qquad e_i = \{0, ..., 1, ..., 0\} \tag{S61}$$

Eq. (S61) means that for each neuron, the distribution of linear responses over different instantiations of the connectivity has a mean given by the linear response in the absence of disorder (due to Eq. (S56)) and the variance Λ_i given by

$$\Lambda_i^2 = \langle \delta r_i^2 \rangle_J - \langle \delta r_i \rangle_J^2 = \langle \delta r_i^2 \rangle_J - \mathscr{F} = \Delta\mathscr{F} \tag{S62}$$

Equations (S56), (S59) and (S60) 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.

5.5 Computation of the response matrix R^0 without disorder

For computing R^0 (given by Eq. (S56)) 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 $u^{(k)} = \frac{1}{N_k} (\underbrace{0, \ldots, 0}_{\sum_{l=1}^{k-1} N_l}, \underbrace{1, \ldots, 1}_{N_k}, \underbrace{0, \ldots, 0}_{N - \sum_{l=1}^{k} N_l})$ and similarly for v and obtain

$$W_0 = V\omega U^\intercal \tag{S63}$$

Where $w^{\alpha\beta}$ and q^α were introduced in Eqs (S33) and (S37), respectively. To obtain R^0 defined in Eq. (S56) 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:

$$R^0 = \left(F^{-1} - W_0\right)^{-1} = \left(F^{-1} - V\omega U^\intercal\right)^{-1} = F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^\intercal F \tag{S64}$$

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

$$R^0_{ij} = \delta_{ij} f'_{\alpha_i} - \frac{f'_{\alpha_i} f'_{\alpha_j}}{N_{\alpha_j}} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S65}$$

$\mathbf{f}'$ 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, 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

$$S_{\alpha_i\alpha_j} = \delta_{\alpha_i\alpha_j} f'_{\alpha_j} - f'_{\alpha_i} f'_{\alpha_j} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S66}$$

Substituting Eq. (S66) into (S65), we obtain an expression for the linear response which will be useful in later sections:

$$R^0_{ij} = \delta_{ij} f'_{\alpha_i} + \frac{S_{\alpha_i\alpha_j}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j} f'_{\alpha_j}}{N_{\alpha_j}} \tag{S67}$$

We point out that the Eq. (S66) is independent of N, and is finite in the limit of large N.

2) case of rank-one ω

In the case of rank-one ω , we cannot invert ω in Eq. (S64) . Instead we write:

$$W_0 = \frac{vu^\intercal}{N} \qquad v = \{1, ..., 1, ..., 1\} \tag{S68}$$

$$u = \{\underbrace{w_1, \cdots, w_1}_{N_1}, \underbrace{w_2, \cdots, w_2}_{N_2}, \cdots, \underbrace{w_n, \cdots, w_n}_{N_n}\} \tag{S69}$$

Using Sherman–Morrison formula we find that

$$R^0 = (F^{-1} - W_0)^{-1} = F + \frac{1}{N}\frac{Fvu^T F}{1 - \frac{u^T F v}{N}} \tag{S70}$$

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

$$R^0_{ij} = \delta_{ij} f'_{\alpha_j} + \frac{1}{N}\frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} = \delta_{ij} f'_{\alpha_j} + \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S71}$$

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

$$S_{\alpha_i \alpha_j} = \delta_{\alpha_i \alpha_j} f'_{\alpha_j} + N_{\alpha_j}\frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S72}$$

We note that the above expression is also finite in the large N limit. Using (S71) and (S72) we conclude that the linear response R^0_{ij} satisfies Eq. (S67) also in this case.

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

In this section we consider fractional perturbations of neural populations, i.e. perturbations applied only to a subset of neurons in each population. We derive a formula for the sum of the linear response of a single neuron in a population α_i to perturbations applied to fractions γ_{α_j} of neurons in populations α_j. Within each perturbed population α_j, we denote the set of perturbed neurons by $\mathscr{P}_{\alpha_j}$. If we perturb γ_{α_j} neurons in a population α_j then we find that the response of the neurons in population α_j that were stimulated have a mean response that depends on whether they were directly stimulated or not (see below).

1) case of invertible ω

In the case of full rank ω (R^0_{ij} is given by (S65)), if we perturb γ_η neurons in populations η the total perturbation vector is given by $\delta h = \{\delta h_1, \delta h_2, \cdots, \delta h_n\}$, 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

$$\mu_i = \sum_j R_{ij}\delta h_j = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} - f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j}f'_{\alpha_j}\left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S73}$$

The expression in Eq. (S73) can be represented as a sum of the mean responses of directly perturbed and non perturbed neurons. The mean response of directly stimulated neurons is given by

$$\mu^{IN}_{i\in\mathscr{P}_{\alpha_i}} = f'_{\alpha_i} - f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j}f'_{\alpha_j}\left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S74}$$

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

$$\mu^{OUT}_{i\notin\mathscr{P}_{\alpha_i}} = -f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j}f'_{\alpha_j}\left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S75}$$

We note that these expressions critically depend on the sign of $\left(\mathbf{f}' - \omega^{-1}\right)^{-1}$. To capture Eq. (S74) and Eq. (S75) as a single equation we define a matrix

$$\begin{aligned}\chi^{\boldsymbol{\gamma}} =& \mathbf{f}'\delta_p - \mathbf{f}'\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\mathbf{f}'\boldsymbol{\gamma} && \text{(S76)}\\ =& \mathbf{f}'\delta_p + \mathbf{f}'\omega\left(\mathbf{f}'^{-1} - \omega\right)^{-1}\boldsymbol{\gamma} && \text{(S77)}\\ =& \mathbf{f}'\delta_p + \mathbf{f}'\omega\chi\boldsymbol{\gamma} && \text{(S78)}\end{aligned}$$

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

In the case when we study the paradoxical response, meaning that we perturb and record activity in the same population we find that using $\sum_{\alpha_j}(\mathbf{f}'^{-1} - \omega)_{\alpha_i\alpha_j}\chi_{\alpha_j\alpha_i} = 1$ (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 rewrite (S73) as

$$\mu_i = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} + f'_{\alpha_i}\gamma_{\alpha_i}[\omega\chi]_{\alpha_i\alpha_i} = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} + \gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i}) \tag{S79}$$

If the response is paradoxical in the Low-dimensional system ($\chi_{\alpha_i\alpha_i} < 0$), the response distribution of non stimulated neurons has a negative mean, and becomes even more negative if the fraction of perturbed increases. If the above term is positive for a small fraction of perturbed cells, it can become negative when the fraction of the perturbed cells increases. We denote the critical fraction of perturbed cells for which the response mean becomes negative by $\gamma^C_{\alpha_i}$ and obtain

$$0 < \frac{f'_{\alpha_i}}{f'_{\alpha_i} - \chi_{\alpha_i\alpha_i}} < \gamma^C_{\alpha_i} < 1 \tag{S80}$$

In Fig. 7, 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

$$R_{ij} = \delta_{ij} f'_{\alpha_j} - \frac{1}{N} \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} = \delta_{ij} f'_{\alpha_j} + \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S81}$$

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

$$\mu_i^{IN} = f'_{\alpha_i} + f'_{\alpha_i} \sum_{\alpha_j} \gamma_{\alpha_j} q_{\alpha_j} \frac{f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} \tag{S82}$$

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

$$\mu_i^{OUT} = f'_{\alpha_i} \sum_{\alpha_j} \gamma_{\alpha_j} q_{\alpha_j} \frac{f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} \tag{S83}$$

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. (S83) is always negative), but the response of those neurons that were stimulated will only be paradoxical when $\mu_i^{IN} < 0$

$$\frac{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{q_{\alpha_j} f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S84}$$

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 (S84) is now

$$\frac{n - \sum_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S85}$$

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. [32]

If now we would normalize the weights in Eq. (S68) as in [32], meaning that $w_j/N \to w_j/(N/n)$ (See Eq. 2 of their paper), the above equation would be

$$\frac{1-\sum_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{f'_{\alpha_j}|w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S86}$$

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

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

From Eq. (S60) 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 (S67). We will first find the general expression and then evaluate for particular cases: For that we write (S60) as

$$\Lambda^2 = \frac{1}{N}\frac{Tr\left(BR^0\Pi_L\Pi_L R^{0\intercal}\right)Tr\left(\Pi_R\Pi_R R^0\Sigma R^{0\intercal}\right)}{1-\frac{1}{N}Tr\left(\Pi_R\Pi_R R^0\Pi_L\Pi_L R^{0\intercal}\right)} = \frac{M.O}{1-D} \tag{S87}$$

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:

$$\Sigma_{jk} = \sum_\eta \sum_{\eta'} \delta_{\alpha_j \eta}\delta_{\alpha_k \eta'}\delta_{k\in\mathscr{P}_{\alpha_k}}\delta_{j\in\mathscr{P}_{\alpha_j}} \tag{S88}$$

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 ($q_\eta = N_\eta/N$):

$$\Lambda^2_{\alpha_l} = \frac{\left(f'^2_{\alpha_l}\kappa_{\alpha_l} + \frac{1}{N}\left(\sum_\eta \frac{\kappa_\eta}{q_\eta}S^2_{\alpha_l\eta} - \frac{f'^2_{\alpha_l}\kappa_{\alpha_l}}{q_{\alpha_l}}\right)\right)\left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta}\gamma_\eta\right)^2 q_{\eta'}\right)}{1-\sum_\eta \nu_\eta\left(q_\eta f'^2_\eta \kappa_\eta + \frac{1}{N}\left(q_\eta\sum_{\eta'}\frac{\kappa_{\eta'}}{q_{\eta'}}S^2_{\eta\eta'} - f'^2_\eta\kappa_\eta\right)\right)} \tag{S89}$$

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

$$\Lambda^2_{\alpha_l} = \frac{f'^2_{\alpha_l}\kappa_{\alpha_l}\left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta}\gamma_\eta\right)^2 q_{\eta'}\right)}{1-\sum_\eta \nu_\eta q_\eta f'^2_\eta \kappa_\eta} \tag{S90}$$

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. (S90) 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:

$$\Lambda^2_{\alpha_l} = \sigma^2 \frac{\left(f'^2_{\alpha_l} + \sum_\eta \frac{S^2_{\alpha\eta}}{N_\eta} - \frac{f'^2_{\alpha_l}}{N_{\alpha_l}}\right)\left(\sum_\eta f'^2_\eta \gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'}\left(\sum_\eta S_{\eta',\eta}\gamma_\eta\right)^2 q_{\eta'}\right)}{1-\sigma^2\sum_\eta\left(q_\eta f'^2_\eta + q_\eta \sum_{\eta'} S^2_{\eta,\eta'}\frac{1}{N_{\eta'}} - \frac{1}{N}f'^2_\eta\right)} \tag{S91}$$

5.8 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

$$\rho^{\text{IN}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda_{\alpha_i}}\exp\left\{-\frac{(x-\mu^{\text{IN}}_{\alpha_i})^2}{2\Lambda^2_{\alpha_i}}\right\} \tag{S92}$$

$$\rho^{\text{OUT}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda_{\alpha_i}}\exp\left\{-\frac{(x-\mu^{\text{OUT}}_{\alpha_i})^2}{2\Lambda^2_{\alpha_i}}\right\} \tag{S93}$$

So the total distribution of responses is

$$\rho_{\alpha_i} = \gamma_{\alpha_i}\rho^{\text{IN}}_{\alpha_i} + (1-\gamma_{\alpha_i})\rho^{\text{OUT}}_{\alpha_i} \tag{S94}$$

Where $\mu^{\text{IN}} = R^{IN}_{j\in\mathscr{P}_{\alpha_j}}$ and $\mu^{\text{OUT}} = R^{OUT}_{j\in\mathscr{P}_{\alpha_j}}$ given by Eqs (S82, S83) for low rank ω or by (S74,S75) for invertible ω, and a variance given by Eq. (S89)

5.9 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[32]). In this case, the distribution of responses will be bimodal, with two delta functions at the values given by Eq. (S79). The density then will be given by the limit of vanishing variance of . (S94)

$$\rho_{\alpha_i}(x) = (1-\gamma_{\alpha_i})\delta(x-\gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i})) + \gamma_{\alpha_i}\delta(x - f'_{\alpha_i} - \gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i})) \tag{S95}$$

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. (S80), 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.

5.10 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

$$\chi^5 = [f_5'^{-1} - \omega_5]^{-1} = -I\begin{bmatrix} \omega^{\text{EE}} - 1/f'_E & \omega^{\text{EP}}\gamma_P & \omega^{\text{EP}}(1-\gamma_P) & \omega^{\text{ES}} & \omega^{\text{EV}} \\ \omega^{\text{PE}} & \omega^{\text{PP}}\gamma_P - 1/f'_P & \omega^{\text{PP}}(1-\gamma_P) & \omega^{\text{PS}} & \omega^{\text{PV}} \\ \omega^{\text{PE}} & \omega^{\text{PP}}\gamma_P & \omega^{\text{PP}}(1-\gamma_P) - 1/f'_P & \omega^{\text{PS}} & \omega^{\text{PV}} \\ \omega^{\text{SE}} & \omega^{\text{SP}}\gamma_P & \omega^{\text{SP}}(1-\gamma_P) & \omega^{\text{SS}} - 1/f'_s & \omega^{\text{SV}} \\ \omega^{\text{VE}} & \omega^{\text{VP}}\gamma_P & \omega^{\text{VP}}(1-\gamma_P) & \omega^{\text{VS}} & \omega^{\text{VV}} - 1/f'_V \end{bmatrix}^{-1} \tag{S96}$$

$$\chi^5_{pp} = [f_5'^{-1} - \omega_5]^{-1}_{pp} = \frac{1}{\det\left[\mathbf{f}'_5{}^{-1} - \omega_5\right]} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f'_E & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}(1-\gamma_P) - 1/f'_P & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f'_s & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f'_V \end{bmatrix} \tag{S97}$$

$$\chi^5_{PP} = \frac{1}{\det\left[\mathbf{f}'_5{}^{-1} - \omega_5\right]} \Big(- \omega^{\mathrm{PE}} \det \begin{bmatrix} \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f'_s & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f'_V \end{bmatrix} \tag{S98}$$

$$+ (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f'_P) \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f'_E & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SS}} - 1/f'_s & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f'_V \end{bmatrix} \tag{S99}$$

$$- \omega^{\mathrm{PS}} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f'_E & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VV}} - 1/f'_V \end{bmatrix} \tag{S100}$$

$$+ \omega^{\mathrm{PV}} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f'_E & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f'_s \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} \end{bmatrix} \tag{S101}$$

dotted

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

$$\chi^5_{PP} = \frac{1}{\det\left[\mathbf{f}'_5{}^{-1} - \omega_5\right]} \left(\omega^{\mathrm{PE}} M_{PE}(1-\gamma_P) - (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f'_P) M_{PP} + \omega^{\mathrm{PS}} M_{PS}(1-\gamma_P) - \omega^{\mathrm{PV}} M_{PV}(1-\gamma_P)\right) \tag{S102}$$

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

$$\chi^5_{PP} = - \frac{\det\left(\mathbf{f}'^{-1} - \omega\right)}{\det\left[\mathbf{f}'_5{}^{-1} - \omega_5\right]} \left(\omega^{\mathrm{PE}} \chi_{EP}(1-\gamma_P) + (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f'_P)\chi_{PP} + \omega^{\mathrm{PS}} \chi_{SP}(1-\gamma_P) + \omega^{\mathrm{PV}} \chi_{VP}(1-\gamma_P)\right) \tag{S103}$$

$$- \frac{\det\left(\mathbf{f}'^{-1} - \omega\right)}{\det\left[\mathbf{f}'_5{}^{-1} - \omega_5\right]} \left([\omega\chi]_{PP}(1-\gamma_P) - 1/f'_P \chi_{PP}\right) \tag{S104}$$

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

$$\chi^5_{PP} = \frac{\det\left(\mathbf{f}'^{-1} - \omega\right)}{\det\left[\mathbf{f}'_5{}^{-1} - \omega_5\right]} \left(\gamma_P [\omega\chi]_{PP} + 1\right) \tag{S105}$$

Given that the mean response of the perturbed population in a high-dimensional system, given by Eq. (S73) (and

also Eq. S126) is $\chi^{\boldsymbol{\gamma}}_{PP} = f'_P(\gamma_P[\omega\chi]_{PP} + 1)$, we obtain that

$$f'_P \chi^5_{PP} = \frac{\det\left(\mathbf{f'}^{-1} - \omega\right)}{\det\left[\mathbf{f'}_5{}^{-1} - \omega_5\right]} \chi^{\boldsymbol{\gamma}}_{PP} \tag{S106}$$

And as both determinants are positive because of linear stability, this two things have the same sign. This calculation, together with Eq. (2), tells us that whenever the mean of the perturbed population is positive, then the sub-circuit without them will be unstable.

5.11 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 (S59 , S60) 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 $D^\eta_i \sim \mathcal{N}(d_\eta, g^2_\eta)$.

The optogenetic targeting matrix Σ, instead of being given by Eq. (S88), will be in this case:

$$\Sigma_{jk} = \sum_\eta \sum_{\eta'} \delta_{\alpha_j \eta} \delta_{\alpha_k \eta'} \delta_{k \in \mathscr{P}_{\alpha_k}} \delta_{j \in \mathscr{P}_{\alpha_j}} (D^{\alpha_i}_i)(D^{\alpha_j}_j) \tag{S107}$$

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 (S59 , S60). Specifically:

$$\delta r_i = \sum_j R^0_{ij} \delta h_j + \Lambda_{\alpha_i} \xi_i \tag{S108}$$

$$= f'_{\alpha_i} D_i - f'_{\alpha_i} \sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f'} - \omega^{-1} \right)^{-1} \right]_{\alpha_i \eta} d_\eta + \Lambda_{\alpha_i} \xi_i \tag{S109}$$

$$\tag{S110}$$

where $\xi_i \sim \mathcal{N}(0,1)$ and Λ_{α_i} is the generalization of Eq. (S89) to disordered perturbations, obtained by replacing Eq. (S107) into (S87) (we note that the only term that needs to be re-computed is the term M).

$$\Lambda^2_{\alpha_i} = \frac{\left(f'^2_{\alpha_i} \kappa_{\alpha_i} + \frac{1}{N} \left(\sum_\eta \frac{\kappa_\eta}{q_\eta} S^2_{\alpha_i \eta} - \frac{f'^2_{\alpha_i} \kappa_{\alpha_i}}{q_{\alpha_i}} \right) \right) \left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta \left((1-\gamma_\eta) d^2_\eta + g^2_\eta \right) q_\eta + \sum_{\eta'} \nu_{\eta'} \left(\sum_\eta S_{\eta',\eta} d_\eta \gamma_\eta \right)^2 q_{\eta'} \right)}{1 - \sum_\eta \nu_\eta \left(q_\eta f'^2_\eta \kappa_\eta + \frac{1}{N} \left(q_\eta \sum_{\eta'} \frac{\kappa_{\eta'}}{q_{\eta'}} S^2_{\eta\eta'} - f'^2_\eta \kappa_\eta \right) \right)} \tag{S111}$$

In the large N limit, this equation reduces to

$$\Lambda^2_{\alpha_i} = \frac{f'^2_{\alpha_i} \kappa_{\alpha_i} \left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta \left((1-\gamma_\eta) d^2_\eta + g^2_\eta \right) q_\eta + \sum_{\eta'} \nu_{\eta'} \left(\sum_\eta S_{\eta',\eta} d_\eta \gamma_\eta \right)^2 q_{\eta'} \right)}{1 - \sum_\eta \nu_\eta q_\eta f'^2_\eta \kappa_\eta} \tag{S112}$$

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

$$\mu^{IN}_{\alpha_i} = f'_{\alpha_i} d_{\alpha_i} - f'_{\alpha_i} \sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1} \right)^{-1} \right]_{\alpha_i \eta} d_\eta \tag{S113}$$

$$\mu^{OUT}_{\alpha_i} = -f'_{\alpha_i} \sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1} \right)^{-1} \right]_{\alpha_i \eta} d_\eta \tag{S114}$$

$$\Lambda^{2,IN}_{\alpha_i} = f'^2_{\alpha_i} g^2_{\alpha_i} + \Lambda^2_{\alpha_i} \tag{S115}$$

$$\Lambda^{2,OUT}_{\alpha_i} = \Lambda^2_{\alpha_i} \tag{S116}$$

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

$$\rho^{\mathrm{IN}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda^{\mathrm{IN}} a_{\alpha_i}} \exp\left\{ -\frac{(x - \mu^{\mathrm{IN}}_{\alpha_i})^2}{2\Lambda^{\mathrm{IN}2}_{\alpha_i}} \right\} \tag{S117}$$

$$\rho^{\mathrm{OUT}}_{\alpha_i} = \frac{1}{\sqrt{2\pi}\Lambda^{\mathrm{OUT}}_{\alpha_i}} \exp\left\{ -\frac{(x - \mu^{\mathrm{OUT}}_{\alpha_i})^2}{2\Lambda^{\mathrm{OUT}2}_{\alpha_i}} \right\} \tag{S118}$$

So the total distribution of responses is

$$\rho_{\alpha_i} = \gamma_{\alpha_i} \rho^{\mathrm{IN}}_{\alpha_i} + (1 - \gamma_{\alpha_i}) \rho^{\mathrm{OUT}}_{\alpha_i} \tag{S119}$$

5.12 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

$$\chi_{\alpha_i,\alpha_j} = \frac{1}{N_{\alpha_i}} \sum_{i\in\alpha_i} \Big(\sum_{j\in\alpha_j} {R^0}_{ij} \Big) = U^{\intercal}_{\alpha_i} R^0 V_{\alpha_j} \tag{S120}$$

Inserting (S64) into the above expression we obtain:

$$\begin{aligned} \chi = U^{\intercal} R V =& U^{\intercal}(F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^{\intercal} F)V & \text{(S121)}\\ =& \mathbf{f}' - \mathbf{f}'\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\mathbf{f}' & \text{(S122)}\\ =& \left({\mathbf{f}'}^{-1} - \omega\right)^{-1} & \text{(S123)} \end{aligned}$$

Which is the definition of χ as in Eq. (S7). We also need to show that the variance vanishes for large N. Writing $B_{ij} = \frac{1}{N^2_{\eta'}}\delta_{\alpha_i\alpha_j}\delta_{\alpha_i\eta'}$, and inserting it and Eq. (S81) into Eq. (S60) we find that :

$$\Lambda_{\eta\eta'} = \frac{\sigma^2}{N} \frac{\left(\sum_{\alpha_i} S^2_{\alpha_i\eta'} N_{\alpha_i}\right)\left(\sum_{\alpha_i} \frac{S^2_{\alpha_i\eta'}}{N_{\alpha_i}}\right)}{1 - \frac{\sigma^2}{N}\left(\sum_{\alpha_i} N_{\alpha_i} {f'_{\alpha_i}}^2 + \sum_{\alpha_i,\alpha_j} S^2_{\alpha_i,\alpha_j} \frac{N_{\alpha_i}}{N_{\alpha_j}} - \sum_{\alpha_i} {f'_{\alpha_i}}^2\right)} \tag{S124}$$

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 γ

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 $\tilde{u}^{(k)} = \frac{1}{N_k}(\underbrace{0,\ldots,0}_{\sum_{l=1}^{k-1} N_l}, \underbrace{1,\ldots,1}_{\alpha_k N_k}, \underbrace{0,\ldots,0}_{N-\sum_{l=1}^{k} N_l})$ and similarly for $\tilde{v}$.

$$\chi^{\boldsymbol{\gamma}}_{\alpha_i,\alpha_j} = \frac{1}{\gamma_{\alpha_i} N_{\alpha_i}} \sum_{i\in\mathscr{P}_{\alpha_i}} \Big(\sum_{j\in\mathscr{P}_{\alpha_j}} {R^0}_{ij} \Big) = \frac{\tilde{U}^{\intercal}_{\alpha_i}}{\gamma_{\alpha_i}} R^0 \tilde{V}_{\alpha_j} = \tilde{U}^{\intercal}_{\alpha_i} R^0 \tilde{V}_{\alpha_j} \tag{S125}$$

Before we had $U^{\intercal}FV = \mathbf{f}'$. 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

$$\begin{aligned}
\chi^{\boldsymbol{\gamma}} = \tilde{U}^\intercal R \tilde{V} =& \tilde{U}^\intercal (F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^\intercal F)\tilde{V} && \text{(S126)}\\
=& \mathbf{f}' - \mathbf{f}'\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\mathbf{f}'\boldsymbol{\gamma} && \text{(S127)}\\
=& \mathbf{f}' + \mathbf{f}'\omega\left(\mathbf{f}'^{-1} - \omega\right)^{-1}\boldsymbol{\gamma} && \text{(S128)}\\
=& \mathbf{f}' + \mathbf{f}'\omega\chi\boldsymbol{\gamma} && \text{(S129)}\\
& && \text{(S130)}
\end{aligned}$$

We notice that the *PP* element of this is

$$\begin{aligned}
\chi^{\boldsymbol{\gamma}}_{\alpha_i\alpha_i} =& f'_{\alpha_i} + f'_{\alpha_i}[\omega\chi]_{\alpha_i\alpha_i}\gamma_{\alpha_i} && \text{(S131)}\\
=& f'_{\alpha_i} + f'_{\alpha_i}(\frac{\chi_{\alpha_i\alpha_i}}{f'_{\alpha_i}} - 1)\gamma_{\alpha_i} && \text{(S132)}\\
=& f'_{\alpha_i} + \gamma_{\alpha_i}\chi_{\alpha_i\alpha_i} - f'_{\alpha_i}\gamma_{\alpha_i} && \text{(S133)}\\
& && \text{(S134)}
\end{aligned}$$

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 as Eq. (S74).

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 HD model
N	Number of neurons in the HD 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 HD system
r	Activity, r^α is the activity in population α

R^0	Linear response of the HD 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
HD	high-dimensional (i.e. N dimensional) model, with 4 populations
HFP	Homogeneous fixed point
LD	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. ISSN 08966273. doi: 10.1016/j.neuron.2009.03.028.

2. 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. ISSN 10974199. doi: 10.1016/j.neuron.2018.01.045.

3. 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. ISSN 10974199. doi: 10.1016/j.neuron.2017.06.019.

4. 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. ISSN 2050084X. doi: 10.7554/eLife.54875.

5. Hillel Adesnik. Synaptic Mechanisms of Feature Coding in the Visual Cortex of Awake Mice. *Neuron*, 95(5): 1147–1159.e4, 2017. ISSN 10974199. doi: 10.1016/j.neuron.2017.08.014.

6. Yashar Ahmadian and Kenneth D. Miller. What is the dynamical regime of cerebral cortex? (March), 2019.

7. Misha Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L Mcnaughton. Paradoxical effects of inhibitory interneurons. *The Journal of Neuroscience*, 17(11):4382–4388, 1997. ISSN 0270-6474.

8. 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. ISSN 2050084X. doi: 10.7554/eLife.49967.

9. 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. doi: 10.1101/2020.06.15.148114.

10. Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, 2016. ISSN 10974199. doi: 10.1016/j.neuron.2016.06.033.

11. 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. *Nature neuroscience*, 16 (8):1068–76, 2013. ISSN 1546-1726. doi: 10.1038/nn.3446.

12. 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. ISSN 10974199. doi: 10.1016/j.neuron.2016.02.037.

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

14. 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. ISBN 1111111111. doi: 10.1371/journal.pcbi.1005576.

15. Jonathan Kadmon and Haim Sompolinsky. Transition to chaos in random neuronal networks. *Physical Review X*, 5(4):1–28, 2015. ISSN 21603308. doi: 10.1103/PhysRevX.5.041030.

16. Tanguy Cabana and Jonathan Touboul. Large Deviations, Dynamics and Phase Transitions in Large Stochastic and Disordered Neural Networks. *Journal of Statistical Physics*, 153(2):211–269, 2013. ISSN 00224715. doi: 10.1007/s10955-013-0818-5.

17. A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk. On the Distribution of Firing Rates in Networks of Cortical Neurons. *Journal of Neuroscience*, 31(45):16217–16226, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.1677-11.2011.

18. Yashar Ahmadian, Francesco Fumarola, and Kenneth D. Miller. Properties of networks with partially structured and partially random connectivity. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 91(1):1–36, 2015. ISSN 15502376. doi: 10.1103/PhysRevE.91.012820.

19. 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. ISSN 10974199. doi: 10.1016/j.neuron.2016.02.037.

20. Julia Veit, Richard Hakim, Monika P. Jadi, Terrence J. Sejnowski, and Hillel Adesnik. Cortical gamma band synchronization through somatostatin interneurons. *Nature Neuroscience*, 20(7):951–959, 2017. ISSN 15461726. doi: 10.1038/nn.4562.

21. Kenneth D Miller and Agostina Palmigiano. Generalized paradoxical effects in excitatory / inhibitory networks. *BioarXiv*, pages 1–10.

22. Yashar Ahmadian, Daniel B. Rubin, and Kenneth D Miller. Analysis of the stabilized supralinear network. *Neural Computation*, 25(8):1994–2037, 2013. ISSN 1530888X. doi: 10.1162/NECO.

23. Christopher M. Bishop. *Pattern Recognition and Machine Learning*, volume 27. 2006. ISBN 0-387-31073-8. doi: 0-387-31073-8.

24. 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. ISSN 14764687. doi: 10.1038/s41586-020-2319-4.

25. 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. ISSN 10974199. doi: 10.1016/j.neuroN.2014.12.026.

26. 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. doi: 10.1101/2020.01.31.929166.

27. 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. ISSN 00280836. doi: 10.1038/nature12676.

28. 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. ISSN 10974172. doi: 10.1016/j.cell.2014.01.050.

29. Leena E. Williams and Anthony Holtmaat. Higher-Order Thalamocortical Inputs Gate Synaptic Long-Term Potentiation via Disinhibition. *Neuron*, 101(1):91–102.e4, 2019. ISSN 10974199. doi: 10.1016/j.neuron.2018.10.049.

30. Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *Journal of Neurophysiology*, 115(3):1399–1409, 2016. ISSN 0022-3077. doi: 10.1152/jn.00732.2015.

31. Terence Tao. Outliers in the spectrum of iid matrices with bounded rank perturbations. *Probability Theory and Related Fields*, 155(1-2):231–263, 2013. ISSN 01788051. doi: 10.1007/s00440-011-0397-9.

32. 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. *The Journal of Neuroscience*, 37(49):12050–12067, 2017. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0963-17.2017.

33. 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.

34. 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. ISSN 00280836. doi: 10.1038/nature11526.

35. Robert Rosenbaum and Brent Doiron. Balanced networks of spiking neurons with spatially dependent recurrent connections. *Physical Review X*, 4(2):021039, may 2014. ISSN 21603308. doi: 10.1103/PhysRevX.4.021039.

36. 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.

37. Y. Fu, J. M. Tucciarone, J. S. Espinosa, N. Sheng, D. P. Darcy, R. A. Nicoll, Z. J. Huang, and M. P. Stryker. A cortical circuit for gain control by behavioral state. *Cell*, 156:1139–1152, 2014.

38. 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. ISSN 08966273. doi: 10.1016/j.neuron.2018.04.017.

39. 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. doi: 10.1101/837567.

40. 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. ISSN 08966273. doi: 10.1016/j.neuron.2010.08.004.

41. 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. ISSN 10974199. doi: 10.1016/j.neuron.2016.12.028.

42. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2005. ISBN 9780471241959. doi: 10.1002/047174882X.

43. Charles L. Lawson and Richard J. Hanson. *Solving least squares problems*, volume 15. Society for Industrial and Applied Mathematics (SIAM), 1995. ISBN 0-89871-356-0. doi: 978-0898713565.

44. 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. ISSN 10974199. doi: 10.1016/j.neuron.2018.03.037.

45. Xavier Didelot, Richard G. Everitt, Adam M. Johansen, and Daniel J. Lawson. Likelihood-free estimation of model evidence. *Bayesian Analysis*, 6(1):49–76, 2011. ISSN 19360975. doi: 10.1214/11-BA602.

46. Donghui Chen and Robert J. Plemmons. Nonnegativity constraints in numerical analysis. *The Birth of Numerical Analysis*, pages 109–140, 2009. doi: 10.1142/9789812836267_0008.

47. 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. *Nature Neuroscience*, 21(6):851–859, jun 2018. ISSN 1097-6256. doi: 10.1038/s41593-018-0143-z.