Common rules underlying optogenetic and behavioral modulation of responses in multi-cell-type V1 circuits

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

[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]Max Planck Institute for Brain Research, Frankfurt, Germany
[4]Biophysics Graduate Group, University of California, Berkeley, United States
[5]Department of Molecular and Cell Biology, University of California, Berkeley, United States
[6]The Helen Wills Neuroscience Institute
[*]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

The visual cortex receives non-sensory inputs containing behavioral and brain state information. Here we propose a parallel between optogenetic and behavioral modulations of activity and characterize their impact on cell-type-specific V1 processing under a common theoretical framework. We infer cell-type-specific circuitry from large-scale V1 recordings and demonstrate that, given strong recurrent excitation, the cell-type-specific responses imply key aspects of the known connectivity. In the inferred models, parvalbumin-expressing (PV), but not other, interneurons have responses to perturbations that we show theoretically imply that their activity stabilizes the circuit. We infer inputs that explain locomotion-induced changes in firing rates and find that, contrary to hypotheses of simple disinhibition, locomotory drive to VIP cells and to SOM cells largely cancel, with enhancement of excitatory-cell visual responses likely due to direct locomotory drive to them. We show that this SOM/VIP cancellation is a property emerging from V1 connectivity structure.

Introduction

The visual cortex of the mouse represents a variety of different signals beyond visual information. Eye position (Meyer et al., 2018; Parker et al., 2022), face (Stringer et al., 2019) and head (Bouvier et al., 2020; Vélez-Fort et al., 2018) movements, and locomotion (Ayaz et al., 2013; Dipoppa et al., 2018; Fu et al., 2014; Keller et al., 2012; Niell and Stryker, 2010) are represented together with visual information in the cortical activity of mouse V1. The integration of locomotory and visual signals is at least partially mediated by the actions of specific interneuronal cell types (Dipoppa et al., 2018; Fu et al., 2014). Three types of GABA-ergic cells – parvalbumin- (PV), somatostatin- (SOM), and vasoactive-intestinal-peptide- (VIP) expressing cells – form a microcircuit characterized by highly specific connectivity (Campagnola et al., 2022; Karnani et al., 2016; Pfeffer et al., 2013), that can simultaneously stabilize and loosely balance strong recurrent excitation (Adesnik, 2017; Ahmadian and Miller, 2021; Ozeki et al., 2009), gate top-down signals (Zhang et al., 2014, 2016), and control the gain of the pyramidal cell population (Pi et al., 2013). In particular, two of these types, SOM and VIP, engage in competitive dynamics whose outcome directly regulates pyramidal cell activity and underlies modulation by contextual (Adesnik et al., 2012; Keller et al., 2020a,b; Mossing et al., 2021; Schnabel et al., 2018; Veit et al., 2017) and behavioral (Ayaz et al., 2013; Dipoppa et al., 2018; Fu et al., 2014; Niell and Stryker, 2010) state. Nevertheless, the mechanisms used by this circuit to simultaneously implement the multiple computations it performs, the dynamical signatures of those computations, and their linkage to the circuit's connectivity are incompletely understood.

Here we propose that optogenetic and behavioral modulations of activity may share common features, and describe them by a common perturbative framework. First, we develop a program for inferring from data recurrent circuit models that capture the full distribution across neurons of each cell type's responses. We apply this framework to calcium recordings of responses to visual stimuli of pyramidal (E) cells and the above three inhibitory cell types in mouse V1 layers 2/3, in stationary (non-locomoting) mice, and infer statistical descriptions of the recurrent circuitry that capture the distributions of responses of these cell types. The models we infer from data exhibit key aspects of the structure of the connectivity found in the mouse visual system – namely weak or no connections from VIP to any cell type except SOM or from SOM to itself (Karnani et al., 2016; Pfeffer et al., 2013) – provided that the recurrent excitation is strong, suggesting that this structure is needed to generate the observed responses given strong recurrent excitation. Second, we develop a theoretical framework for quantitatively predicting the distributions of responses of each cell type in model circuits to patterned, full or partial optogenetic perturbations. We show that if the mean response of a given cell-type population to a perturbation of that cell type is "paradoxical" – a negative response to positive stimulation – then that population is stabilizing the circuit. We predict that PV but not SOM interneurons show such "paradoxical" responses, and do so across stimulus conditions. This is consistent with convergent experimental (Veit et al., 2017) and theoretical (Bos et al., 2020) arguments that PV interneurons, but not SOM interneurons, play a major role in circuit stabilization. Furthermore, the fraction of PV cells that respond paradoxically is found to depend non-monotonically on the fraction of PV cells that are stimulated, with the turnaround point determined by the fraction of PV cells that are necessary to stabilize the circuit. This *fractional paradoxical effect* opens a new avenue for experimental inquiry.

Finally, under the hypothesis that locomotion acts by adding inputs of different strengths to the different cell types in our inferred circuits, we leverage our theoretical framework to infer the locomotion-induced inputs that can account for the cell-type specific distributions of response changes observed with locomotion. We find that these response changes require that locomotion induces strong inputs to both VIP and to SOM, whose effects on non-VIP cells largely cancel. This cancellation explains the lack of locomotory response of E cells in the absence of visual stimulation, despite the disinhibitory effect expected from the locomotory activation of VIP cells (Dipoppa et al., 2018; Pakan et al., 2016). We show that this cancellation of the effects of VIP and SOM stimulation is a general property of the inferred circuit structure, and is reliably seen at the level of single-cell

responses as well as at the population level. Our results suggest that locomotion-induced increases in pyramidal cell activity in the presence of visual inputs is induced by a locomotion-induced increase in visual drive to E cells, consistent with the fact that acetylcholine release, which is induced by locomotion (Fu et al., 2014), increases the gain of thalamocortical synapses (Gil et al., 1997; Hsieh et al., 2000). This is a first example of a more general approach to inferring inputs underlying behavioral modulations of neural responses.

Results

Inference of large scale models from multi-cell-type data recover V1 connectivity structure

In order to investigate the mechanisms that underlie responses to behavioral and optogenetic perturbations in mouse V1, we build large-scale, cell-type-specific models by inferring the model parameters from data. We consider the response to visual stimuli of varying contrasts in neurons of layer 2/3 of primary visual cortex (V1) of awake, head-fixed mice. Specifically, we study the responses, as assayed by calcium imaging, 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 contrasts.

Our large-scale model is a fully recurrent network composed of four populations representing the four cell types, with many units of each cell type (Fig. 1**a**). Each unit has a rectified power-law input-output function (Ahmadian et al., 2013) and receives a baseline input to account for the spontaneous activity, while feed-forward inputs carrying visual information only target E and PV cells. In order to find parameters for this model, we begin by first inferring the parameters of a simplified recurrent circuit model with only four units, one unit per population, each representing the mean activity of its cell type's population in the case in which there is no heterogeneity among the cells within each population (see Fig. 1**b**, fitting pipeline). From the measured means and standard errors of the responses of each cell type to each contrast, we sample millions of sets of surrogate curves of mean response vs. contrast for the four cell types, and use non-negative least squares to fit a model to each surrogate set (following Renner et al. (2020), see Fig. S1 and Methods section 3.2). We then select, from these millions of possible models, some hundreds of data-compatible 4-unit models for which the network steady-states provide the best fit.

We then fit the large-scale model. Its parameters are the means and variances of the baseline inputs to all cell types, of the contrast-driven inputs to E and PV, and of the connectivities from each cell type to itself and to each of the other cell types. To fit this model, we use the parameters of the hundreds of fitted 4-unit models as initial, or *seed* parameters for the means, and search over priors to find the variances (see Methods section 4.4). To fit this model, we leverage two facts. First, in large-scale circuit models in which neurons have a power-law input-output function, there is an explicit expression describing the distributions of activity that the population of each cell type produces (Roxin et al., 2011) given the distribution of its total inputs (recurrent and external). Second, for a fixed set of parameters, the distributions of activities the model will produce can be computed self-consistently through mean-field theory (Cabana and Touboul, 2013; Kadmon and Sompolinsky, 2015). From these two sets of analytic expressions, we create an explicit error function that quantifies how different the measured distributions of activity of all cell types at all contrasts are from those distributions produced by the large-scale model with a given set of parameters (see Methods sections 4 and 4.4 for a full description of the method). We construct a family of large scale models that match well the experimentally measured distributions of responses of all of the cell types (see below and Fig. S1) by choosing those that minimize that error function (Fig. 1**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 tails of the distributions seen in calcium data. These results do not depend on the exact shape of the distribution of weights as long as these distributions are bounded. Remarkably, even models with log-normally distributed synaptic weights reproduce

the data similarly well when using the exact same statistics as the models shown here (Fig. S1).

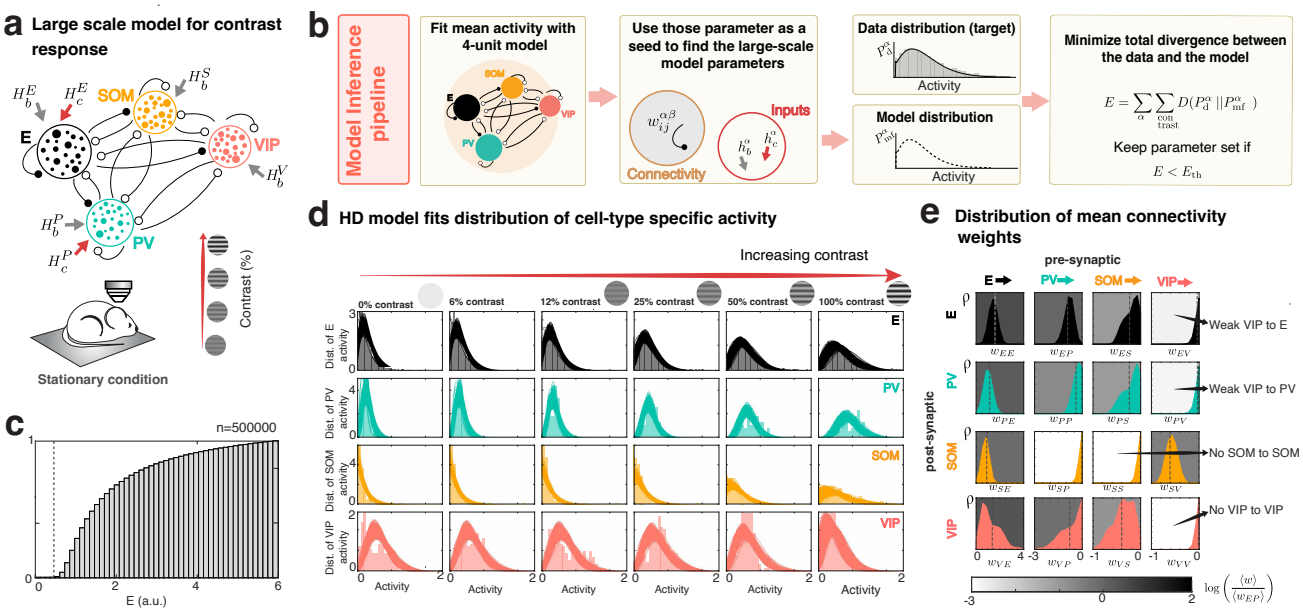

Figure 1 **Large scale model of mouse V1 a)** Large scale model composed of pyramidal (E, black), PV (turquoise), SOM (orange) and VIP (pink) populations. Inputs to each cell type are composed of a spontaneous activity baseline H_b and a stimulus related current, H_c modeling the feed-forward inputs from layer 4 targeting only E and PV (see also Fig. S1). **b)** Inference pipeline. Left: Model of four populations (see also (Renner et al., 2020)) describing only the mean activity. This model is fitted to the mean activity of the data (see Fig. S1 for details). Middle and middle right: we generate large-scale models by sampling from prior distributions of parameters given by the 4 unit model fits, computing the distributions of responses produced by the large-scale model with those parameters and finally comparing the model distributions we obtain, to the data distributions using an error function. Right: The error is given by the sum of the Kullback-Leibler divergences of the distributions given by the model (P_{mf}) and the data (P_{d}) for all cell types and contrast values, which can be found explicitly. We build a suitable family of models by only accepting errors smaller than a threshold of 0.5 (top 0.005%) (in other words we perform approximate Bayesian computation (Sisson et al., 2018)). **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 section 4.4 for details). **d)** 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 absolute value of the mean of each distribution. As in experiments (see Fig.S1), the models are found to have, at most, very weak recurrent SOM and VIP connections and connections from VIP to E and PV.

We required the model to have strong recurrent excitation (strong enough to be an inhibition-stabilized network, or ISN, described further below, see Fig. S1**c-d**). Beyond that, this optimization takes as sole input the response data and uses no prior information on the synaptic structure. It is not obvious that any meaningful synaptic structure should be recoverable from such a procedure. Yet, across models that fit the response distributions, the structure of the inferred connectivity matrices has a striking resemblance to that reported experimentally (Fig. 1**e**, see also S1**b**). In particular, recurrent connections within the SOM population and the VIP population were weak or absent, and VIP interneurons had weak or absent connections to all other cell types except SOM interneurons; these are both robustly reported features of the connectivity of mouse V1 L2/3 (Campagnola et al., 2022; Karnani et al., 2016; Pfeffer et al., 2013).

PV stabilizes the cortical circuit and is the only cell-type with paradoxical response

Cortical circuits have strong recurrent excitation that is stabilized and loosely balanced by inhibition (Adesnik, 2017; Ahmadian and Miller, 2021; Kato et al., 2017; Moore et al., 2018; Ozeki et al., 2009; Sanzeni et al., 2020). In models of one excitatory and only one inhibitory population, inhibitory stabilization implies that an increase in the input drive to the inhibitory population will elicit a "paradoxical" *decrease* of the inhibitory steady-state activity, along with a decrease of the excitatory activity (Ozeki et al., 2009; Tsodyks et al., 1997). In models with multiple cell types, if a perturbation only to GABAergic cells elicits a decrease in excitatory-cell activity along with a paradoxical decrease in the inhibition received by E cells – the latter would necessarily be true if there was a decrease in the activity of all GABAergic cells that project to excitatory cells – then the circuit is an ISN (Litwin-Kumar et al., 2016; Rubin et al., 2015). 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.

What then are the conditions that need to be satisfied for a set of cells to respond paradoxically to stimulation? Answering this question will lead us to an understanding of how stabilization of strong recurrent excitation is distributed across the different inhibitory cell types. We show theoretically, that paradoxical response of a given cell-type population implies that the circuit without that population is unstable (see Fig. 2). In other words, if a cell-type responds paradoxically, then if we would clamp the activity of that cell-type, not allowing it to change to track the activity of the rest of the circuit, a small perturbation to the circuit would make the activity grow or collapse unchecked. Furthermore, we find that because of the structure of the connectivity, a paradoxical response of SOM directly implies that the E-PV sub-circuit is unstable. A summary of the results derived in this section are: we find that i) the mean response of PV to its own stimulation is paradoxical in almost all the models that fit the data across stimulus conditions, and that ii) almost all models without PV are unstable, while iii) SOM and VIP don't have a mean paradoxical response, and iv) sub-circuits without SOM and/or VIP are always stable. These results are consistent with experimental work showing that strong perturbations to PV destabilize the dynamics in V1 while strong perturbations to SOM do not (Veit et al., 2017).

In order to demonstrate these results theoretically, we establish a formalism that allows us to study response to (small) perturbations of arbitrary fractions of cells of each cell type, and how the response is related to the stability of the network without that cell type, in large-scale models (Fig. 2**a**(I)). The theory relies on a simplifying approximation, which is that the gain (the slope of the input/output function) of each cell is the same for all cells in a given cell-type population. Under this approximation, which we refer to as the *homogeneous fixed point approximation* (hereinafter HFA, see Methods section 5.2), and by building on recent random matrix theory results (Ahmadian et al., 2015) (see Methods section 5), we are able to obtain explicit expressions for the behavior of the mean (See. Eq. S84 and section 5.5 of the Methods) and the variance (Eq.S108 in section 5.6 of the Methods) of the distributions of responses in each population to either full or partial, and either homogeneous or heterogeneous, perturbations. Importantly, we find that under the HFA: i) the mean responses to a perturbation of all neurons of a given cell type are equal to the responses in a corresponding 4-unit model to a perturbation of that cell type, and ii) when that mean response is paradoxical, the sub-circuit without that cell type is unstable (Fig. 2a(III); see Methods sections 5.11, 6). Thus, we can use the HFA as an analytic bridge between the simple 4-unit model and the fully nonlinear large-scale model.

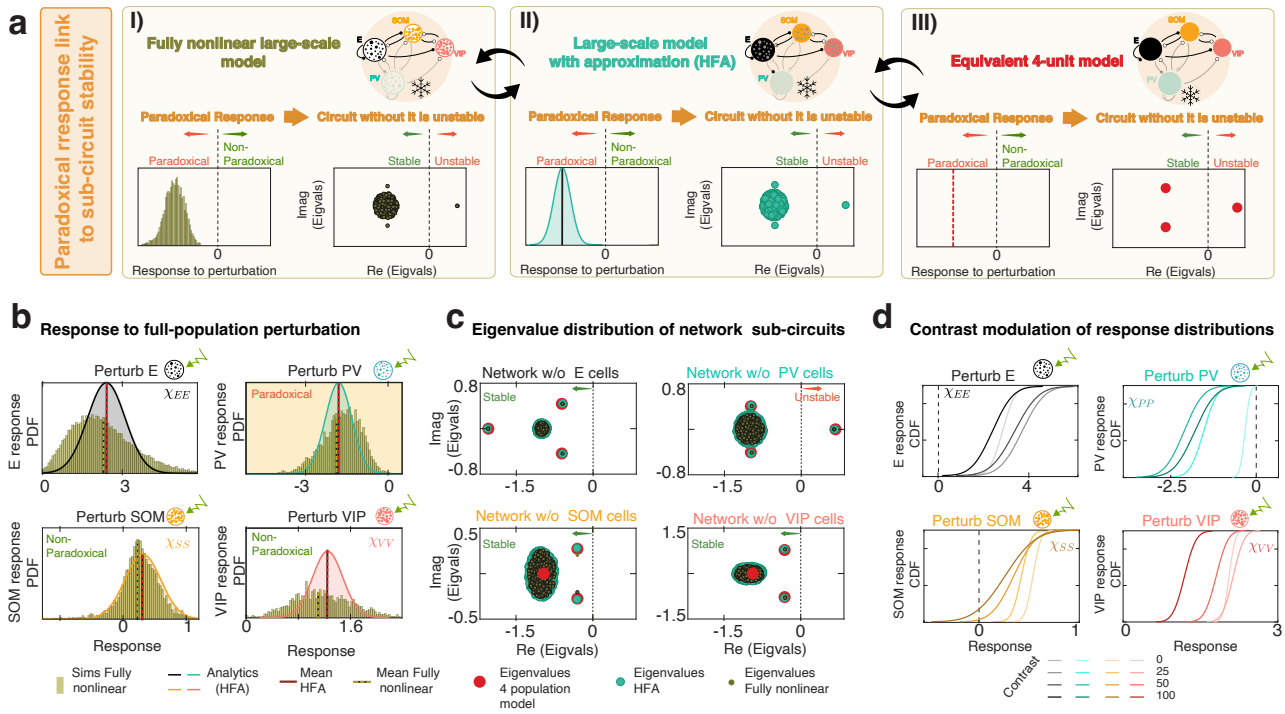

Figure 2 **Paradoxical response and circuit stabilization of inferred models.** **a)** Graphic summary of the relation between stability and paradoxical responses. (I): We observe a paradoxical PV response in the fully nonlinear system seems to imply instability of the network without that cell type (one eigenvalue has positive real part). (II) We find that under the HFA (see text), the distribution of responses has mean (red line) given by the response of the 4-unit model, and a variance that can be computed exactly. Furthermore, the outlier eigenvalues in this case are given by the 4-unit model. (III) In the 4-unit model, a negative (paradoxical) response of a cell type to a positive perturbation (red dashed line, left) provably implies that the circuit without that cell type is unstable (see Eq. 1), which means that at least one eigenvalue of the Jacobian of the circuit without that cell type has a positive real part (right). **b)** Distribution of responses of each cell type to its own full-population perturbations for a stimulus of 100% contrast, for the fully nonlinear large-scale system (green, dashed green line is its mean), and the analytical distributions obtained under the HFA (colored histograms, red dashed line is its mean). The responses of E, VIP, and most SOM cells are not paradoxical, while all cells in the PV population respond paradoxically to PV stimulation. When VIP projects only to SOM, which is true in most our models, the lack of a paradoxical response in SOM indicates that the E-PV circuit is stable. **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. Only the sub-circuit without the PV population is unstable (has a positive eigenvalue). The eigenvalues of the fully nonlinear system overlap with those of the system under the HFA, which coincide with those of the 4-unit model. **d)** Cumulative distribution of responses (HFA) to full-population perturbation in the presence of a visual stimulus for varying stimulus contrast. These results hold in the family of models that fit the data, see Fig. S2.

To state these results mathematically, we first consider the steady-state responses to a small perturbation of a 4-unit circuit, initially at a fixed point (a steady-state response to a fixed input). The responses are described by the linear response matrix χ. The response of cell type α to a perturbative input to itself is given by the diagonal entry $\chi_{\alpha\alpha}$ multiplied by the input. When $\chi_{\alpha\alpha}$ is negative, this response is paradoxical – a negative response to a positive input. $\chi_{\alpha\alpha}$ in turn can be written as a function of the Jacobian, J_α, of the sub-circuit without the cell-type α, at the fixed point (see also Methods sections 3.3 and 3.3.2):

$$\chi_{\alpha\alpha} = k_\alpha \det(-J_\alpha) \qquad \text{with } k_\alpha > 0 \qquad \alpha = \{E, P, S, V\} \tag{1}$$

The Jacobian determines the stability of the sub-circuit. For a fixed point of a system to be stable, all eigenvalues of the negative Jacobian have to have positive real part (corresponding to all eigenvalues of the Jacobian having negative real part), and thus its determinant – the product of its eigenvalues – must be positive. Thus, if the response of cell type α at a given fixed point is paradoxical ($\chi_{\alpha\alpha} < 0$), then, at that fixed point, the sub-circuit without that cell type is unstable ($\det(-J_\alpha) < 0$), *i.e.* at least one eigenvalue of the Jacobian of the sub-circuit has positive real part (Fig. 2a). This insight is a simple generalization of the two-population ISN network – there, when the sub-circuit without I cells was unstable, then I cells responded paradoxically – that links cell-type specific paradoxical response to sub-circuit stability more generally (Fig. 2**a**, left). We furthermore find that when VIP projects only to SOM (as is at least approximately true in mouse V1), then a paradoxical response in the SOM population indicates not only that the E-PV-VIP sub-circuit is unstable, but that the E-PV sub-circuit by itself is unstable (see Eq. S14). These simple findings in the model with 4 units apply more generally: we find that in the large scale model, under the HFA, whenever the mean response of a cell-type to a perturbation of the entire population is paradoxical, the system without that population will be unstable (Fig. 2**a**, middle, see theoretical argument for this in Methods section 6), and we have found this to hold in numerical simulations of the fully nonlinear system (Fig. 2**a**, right).

We find (as illustrated for an example model at 100% contrast, Figs. 2b-c) that in all three models (the 4-unit model, the large-scale model with the HFA, and the fully nonlinear large-scale model) the mean response is only paradoxical (negative) for PV cells, and that therefore the system without PV is unstable (it has one positive eigenvalue, top-right panel). The PV responses remain paradoxical, and that of the other cell types non-paradoxical, at all stimulus contrasts (Figure 2**d**). These results hold in almost all the models that fit the data (see S2**a-b**), suggesting a fundamental role of PV in circuit stabilization in the family of models we derived from the data. Furthermore, we find that, in the great majority of solutions, no other interneuron type has a mean paradoxical response, and that sub-circuits without them (*i.e.,* without SOM and/or VIP) are always stable (Fig. S2**a-d**).

In summary, 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. 2**b,c**). This is consistent with experimental work showing that suppression of PV activity destabilizes V1 while suppression of SOM activity does not (Veit et al., 2017)

Fractional paradoxical effect

Motivated by the fact that in optogenetic experiments with viral injection only a fraction of cells is perturbed, we next investigated how many PV cells need to be stimulated to observe the paradoxical effect in the population of PV cells (Fig. 3). We find, that, when targeting only a fraction of the cells in the circuit, there is a range in which the more strongly we probe the system to find the paradoxical effect, the less likely we are to find it. Specifically, we find that there is a range in which the more PV cells we perturb (which experimentally can be done in a controlled manner via holographic optogenetics, Adesnik and Abdeladim, 2021) the less likely we are to observe a paradoxical effect by measuring the activity of the population. We name this *the fractional paradoxical effect*. Furthermore we find, that under the HFA, the perturbed cells only have a mean paradoxical response when the sub-circuit without the perturbed cells becomes unstable, generalizing the above findings to partial perturbations.

Mathematically, we establish these results as follows. Under the HFA, the distribution of responses of the entire population to partial perturbations is bimodal, given by a mixture of two Gaussians (Fig. 3**a**, turquoise). One Gaussian distribution corresponds to the perturbed cells (red dashed line), while the other corresponds to the unperturbed population (green dashed line, see Eq. S113).For the PV population, the unperturbed cells always

show a negative mean response to the perturbation. 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. 3**b**) and the sub-circuit without those perturbed cells is stable (all the eigenvalues of the sub-circuit's Jacobian have negative real part (Fig. 3**a**, bottom left; Fig. S3). As the fraction of perturbed PV cells increases, the mean response of the perturbed population moves towards negative values. However, since the fraction of perturbed PV cells is increasing, then while their response remains positive, the fraction of negatively responding PV cells is decreasing. With continued growth of the perturbed fraction, their response ultimately changes sign, as the sub-circuit without the perturbed cells loses stability (the maximum eigenvalue of the Jacobian of the sub-circuit without the perturbed cells becomes positive, Fig. 3**a**, bottom right). This gives rise to a curious phenomenon: with increasing fraction of PV cells perturbed, the fraction of PV cells responding negatively shows non-monotonic behavior (Fig. 3**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. This *fractional paradoxical effect* extends the concept of critical fraction (the fraction of inhibitory neurons that must be perturbed in order to see a paradoxical effect, developed by Sadeh et al. (2017)), to the case in which the neurons have heterogeneous connectivity. The lower panels of Figure 3**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 1. Intriguingly, in the models that fit the data, PV has a fractional paradoxical response at all contrasts. Recent experiments (Sanzeni et al., 2020) 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 affects 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 this property is independent of the stimulus contrast.

To understand theoretically the relationship between fractional paradoxical response and sub-circuit stability, we considered a 5-unit network (Fig. 3**c**, top right), with two PV populations, a perturbed one (red) and an unperturbed one (green). For each fraction of perturbed cells, the connectivity of this network is chosen such that its response to perturbations is mathematically equivalent to the mean response to that partial perturbation in the large-scale system under the HFA (see Sec. 5.9). As predicted by Eq. 1, whenever the response of the perturbed PV population in the 5-unit model system becomes negative (paradoxical), the sub-system composed of all of the non-perturbed populations loses stability; and, as in Fig. 2, in both the HFA and the fully nonlinear system, the outlier eigenvalues (those with largest real part) of the Jacobian of the sub-circuit of non-perturbed cells will be very close to those in the 5D system (Fig. 3**d**, top). Thus, when the mean response of the perturbed PV population in the HFA becomes negative, the sub-system without those perturbed neurons will become unstable (see Figs. 3**d**, S3; and Methods Sect. 5.9).

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 for small numbers of perturbed cells, and will lose stability only after a critical fraction of cells are stimulated. Note 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 (Fig. 3**d**, bottom, upper right quadrant). 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, and thus without evoking a paradoxical response (Eq. 1; see Miller and Palmigiano, 2020 for a clarification). In sum, although instability does not imply a paradoxical response, observation that a set of perturbed cells shows a mean paradoxical response tells us that the circuit without the perturbed cells is unstable.

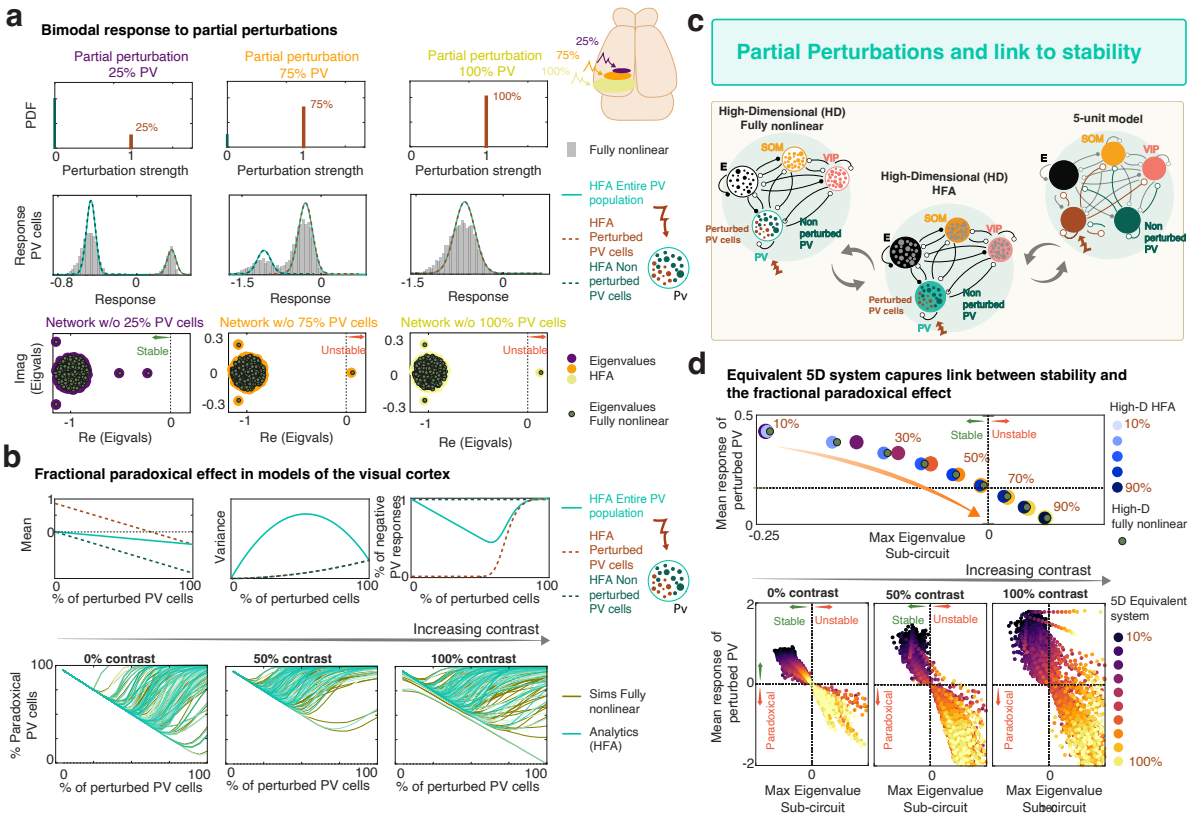

Figure 3 **Fractional paradoxical effect and link to sub-circuit stability: a)** Top: distribution of perturbation strengths, when perturbing 25% (left), 75% (middle) and a 100% (right) of the PV population. Middle row: Partial perturbations result in a bimodal distribution of responses. Simulations (gray) are well described by the mixture of two Gaussians predicted by the HFA (turquoise, solid line). 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. 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 sub-circuit changes sign, indicating that the non-perturbed sub-circuit has become unstable, 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 responses (dashed green) 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 non-perturbed (dashed green) and perturbed (dashed red) monotonically increase, while the distance between the two peaks, and thus the variance associated with this distance, monotonically decreases (not shown), resulting in a non-monotonic change in variance of the full distribution (turquoise). Top right: Fraction of negative responses as a function of the fraction of stimulated cells shows a non-monotonic dependence (turquoise), 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 for different data-fit models (green) are in good agreement with calculations from the HFA for the same models (turquoise). **c)** Linking response and stability across models. A fully nonlinear system can be linked to a large-scale system of lower complexity via the HFA. The mean response to a partial perturbation in the HFA can be mapped to the response to perturbation of one PV sub-population in a 5-unit system with two PV sub-populations, a perturbed one (red) and an unperturbed one (green) see, Eq. (S124). **d)** Top: Mean response of the perturbed PV population as a function of the outlier eigenvalue with largest real part of the Jacobian of the unperturbed subcircuit, for different fractions of perturbed PV cells (numbers in graph indicate this fraction). This is shown for the fully nonlinear system (green), HFA (blue colors) and the equivalent 5D system (purple/orange palette , legend in bottom panel). The mean responses become negative when the maximum eigenvalue real part crosses zero, indicating instability of the non-perturbed sub-circuit. Bottom: Mean response of the perturbed PV population, for varying fractions of perturbed cells (indicated by colors), as a function of the value of the outlier eigenvalue in the equivalent 5D model obtained from different models that fit the data, for three different values of the contrast (see also Methods section 5.9).

Theoretical foundation of VIP and SOM mediated disinhibition

VIP and SOM interneurons engage in competitive dynamics that regulate pyramidal cell activity via direct inhibition and disinhibition (Tremblay et al., 2016). We find theoretically that, when VIP projects only to SOM (as is at least approximately true in mouse V1, Campagnola et al., 2022; Karnani et al., 2016; Pfeffer et al., 2013, and as we observe in the connectivity matrix we inferred from data, Fig. 1), the response of E, PV and SOM to a change in the SOM input is negatively proportional to the response to a change in the input to VIP, with the same proportionality constant for all three cell types. VIP response shows the same proportionality, plus an additional direct VIP response to VIP drive. Thus, simultaneous drive to SOM and to VIP can cancel in their effects on E, PV, and SOM cells, causing only an increase in VIP activity. While drive to SOM or to VIP causes changes of opposite sign, it is theoretically possible for either one to have the positive sign, *i.e.* to evoke disinhibition. This depends on which is dominant, SOM inhibition of E cells or of PV cells, since the latter disinhibits E cells. This choice corresponds to a difference observed between layers in mouse V1. In layer 2/3, disinhibition is mediated by VIP: When VIP activity increases, it inhibits SOM, releasing E cells from its inhibition (Pi et al., 2013) and increasing E activity (Attinger et al., 2017; Fu et al., 2014; Karnani et al., 2016; Keller et al., 2020a; Mossing et al., 2021; Pi et al., 2013; Williams and Holtmaat, 2019). In layer 4, in contrast, SOM interneurons evoke disinhibition, because SOM projects preferentially to PV instead of E in this layer (Xu et al., 2013). Finally, although we can mathematically show this negative proportionality of responses only for the mean responses of each cell type, we find in simulations that this proportionality of response also holds true at the single cell level.

To establish these results, we start with the analysis of the 4-unit model, for a generic connectivity that satisfies the condition that VIP projects only to SOM but is otherwise arbitrary. We find mathematically (see Methods section 3.4) that the response of any given cell type to a VIP perturbation is negatively proportional to the response to a SOM perturbation, with a common proportionality constant across the four cell types (Fig. 4**a**, see also Eq. S15). In the case of VIP, there is an additional response component determined by its own gain. More concretely, if f'_V is the gain of VIP at a particular steady-state configuration (that is, the slope of its input/output function f_V, which determines its firing rate at the steady state) and ω_{SV} is the synaptic weight from VIP to SOM, then

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

where $\delta_{\alpha V}$ is the Kronecker delta ($\delta_{\alpha V} = 1$, $\alpha = V$; $= 0$, otherwise).

We refer to these equalities as *Disinhibition Symmetries* (Fig. 4**a-c**). Because the mean response of a cell-type population to the perturbation of all neurons in another cell-type population is (under the HFA) given by the response of the 4-unit circuit (see Methods section 5.11), these symmetries also apply to the mean of the distributions in the large scale system under the HFA. Figure 4**b** shows the response distributions of an example model from Figure 1**d**, to perturbations of the entire SOM and VIP populations. The distributions obtained in simulations of the fully non-linear system (green histogram) are in good agreement with the analytic results under the HFA, and exhibit the disinhibition symmetries. Mean responses across the 360 models that best fit the data all accord well with the disinhibition symmetries (Fig. 4**c**), both in models under the HFA (colored circles) and the fully nonlinear networks (triangles), as expected since the data-compatible models all exhibit only weak connections from VIP to cell types other than SOM (Fig. 1**e**). All of these models predict the VIP-mediated disinhibition observed in L2/3.

Finally, we show that, surprisingly, the mathematical understanding provided by the disinhibition symmetries at the mean level, also apply at the single cell level. Figure 4**d** shows, for a single fully-nonlinear network example, that the response of each cell to a perturbation of the full SOM population is negatively proportional to the response of that cell to a perturbation of the full VIP population, with a proportionality factor approximated as

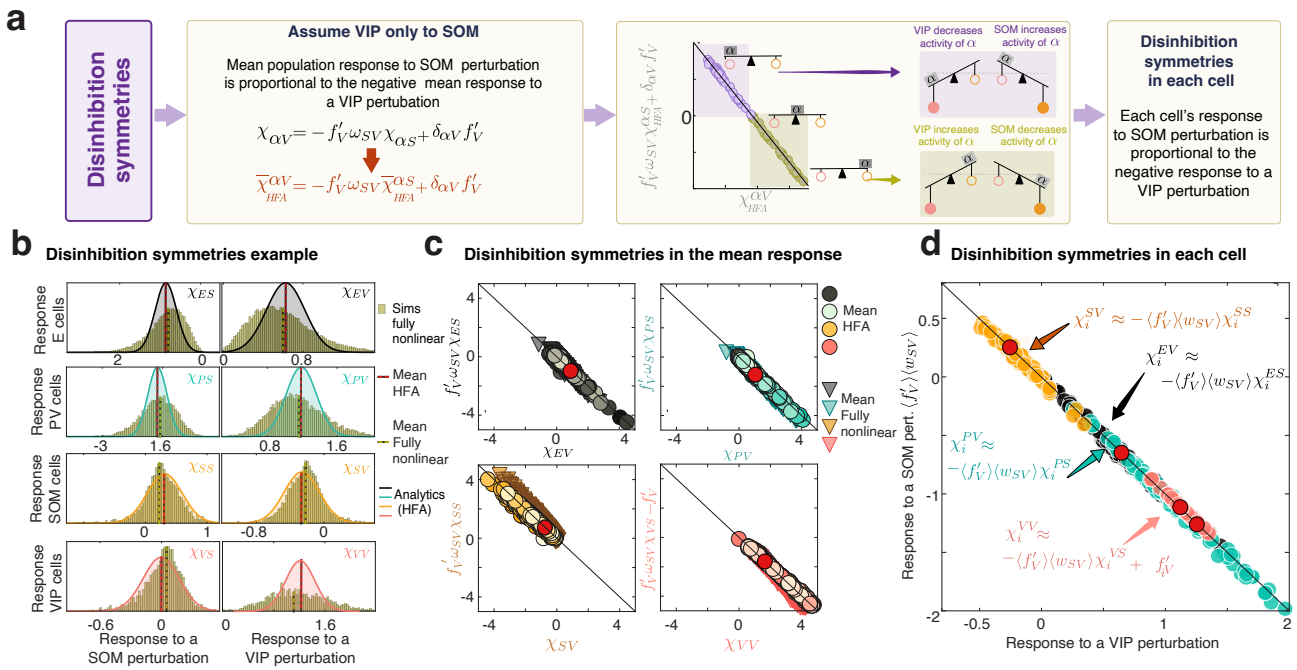

Figure 4 Disinhibition symmetries. **a)** Left: A parameter-independent relation holds true between the response of pyramidal cells (E), PV interneurons (P), SOM interneurons (S), and VIP interneurons (V) to perturbation of SOM and VIP under the assumption of VIP projecting only to SOM. Middle: The response of a cell type α to VIP perturbation, vs. the response to a SOM perturbation multiplied by the coefficient in Eq. (2), are negatively proportional. 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). **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 large-scale system, while colored histograms and corresponding lines are the analytical results obtained using the HFA. Note the similarity of mean response in the fully nonlinear model (black dashed lines) and the HFA (red dashed lines). **c)** Opposite and proportional mean responses in the fully nonlinear system ($\bar{\chi}_{\text{HD}}$, triangles) and under the HFA ($\bar{\chi}_{\text{HFA}}$, circles) to perturbations of SOM and VIP, for E (top left), PV (top right), SOM (bottom left) and VIP (bottom right), for the best 360 model fits. Because these data-compatible connectivities have only small values of connections weights from VIP to E, PV and SOM (Fig. 1**e**), they all show this symmetry between responses to SOM and VIP perturbations. These best-fit models all display VIP-mediated disinhibition and SOM-mediated inhibition of E and PV cells (compare E and PV responses to green quadrant in **a**). **d)** Disinhibition symmetries at the single neuron level. Responses of randomly chosen single cells (for the best model fit) of type E (black), PV (turquoise) and SOM (orange) and VIP (pink) cells to a perturbation of the entire VIP population ($\chi_i^{\alpha V}$), vs. the responses of those same cells to a perturbation of the full SOM population ($\chi_i^{\alpha S}$) multiplied by the factor: $\langle f_V' \rangle \langle w_{SV} \rangle$ where the average is over the population (and for VIP cells, adjusted for self-excitation by VIP stimulation). The response symmetries hold at the single cell level in the fully nonlinear system.

the average connectivity from SOM to VIP ($\langle w_{SV} \rangle$) and the average gain of VIP cells ($\langle f_V' \rangle$). These responses are perfectly anti-correlated (see also Fig. S4 for log-normally distributed weights and connectivities in which VIP projects to all cell types).

The disinhibition symmetries 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 opposing, proportional effects on E, PV and SOM of small VIP versus SOM perturbations, with the same proportionality constant for all, are strong predictions resulting from our analysis.

VIP-SOM cancellation underlies behavioral modulations of activity

We next asked, if behavioral modulations of activity act as an extra input –a perturbation– to the local circuit, what are the characteristics of those inputs? We addressed this inverse problem – asking what is the distribution of inputs targeting each cell-type population that would lead to observed changes in activity – to study the mechanisms underlying a known behavioral perturbation: locomotion. We examined the difference between each cell's activity in the locomotion and the stationary condition in the experimental data (Fig. 5**a**). We found, consistent with previous findings (Fu et al., 2014; Pakan et al., 2016), that in the absence of visual stimulation the mean change in activity is only positive for VIP cells (Fig. 5**a** and **c**); that of E cells is not. This fact leads to a puzzle: why does the large change in VIP activity not cause an increase in E activity via VIP-mediated disinhibition as widely reported in layer 2/3, even in the absence of visual stimulation, (Mossing et al., 2021)?

To answer this question we expanded the framework described in previous sections to include heterogeneous perturbations (see Methods section 5.10). The obtained expressions, which depend on the stimulus contrast only implicitly via the population's gain, allowed us to map the statistics (mean and variance) of a given perturbation, to the statistics (mean and variance) of the changes in the recurrent-network activity induced by that perturbation. By fitting these expressions to the changes in activity induced by locomotion (Fig. 5**c**, left panels), we could infer, for each model in the family of models that fit the stationary data (see Fig. 1), the distribution of inputs to each cell type population that would generate distributions of responses in the circuit identical to those produced by locomotion at each value of the stimulus contrast (Fig. 5**c**, right panels).

We found that, to account for the locomotion-induced activity changes, the inputs to VIP must have a large mean and variance, as discussed in Fu et al. (2014). Surprisingly, we found that inputs to SOM also had to be strong. This does not imply that SOM activity increases with locomotion. Instead, in the absence of visual stimulation, the effects of the direct locomotory input to SOM are largely canceled by inhibitory inputs coming from VIP. This is an example of the disinhibition symmetries described in the previous section. It is this cancellation at the level of the SOM output that leads to the zero mean change in pyramidal cell activity in the absence of visual stimulation.

We further found that this cancellation holds when the network receives visual inputs, across all contrasts. This is true even though the mean change in activity induced by locomotion and its variance increase with stimulus contrast for all cell-types (Fig. 5**c**, left panels). According to equation 2, for this cancellation to happen, the mean locomotory input to SOM I_S and to VIP I_V should fulfill (see Eq. S30).

$$\frac{I_S}{I_V} = f'_{\mathrm{V}} \omega^{\mathrm{SV}} \tag{3}$$

Plotting $\frac{I_S}{I_V}$ vs. $f'_{\mathrm{V}} \omega^{\mathrm{SV}}$ shows a scatter plot of this relation for all models that fit the data for all stimulus contrasts. We find that the inferred mean inputs cluster around the diagonal line, indicating that the inputs to SOM and to VIP largely cancel in their effects (Figure 5**d**, top). Directly examining the changes in mean E activity induced by the locomotory inputs to SOM and to VIP we find that these changes largely cancel, with only a few models showing weak disinhibition (Figure 5**d**, bottom). To infer the locomotory inputs, we assumed that the mean and standard deviation of the inputs varied as the square-root of stimulus contrast, but these results hold equally if instead they are assumed to vary linearly in, or to be independent of, the stimulus contrast (Fig. S5**b**).

If the effects of locomotory inputs to VIP and to SOM cancel, why does locomotion increase E firing rates in the presence of a visual stimulus? According to the inferred inputs, this is due to a small direct locomotory input to E cells, which increases with contrast (Fig. 5**c**, 3rd panel).

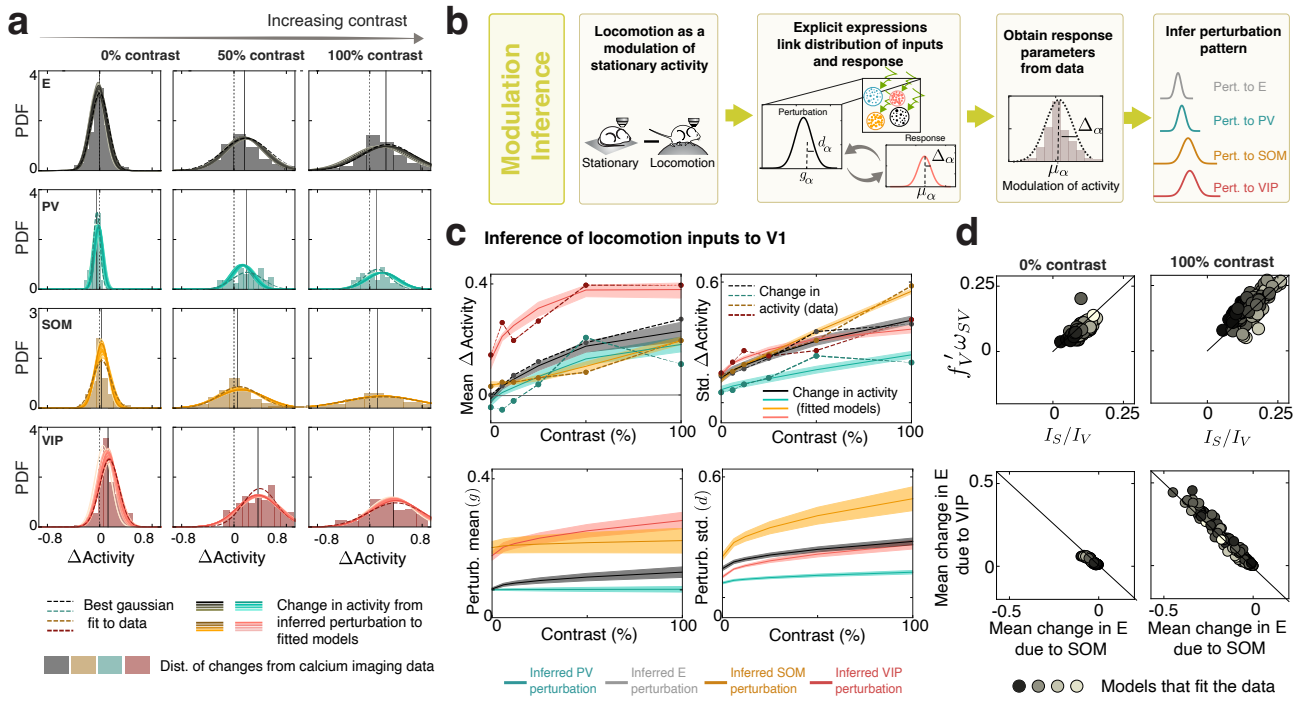

Figure 5 **Inferring locomotory-related inputs reveals VIP-SOM cancellation.** **a)** Distribution of Δ activity (difference between each cell's activity in the locomotion and the stationary condition) for each cell type and different values of the stimulus contrast. Dashed lines indicate best Gaussian fits. Solid lines are fits from the explicit expressions (see Eqs. S126 and S129) **b)** Scheme of how to infer the cell-type specific perturbations (Gaussians of mean g_α and standard deviation d_α for $\alpha = \{\text{E,PV,SOM,VIP}\}$) that give rise to the distributions of Δ activity (with mean μ_α and standard deviation Δ_α). **c)** Left two panels: Mean 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). Full lines are the mean over all model fits. Shaded areas indicate the 95 percentile of the distribution of those changes in activity over the model fits, obtained as described in **b**, for the family of models that fit the stationary data. Right two panels: The mean (g) and the standard deviation (d) of the inferred perturbations. The shaded area indicates the 5th through 95th percentile of the distribution of the values of the mean and std of the inputs over different models. **d)** The ratio of the mean inferred inputs to SOM and to VIP I_S/I_V are proportional to the mean VIP gain (f'_V) times the mean strength of the connection from VIP to SOM ((ω_V), different for each model). This is the case even when the stimulus contrast is large and the mean change in E due to locomotion no longer vanishes. Bottom: Change in E activity due locomotory inputs to VIP Vs due to locomotory inputs to SOM.

Discussion

Experimental work has shown that multiple non-sensory signals are represented in areas previously thought to be highly specialized for sensory processing. Behavioral states such as thirst (Campagnola et al., 2022), hunger (Padamsey et al., 2022), or fear (Li et al., 2019) as well as multiple motor signals (Attinger et al., 2017; Bouvier et al., 2020; Dipoppa et al., 2018; Keller et al., 2012; Liska et al., 2022; Meyer et al., 2018; Miura and Scanziani, 2021; Musall et al., 2019; Niell and Stryker, 2010; Pakan et al., 2016; Stringer et al., 2019; Vélez-Fort et al., 2018), are represented in the sensory cortex of the mouse. Here we investigate the effect of two non-sensory signals, optogenetic perturbations and locomotion, in modulating cell-type-specific activity in the visual cortex of the mouse in the presence of different visual stimuli. We proposed a parallel between these modulations of activity, and developed a common theoretical framework to study responses to such cell-type-specific, full or partial perturbations in large-scale heterogeneous circuits. This framework is composed of three stages i) inference of heterogeneous, cell-type-specific recurrent model circuits from distributions of sensory responses; ii) characterization of responses to perturbations of the model circuits; and iii) inference of cell-type-specific

inputs that can account for the distributions of activity changes induced by locomotion or other behavioral states.

We leveraged the mathematical tractability of large-scale recurrent network models to find a family of circuit models that match the cell-type-specific distributions of visual responses observed across stimuli of varying contrasts. These models are specified by the statistics (means and variance) of the connectivity between each pair of cell types and of the stimulus-specific inputs to each cell type. We found that, provided the network is constrained to be inhibition-stabilized (i.e., has strong, potentially destabilizing recurrent excitation, Ozeki et al., 2009; Tsodyks et al., 1997, as observed across mouse cortex, Sanzeni et al., 2020), this method robustly recovers a specific connectivity structure found in mouse V1: few or no connections of VIP cells to cell types other than SOM, or of SOM cells to themselves (Campagnola et al., 2022; Karnani et al., 2016; Pfeffer et al., 2013) (see Figs. 1 and S1).

We found that these models, despite spanning a broad distribution of parameters, respond similarly to perturbations, and that cell-type-specific responses reveal mechanisms of circuit stabilization. Specifically, we established that, if stimulation of all of, or a subpopulation of, a cell type evokes a "paradoxical" response – meaning that the stimulated cells show a negative mean steady state response to positive stimulation – then the circuit without the stimulated cells is unstable. In our data-fit models, PV cells, and not any other cell type, are circuit stabilizers and show a paradoxical response, consistent with other theoretical results (Bos et al., 2020) and with experimental findings that suppression of PV but not SOM cells leads to epileptiform activity (Veit et al., 2017). Furthermore, we predict a “fractional paradoxical effect" for partial perturbations of PV cells (or any cell type showing a paradoxical response to full-population perturbation): with increasing proportion of the population stimulated, the fraction of the population responding negatively changes non-monotonically. The fraction of negative responses decreases as a function of the fraction of stimulated cells until shortly before the stimulated population is big enough that the network is unstable without it, meaning that suppression of that population leads to epileptiform activity. This prediction can be directly tested by simultaneously monitoring (via calcium imaging) and perturbing (via holographic optogenetic stimulation) an increasing number of PV cells.

In layer 2/3 of the visual cortex of the mouse, SOM and VIP interneurons engage in competitive dynamics that underlie multiple forms of contextual modulations of activity (Keller et al., 2020a,b; Mossing et al., 2021; Veit et al., 2017). We showed that, because VIP projects almost exclusively to SOM cells (Campagnola et al., 2022; Karnani et al., 2016; Pfeffer et al., 2013), the effect of a VIP perturbation on the mean activity of other cell types is proportional to, but opposite in sign to, their mean response to SOM perturbation, with the same proportionality constant for all cell types (VIP cells show the same proportional response, plus an additional response to their own stimulation). This means that the effect of simultaneous inputs to both SOM and VIP can largely cancel for all cell types except VIP interneurons. Furthermore, while our theory established this cancellation at the level of the mean responses of each cell type, we found in simulations that this cancellation reliably characterized responses at the single-cell level. Remarkably, this cancellation is the mechanism that we found underlies locomotory modulations of activity.

We used our perturbative framework to infer the distribution of locomotion-induced inputs to the four cell types from their distributions of locomotion-induced response changes. We found that, to account for these response distributions of all four cell types, such a cancellation of VIP and SOM influences must be taking place. This cancellation accounts for the large increase in VIP firing rates while producing little change in other types. As a result, locomotory increases in excitatory-cell firing rates are inferred to be primarily due to direct locomotory drive to excitatory cells, which is consistent with the finding that acetylcholine, which is released in V1 in response to locomotion (Fu et al., 2014), increases the gain of thalamocortical inputs (Gil et al., 1997; Hsieh et al., 2000) (further discussed below). This contrasts with the hypothesis that locomotion-driven increases in E-cell visual responses are due to disinhibition caused by activation of VIP cells (Fu et al., 2014).

Note that experimental findings that activation of VIP cells increased excitatory firing rates, while suppressing VIP activity prevented the effects of locomotion (Fu et al., 2014), do not contradict our prediction: either manipulation imbalances the VIP-SOM competition, activating or suppressing the disinhibitory circuit and thus altering excitatory cell activity, yet those effects may not be driving the normal effects of locomotion.

Our findings are in line with experimental and theoretical work suggesting that locomotory inputs in V1 influence all of the cell types. These locomotion-related signals have multiple sources, each with different projection patterns to the different cell types. Cholinergic inputs from the basal forebrain, which have been shown to at least partially underlie the modulations of activity induced by locomotion (Fu et al., 2014; Lee et al., 2014, but see Polack et al., 2013, target many or all interneuronal types (Arroyo et al., 2014; Colangelo et al., 2019; Fanselow et al., 2008; Kawaguchi, 1997; Kuchibhotla et al., 2017; but see Yaeger et al., 2019). Noradrenergic input from the locus coeruleus, which targets multiple cell types (Kawaguchi and Shindou, 1998), also is implicated in locomotory response changes (Polack et al., 2013). Locomotory inputs can potentially also arrive to V1 from motor cortex, which sends direct projections to V1, mostly strongly targeting excitatory and PV cells (Leinweber et al., 2017). Finally, as locomotory inputs are already present in the lateral geniculate nucleus (Busse, 2020; Erisken et al., 2014), and locomotion-induced acetylcholine release increases the gain of thalamocortical synapses (Gil et al., 1997; Hsieh et al., 2000), we expect that visual drive to L4, and thus to L2/3, carries mixed locomotory and visual signals. In particular, an increase in gain of thalamocortical synapses could induce a multiplicative scaling of visual responses, consistent with our finding that the locomotory input to E cells should be increasing with stimulus contrast (note that we found that locomotory inputs targeting SOM and VIP can be independent of the visual stimulus, see Fig. S5).

The family of models we obtained describe the responses to small visual stimuli, averaged over stimulus orientations. This allowed us to ignore the distributions of neurons across space and preferred orientation and the dependencies of connectivity on differences in these features (*e.g.*, Cossell et al., 2015; Ko et al., 2011; Rossi et al., 2020). Future work will be needed to extend the framework proposed here to embrace space, orientation, or other features on which connectivity may depend. It is in principle possible that larger visual stimuli, which strongly engage SOM cells (Adesnik et al., 2012), would lead SOM to play a role in stabilizing the circuit. However, optogenetic suppression of SOM cells in the presence of a large visual stimulus do not show signatures of instability (Veit et al., 2017), and we expect that the predictions proposed here are robust to changes in stimulus configuration.

Finally, the models developed here could be applied to study the differences in the effects of optogenetic and behavioral modulation across different cortical areas. Auditory cortex, for example, is an area that shares multiple organizing principles with V1, such as a VIP→SOM→E disinhibitory circuit (Bigelow et al., 2019), inhibitory stabilization (Kato et al., 2017; Moore et al., 2018) and SOM engagement in suppression by a stimulus surround (Lakunina et al., 2020). Nevertheless, the effect of locomotion in A1 is to suppress excitatory cells, rather than to enhance their firing as in V1 (Zhou et al., 2014), and this effect is not mediated by VIP (Bigelow et al., 2019). It is conceivable that, instead of a different mechanism underlying behavioral modulations of activity across sensory areas, differences in the degree of cancellation between SOM and VIP projections to the pyramidal cell population could lead to the observed differences. More generally, our work provides a general framework to study both behavioral and optogenetic modulations of activity and how those are shaped by the responses of sensory stimuli.

Acknowledgements

A.P. would like to acknowledge the support of the Swartz Foundation Fellowship for Theory in Neuroscience 2019-4 and 2020-6. 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. FF was supported by RIKEN Center for Brain Science, Brain/MINDS from AMED under Grant No. JP20dm020700, and JSPS KAKENHI Grant No. JP18H05432. 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. This work was supported by the New York stem cell foundation. All authors would like to thank L. Abbott, B. Doiron, G. Handy, A.L. Kumar, and L. Mazzucato for useful feedback on earlier versions of 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.

References

H. 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. Publisher: Elsevier Inc.

H. Adesnik and L. Abdeladim. Probing neural codes with two-photon holographic optogenetics. *Nature Neuroscience*, 24(10):1356–1366, 2021. ISSN 15461726. doi: 10.1038/s41593-021-00902-9. Publisher: Springer US.

H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, and M. Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–230, 2012. ISSN 00280836. doi: 10.1038/nature11526. arXiv: 1011.1669v3 Publisher: Nature Publishing Group ISBN: 9788578110796.

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

Y. Ahmadian, D. B. Rubin, and K. D. Miller. Analysis of the stabilized supralinear network. *Neural Computation*, 25(8):1994–2037, 2013. ISSN 1530888X. doi: 10.1162/NECO. arXiv: 1309.2848v1 ISBN: 9781467312264.

Y. Ahmadian, F. Fumarola, and K. 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. arXiv: 1311.4672v2 ISBN: 9781467312264.

S. Arroyo, C. Bennett, and S. Hestrin. Nicotinic modulation of cortical circuits. *Front Neural Circuits*, 8:30, 2014.

A. Attinger, B. Wang, and G. B. Keller. Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex. *Cell*, 169(7):1291–1302.e14, 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.05.023. Publisher: Elsevier.

A. Ayaz, A. B. Saleem, M. L. Schölvinck, and M. Carandini. Locomotion controls spatial integration in mouse visual cortex. *Current Biology*, 23(10):890–894, 2013. ISSN 09609822. doi: 10.1016/j.cub.2013.04.012. ISBN: 1879-0445 (Electronic)\r0960-9822 (Linking).

J. Bigelow, R. J. Morrill, J. Dekloe, and A. R. Hasenstaub. Movement and VIP interneuron activation differentially modulate encoding in mouse auditory cortex. *eNeuro*, 6(5), 2019. ISSN 23732822. doi: 10.1523/ENEURO.0164-19.2019.

H. Bos, A.-M. Oswald, and B. 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.

G. Bouvier, Y. Senzai, and M. Scanziani. Head Movements Control the Activity of Primary Visual Cortex in a Luminance-Dependent Manner. *Neuron*, 108(3):500–511.e5, 2020. ISSN 10974199. doi: 10.1016/j.neuron.2020.07.004. Publisher: Elsevier Inc.

L. Busse. The influence of locomotion on sensory processing and its underlying neuronal circuits. *Neuroforum*, 24(1):A41–A51, 2020. ISSN 23637013. doi: 10.1515/nf-2017-A046.

T. Cabana and J. 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.

L. Campagnola, S. C. Seeman, T. Chartrand, L. Kim, A. Hoggarth, C. Gamlin, S. Ito, J. Trinh, P. Davoudian, C. Radaelli, M.-H. Kim, T. Hage, T. Braun, L. Alfiler, J. Andrade, P. Bohn, R. Dalley, A. Henry, S. Kebede, M. Alice, D. Sandman, G. Williams, R. Larsen, C. Teeter, T. L. Daigle, K. Berry, N. Dotson, R. Enstrom, M. Gorham, M. Hupp, S. Dingman Lee, K. Ngo, P. R. Nicovich, L. Potekhina, S. Ransford, A. Gary, J. Goldy, D. McMillen, T. Pham, M. Tieu, L. Siverts, M. Walker, C. Farrell, M. Schroedter, C. Slaughterbeck, C. Cobb, R. Ellenbogen, R. P. Gwinn, C. D. Keene, A. L. Ko, J. G. Ojemann, D. L. Silbergeld, D. Carey, T. Casper, K. Crichton, M. Clark, N. Dee, L. Ellingwood, J. Gloe, M. Kroll, J. Sulc, H. Tung, K. Wadhwani, K. Brouner, T. Egdorf, M. Maxwell, M. McGraw, C. A. Pom, A. Ruiz, J. Bomben, D. Feng, N. Hejazinia, S. Shi, A. Szafer, W. Wakeman, J. Phillips, A. Bernard, L. Esposito, F. D. D'Orazi, S. Sunkin, K. Smith, B. Tasic, A. Arkhipov, S. Sorensen, E. Lein, C. Koch, G. Murphy, H. Zeng, and T. Jarsky. Local connectivity and synaptic dynamics in mouse and human neocortex. *Science*, 375(6585):eabj5861, Mar. 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abj5861.

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

C. Colangelo, P. Shichkova, D. Keller, H. Markram, and S. Ramaswamy. Cellular, Synaptic and Network Effects of Acetylcholine in the Neocortex. *Front Neural Circuits*, 13:24, 2019.

L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403, 2015. ISSN 14764687. doi: 10.1038/nature14182.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2005. ISBN 978-0-471-24195-9. doi: 10.1002/047174882X. arXiv: ISBN 0-471-06259-6 Publication Title: Elements of Information Theory ISSN: 15579654.

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

M. Dipoppa, A. Ranson, M. Krumin, M. Pachitariu, M. Carandini, and K. 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. Publisher: Elsevier Inc. ISBN: 0896-6273.

S. Erisken, A. Vaiceliunaite, O. Jurjut, M. Fiorini, S. Katzner, and L. Busse. Effects of locomotion extend throughout the mouse early visual system. *Current Biology*, 24(24):2899–2907, 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.10.045.

E. E. Fanselow, K. A. Richardson, and B. W. Connors. Selective, state-dependent activation of somatostatin-expressing inhibitory interneurons in mouse neocortex. *J Neurophysiol*, 100(5):2640–2652, Nov 2008.

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(6):1139–1152, 2014. ISSN 10974172. doi: 10.1016/j.cell.2014.01.050. arXiv: NIHMS150003 ISBN: 1097-4172 (Electronic)\r0092-8674 (Linking).

Z. Gil, B. W. Connors, and Y. Amitai. Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, 19(3):679–686, 1997. ISSN 08966273. doi: 10.1016/S0896-6273(00)80380-3.

C. Y. Hsieh, S. J. Cruikshank, and R. Metherate. Differential modulation of auditory thalamocortical and intracortical synaptic transmission by cholinergic agonist. *Brain Res.*, 880(1-2):51–64, 2000. ISSN 00068993. doi: 10.1016/S0006-8993(00)02766-9.

J. Kadmon and H. Sompolinsky. Transition to chaos in random neuronal networks. *Physical Review X*, 5(4): 1–28, 2015. ISSN 21603308. doi: 10.1103/PhysRevX.5.041030. arXiv: 1508.06486 ISBN: 0044122374014.

M. M. Karnani, J. Jackson, I. Ayzenshtat, J. Tucciarone, K. Manoocheri, W. G. Snider, and R. 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. Publisher: Elsevier Inc.

H. K. Kato, S. K. Asinof, and J. 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. Publisher: Elsevier Inc.

Y. Kawaguchi. Selective cholinergic modulation of cortical GABAergic cell subtypes. *J Neurophysiol*, 78(3): 1743–1747, 1997.

Y. Kawaguchi and T. Shindou. Noradrenergic excitation and inhibition of GABAergic cell types in rat frontal cortex. *J Neurosci*, 18(17):6963–6976, 1998.

A. J. Keller, M. Dipoppa, M. M. Roth, M. S. Caudill, A. Ingrosso, K. D. Miller, and M. Scanziani. A Disinhibitory Circuit for Contextual Modulation in Primary Visual Cortex. *Neuron*, 108(6):1181–1193.e8, 2020a. ISSN 10974199. doi: 10.1016/j.neuron.2020.11.013. Publisher: Elsevier Inc.

A. J. Keller, M. M. Roth, and M. Scanziani. Feedback generates a second receptive field in neurons of the visual cortex. *Nature*, 582(7813):545–549, 2020b. ISSN 14764687. doi: 10.1038/s41586-020-2319-4. Publisher: Springer US.

G. B. Keller, T. Bonhoeffer, and M. Hübener. Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron*, 74(5):809–815, 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.03.040.

A. G. Khan and S. B. Hofer. Contextual signals in visual cortex. *Current Opinion in Neurobiology*, 52:131–138, 2018. ISSN 18736882. doi: 10.1016/j.conb.2018.05.003. Publisher: Elsevier Ltd.

H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011. ISSN 00280836. doi: 10.1038/nature09880.

K. V. Kuchibhotla, J. V. Gill, G. W. Lindsay, E. S. Papadoyannis, R. E. Field, T. A. H. Sten, K. D. Miller, and R. C. Froemke. Parallel processing by cortical inhibition enables context-dependent behavior. *Nat. Neurosci.*, 20(1):62–71, 2017. ISSN 15461726. doi: 10.1038/nn.4436.

A. A. Lakunina, M. B. Nardoci, Y. Ahmadian, and S. Jaramillo. Somatostatin-expressing interneurons in the auditory cortex mediate sustained suppression by spectral surround. *Journal of Neuroscience*, 40(18): 3564–3575, 2020. ISSN 15292401. doi: 10.1523/JNEUROSCI.1735-19.2020.

A. M. Lee, J. L. Hoy, A. Bonci, L. Wilbrecht, M. P. Stryker, and C. M. Niell. Identification of a brainstem circuit regulating visual cortical state in parallel with locomotion. *Neuron*, 83:455–466, 2014.

M. Leinweber, D. R. Ward, J. M. Sobczak, A. Attinger, and G. B. Keller. A Sensorimotor Circuit in Mouse Cortex for Visual Flow Predictions. *Neuron*, 95(6):1420–1432.e5, 2017. ISSN 10974199. doi: 10.1016/j.neuron.2017.08.036. Publisher: Elsevier Inc.

Z. Li, A. Yan, K. Guo, and W. Li. Fear-Related Signals in the Primary Visual Cortex. *Current Biology*, 29(23): 4078–4083.e2, Dec. 2019. ISSN 09609822. doi: 10.1016/j.cub.2019.09.063.

J. P. Liska, D. P. Rowley, T. T. K. Nguyen, J.-O. Muthmann, D. A. Butts, J. L. Yates, and A. C. Huk. Running modulates primate and rodent visual cortex via common mechanism but quantitatively distinct implementation. *bioRxiv*, page 2022.06.13.495712, 2022.

A. Litwin-Kumar, R. Rosenbaum, and B. 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. ISBN: 1522-1598 (Electronic)\r0022-3077 (Linking).

A. F. Meyer, J. Poort, J. O'Keefe, M. Sahani, and J. F. Linden. A Head-Mounted Camera System Integrates Detailed Behavioral Monitoring with Multichannel Electrophysiology in Freely Moving Mice. *Neuron*, 100 (1):46–60.e7, 2018. ISSN 10974199. doi: 10.1016/j.neuron.2018.09.020. Publisher: Elsevier Inc.

K. D. Miller and A. Palmigiano. Generalized paradoxical effects in excitatory / inhibitory networks. *BioarXiv*, pages 1–10, 2020.

S. K. Miura and M. Scanziani. Erasing motion: Scrambling direction selectivity in visual cortex during 1 saccades. *bioRxiv*, page 2021.03.30.437338, 2021.

A. K. Moore, A. P. Weible, T. S. Balmer, L. O. Trussell, and M. 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. Publisher: Elsevier Inc.

D. P. Mossing, J. Veit, A. Palmigiano, K. D. Miller, and H. Adesnik. Antagonistic inhibitory subnetworks control cooperation and competition across cortical space. *bioRxiv*, 2021. doi: 10.1101/2021.03.31.437953.

S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22(10):1677–1686, 2019. ISSN 15461726. doi: 10.1038/s41593-019-0502-4. Publisher: Springer US.

C. M. Niell and M. P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, 65(4):472–479, Feb. 2010. ISSN 08966273. doi: 10.1016/j.neuron.2010.01.033. Publisher: Elsevier Ltd ISBN: 1097-4199 (Electronic)\n0896-6273 (Linking).

H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster. Inhibitory Stabilization of the Cortical Network Underlies Visual Surround Suppression. *Neuron*, 62(4):578–592, 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.03.028. Publisher: Elsevier Ltd ISBN: 0896-6273.

Z. Padamsey, D. Katsanevaki, N. Dupuy, and N. L. Rochefort. Neocortex saves energy by reducing coding precision during food scarcity. *Neuron*, 110(2):280–296.e10, Jan. 2022. ISSN 08966273. doi: 10.1016/j.neuron.2021.10.024.

J. M. Pakan, S. C. Lowe, E. Dylda, S. W. Keemink, S. P. Currie, C. A. Coutts, and N. L. Rochefort. Behavioral-state modulation of inhibition is context-dependent and cell type specific in mouse visual cortex. *eLife*, 5(AUGUST):1–18, 2016. ISSN 2050084X. doi: 10.7554/eLife.14985.

P. R. L. Parker, E. T. T. Abe, E. S. P. Leonard, D. M. Martins, and C. M. Niell. Joint coding of visual input and eye/head position in V1 of freely moving mice. *bioRxiv*, page 2022.02.01.478733, 2022.

C. K. Pfeffer, M. Xue, M. He, Z. J. Huang, and M. 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. arXiv: NIHMS150003 Publisher: Nature Publishing Group ISBN: 1546-1726 (Electronic)\n1097-6256 (Linking).

H. J. Pi, B. Hangya, D. Kvitsiani, J. I. Sanders, Z. J. Huang, and A. Kepecs. Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477):521–524, 2013. ISSN 00280836. doi: 10.1038/nature12676.

P. O. Polack, J. Friedman, and P. Golshani. Cellular mechanisms of brain state-dependent gain modulation in visual cortex. *Nat. Neurosci.*, 16(9):1331–1339, sep 2013. ISSN 10976256. doi: 10.1038/nn.3464.

S. Renner, N. Kraynyukova, Y. Bauer, G. Born, A. H. Kotkat, X. Liu, M. Spacek, G. Tushev, T. Tchumatchenko, and L. Busse. Inference of network connectivity from responses to briefly flashed gratings in mouse V1 using a stabilized supralinear network model (SSN). *Bernstein Conference Abstract*, 16(1982):10–11, 2020. doi: 10.12751/nncn.bc2020.0111.

L. F. Rossi, K. D. Harris, and M. Carandini. Spatial connectivity matches direction selectivity in visual cortex. *Nature*, 588(7839):648–652, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2894-4.

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. ISBN: 1529-2401 (Electronic)\n0270-6474 (Linking).

D. B. Rubin, S. D. VanHooser, and K. D. Miller. The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2015. ISSN 10974199. doi: 10.1016/j.neuroN.2014.12.026. ISBN: 0896-6273.

S. Sadeh, R. A. Silver, T. Mrsic-Flogel, and D. R. 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.

A. Sanzeni, B. Akitake, H. C. Goldbach, C. E. Leedy, N. Brunel, and M. H. Histed. Inhibition stabilization is a widespread property of cortical networks. *eLife*, 9:1–39, 2020. ISSN 2050084X. doi: 10.7554/eLife.54875.

U. Schnabel, L. Kirchberger, E. van Beest, S. Mukherjee, A. Barsegyan, J. Lorteije, C. van der Togt, M. Self, and P. Roelfsema. Feedforward and feedback processing during figure-ground perception in mice. *bioRxiv*, (October):456459, 2018. doi: 10.1101/456459.

S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437), 2019. ISSN 10959203. doi: 10.1126/science.aav7893.

T. 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. arXiv: 1012.4818.

R. Tremblay, S. Lee, and B. 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. Publisher: Elsevier Inc. ISBN: 0896-6273.

M. Tsodyks, W. E. Skaggs, T. J. Sejnowski, and B. L. Mcnaughton. Paradoxical effects of inhibitory interneurons. *The Journal of Neuroscience*, 17(11):4382–4388, 1997. ISSN 0270-6474.

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

M. Vélez-Fort, E. F. Bracey, S. Keshavarzi, C. V. Rousseau, L. Cossell, S. C. Lenzi, M. Strom, and T. W. Margrie. A Circuit for Integration of Head- and Visual-Motion Signals in Layer 6 of Mouse Primary Visual Cortex. *Neuron*, 98(1):179–191.e6, 2018. ISSN 10974199. doi: 10.1016/j.neuron.2018.02.023.

L. E. Williams and A. 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. Publisher: Elsevier Inc.

H. Xu, H. Y. Jeong, R. Tremblay, and B. Rudy. Neocortical Somatostatin-Expressing GABAergic Interneurons Disinhibit the Thalamorecipient Layer 4. *Neuron*, 77(1):155–167, 2013. ISSN 08966273. doi: 10.1016/j.neuron.2012.11.004. arXiv: NIHMS150003 Publisher: Elsevier ISBN: 3143627344.

C. E. Yaeger, D. L. Ringach, and J. T. Trachtenberg. Neuromodulatory control of localized dendritic spiking in critical period cortex. *Nature*, 567(7746):100–104, 2019. ISSN 14764687. doi: 10.1038/s41586-019-0963-3.

S. Zhang, M. Xu, T. Kamigaki, J. P. H. Do, W. C. Chang, S. Jenvay, K. Miyamichi, L. Luo, and Y. Dan. Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, 345(6197): 660–665, 2014. ISSN 10959203. doi: 10.1126/science.1254126. arXiv: 1011.1669v3 ISBN: 1853467960.

S. Zhang, M. Xu, W. C. Chang, C. Ma, J. P. Hoang Do, D. Jeong, T. Lei, J. L. Fan, and Y. Dan. Organization of long-range inputs and outputs of frontal cortex for top-down control. *Nature Neuroscience*, 19(12):1733–1742, 2016. ISSN 15461726. doi: 10.1038/nn.4417.

M. Zhou, F. Liang, X. R. Xiong, L. Li, H. Li, Z. Xiao, H. W. Tao, and L. I. Zhang. Scaling down of balanced excitation and inhibition by active behavioral states in auditory cortex. *Nature Neuroscience*, 17(6):841–850, 2014. ISSN 15461726. doi: 10.1038/nn.3701.

Methods

1 Methods summary . . . 26

2 Data Collection and Analysis . . . 27

3 Low-dimensional circuit models (LD model or 4-unit model) . . . 28

3.1 Definition . . . 28

3.2 Fitting of mean activity . . . 28

3.3 Linear response and paradoxical effects . . . 29

3.4 Disinhibition symmetries (VIP projects only to SOM) . . . 31

3.5 Disinhibition symmetries underlying behavioral modulations of activity. . . . 32

3.6 Equivalent class for low-dimensional circuits . . . 32

4 High-dimensional circuit models . . . 32

4.1 Definition . . . 32

4.2 Set-up and mean field equations . . . 33

4.3 Mean field perturbation . . . 34

4.4 Data Fitting . . . 34

4.5 Equivalent class for high-dimensional circuits . . . 36

5 Analytical approach to linear response of disordered networks . . . 36

5.1 Set up . . . 36

5.2 Homogeneous fixed point approximation (HFA) . . . 37

5.3 General framework to compute the linear response in networks in the HFA . . . 37

5.4 Computation of the response matrix R^0 without disorder . . . 38

5.5 Response distribution to partial (homogeneous) perturbations: Mean term . . . 40

5.6 Response distribution to partial (homogeneous) perturbations: Variance term . . . 43

5.7 Response distribution to partial (homogeneous) perturbations: Full Distribution . . . 45

5.8 Simple description of the fractional paradoxical effect . . . 46

5.9 Fractional paradoxical effect and link to a 5D low-dimensional system. . . . 46

5.10 Response distribution to partial and non-homogeneous perturbations. 48

5.11 Link to the low-dimensional system linear response 49

6 Eigenvalue Spectrum of the Jacobian in the HFA 50

7 Numerics 51

7.1 Network simulations 51

7.2 Linear response 52

7.3 Inference of locomotion inputs 52

1 Methods summary

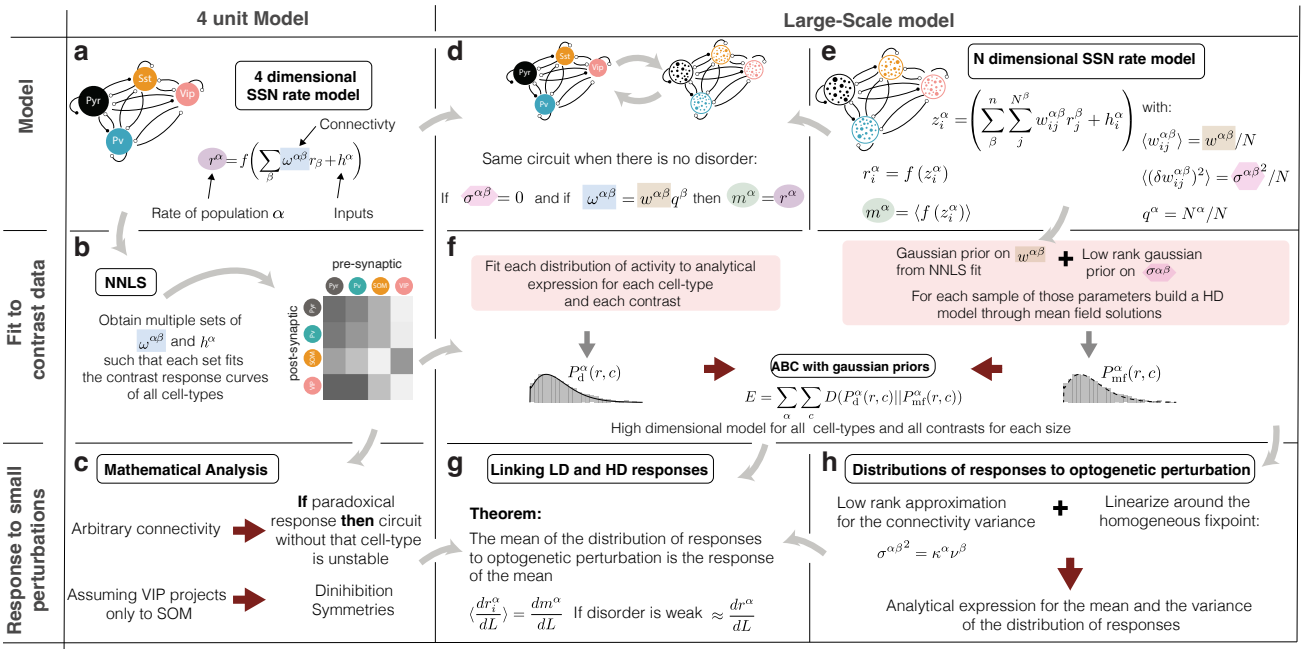

Figure 6 **Outline of this paper. a-b)** 4-unit circuit): circuit describing the mean activity of E, PV, SOM, and VIP cells when presented with stimuli of different contrasts. By using non-negative least squares (NNLS) we find the parameters to fit the contrast response of all four cell types. **c)** We find relations between stability and responses to optogenetic perturbations and, assuming that VIP only projects to SOM and SOM does not project to itself, find hidden structure in the response matrix. These findings are applied to the models that fit the data. **d-e)** Large-scale model: There are now many neurons of each cell type. When there is no disorder - all the connections of a given type (as defined by presynaptic and postsynaptic cell types) have the same strength, and all cells of a given type receive the same external input, the large-scale circuit describes the same dynamics as the 4-unit circuit described in (a), provided that the parameters are chosen appropriately. Inclusion of disorder changes the mean activity. **f)** We use Approximate Bayesian Computation, also known as ABC, to fit the high-dimensional system. Firstly, the models we use have analytical expressions for their distributions of activity. We fit such an expression to the measured distribution of activity of each cell type for each stimulus condition. Secondly, we build large-scale models with parameters sampled from a distributions with priors on the means obtained from the NNLS analysis. For each model, we use mean field theory to compute the resulting activity distributions. By minimizing the Kullback-Leibler divergence (Cover and Thomas, 2005) between these two sets of activity distributions (the one obtained from the data and the one obtained from the model family), we find the models that best approximate the activity distributions 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 and compute the entire distribution of responses to an arbitrary pattern of optogenetic stimulation.

We develop a two-stage program for the prediction of responses to perturbations of large-scale circuits with multiple inhibitory types (Fig. 6). In a *first stage*, we build large-scale models that account for the entire distribution of cell-type-specific responses observed in the data. For that, we start by fitting a 4-unit dynamical system (see Sec. 3.1) to the mean responses observed experimentally in the four cell types (excitatory(E), PV, SOM, VIP) in layer 2/3 of the mouse visual cortex, in response to stimulation by a small (5 degree diameter) visual stimulus of varying contrast (for data details see Sec. 2), while the animal is in a stationary condition. These fits are obtained by finding parameters that are a solution to a non-negative least squares (NNLS) problem (Renner et al., 2020) (see Sec. 3.2). These simple models allow for a mathematical understanding of the response to perturbations to different cell types, and to make parameter-independent predictions (see Secs. 3.3, 3.4 and Fig. 6**c**). Then, we use these 4-unit model fits as a parameter seed to build a family of large-scale

models (Fig. 6**d-e**, see Sec. 4) that captures the full distributios of activity for each cell type and stimulus contrast. To find the parameters of these models we leverage their mathematical tractability (see Sec. 4.4). First, as the distribution of activity has a tractable analytical form (Roxin et al., 2011) that depends on the mean and variance of the input currents to each cell-type population, we can obtain that mean and variance for each cell-type and stimulus condition by fitting that distribution to the data via maximum likelihood (Fig. 6**f**, left). This gives us the target mean and variance of the input currents for each cell type that the model needs to be able to self-consistently create. Given some model parameters (Kadmon and Sompolinsky, 2015) (*e.g.*, means and variances of connection strengths and external inputs), the means and variances of the input currents can be computed self-consistently through Mean-Field (MF) theory (Fig. 6**f**, right). Now, to find the parameters that make the model capture the data, we explicitly write the Kullback-Leibler divergence between the activity distributions that fits the data and the distributions that the model creates. By sampling parameters from distributions created from the 4-unit fits, and by choosing a suitable threshold on this distance (0.5) (Didelot et al., 2011), we find large-scale models whose distributions of activity across cell types and stimulus contrasts reproduce those observed experimentally (Fig. 6**f**, middle). These fits make predictions about the mean and variance of the connection strengths between neurons of any two given cell types, and of the external inputs each cell-type receives as a function of the visual stimulus.

In a second stage, we build on random matrix theory results (Ahmadian et al., 2015) (see Sec. 5.1 and 5.3) to analytically compute, under the homogeneous fixed point approximation (HFA, i.e. that all neurons in each cell type's population have the same gain, see Fig. 6**h** and Sec. 5.2), the distribution of neuronal responses to patterned perturbations. We show that under this approximation, the mean response to an excitatory perturbation to all cells in a cell-type population in the large-scale system has a mean given by the response of the system without disorder to that same perturbation (see Sec. 5.5). Furthermore, this response is itself mathematically identical to the response of an equivalent 4-unit system (see Sec. 5.4 and Fig. 6**g**). This means that the parameter-independent predictions of the response to perturbations in the low dimensional case (4-unit model), carry over to the mean response to a full population perturbation in the large-scale system. Furthermore, under the HFA, partial perturbations lead to a bimodal distribution of responses given by a mixture of Gaussians (see Sec. 5.7), whose mean (see Sec. 5.5) and variance (see Sec. 5.6) can be computed exactly.

We rely on a Theorem by T. Tao (Tao, 2013) to link the stability of the network subsystems and the paradoxical response of a given cell-type. First we show that in the 4-unit system, if the sign of the response of a cell-type to its own perturbation is negative, that implies that the circuit without that cell-type is unstable. Under the HFA, what can be shown is that the bulk eigenvalue distribution lies in a ball centered at -1, and contains only stable eigenvalues (those with real part < 0); and that, by Tao's theorem, there are the outlier eigenvalues are given by the eigenvalues of the 4-unit system (see Sec. 6), which may be unstable. In other words, the outlier eigenvalues are the ones determining the stability of the system. Therefore, the insights about the relationship between stability and paradoxical responses obtained in the 4-unit system also carry over to the large-scale model under the HFA. Furthermore, as the response of the large circuit to partial perturbations is given by a mixture of Gaussians, and the mean of each Gaussian composing the mixture can be described by the response of a 5 dimensional circuit (four non-perturbed cell-type populations and a perturbed population, see Sect. 5.9), the link between response to perturbations and stability can be also studied in the case of practical perturbations.

2 Data Collection and Analysis

All the data presented here was collected by Daniel Mossing, refer to Mossing et al. (2021) for details about data collection. We analyzed the activity of 2664 pyramidal cells, 57 PV cells, 203 SOM cells and 169 VIP cells, whose receptive field was within 10 degrees of the stimulus center. The analyzed conditions were small drifting gratings (5 degree) for different contrast conditions (0%,6%,12%,25%,50%,100%). The cells were selected to

have a significant response that differed from the baseline (pvalue<0.05). The activity shown in Figure 1 is the mean over trials and stimulus orientations. We did not select cells that were aligned to the orientation of the stimulus. While the mouse was free to run, the analysis focused on stationary data fundamentally, i.e. trials running speed was defined smaller than 1 cm/sec as in (Mossing et al., 2021) for all figures except in Figs. 5 and S5, where we took the difference between locomotion and non-locomotion conditions. In this case, the locomotion threshold was taken to be 5 cm/sec.

3 Low-dimensional circuit models (LD model or 4-unit model)

We consider a network of $n = 4$ units, each describing the activity r^α of a particular cell type's population α, with $\alpha = \{\text{E, PV, SOM, VIP}\}$ in layer 2/3 of the visual cortex of the mouse.

3.1 Definition

The network integrates input currents z^α in the following way

$$\tau^\alpha \dot{r^\alpha} = -r^\alpha + f(z^\alpha) \qquad z^\alpha = \left(\sum_\beta^n \omega^{\alpha\beta} r^\beta + h^\alpha(c) \right) \tag{S4}$$

where τ^α is the relaxation time scale, $\omega^{\alpha\beta}$ is the connectivity matrix, and $f(z) = [.]_+^\xi$ is the activation function. We use $\xi = 2$ 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:

$$h^\alpha(c) = h_b^\alpha + h_c^\alpha c \tag{S5}$$

3.2 Fitting of mean activity

To simultaneously fit the mean rates of all four cell-types at all contrast values (six in total $c = \{0, 6, 12, 25, 50, 100\}$), we consider the steady-state equations corresponding to (S4). 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 (S4) becomes a linear equation with respect to the connectivity parameters:

$$\sqrt{r^\alpha} = \sum_\beta^n \omega^{\alpha\beta} r^\beta + h^\alpha(c) \tag{S6}$$

As stimulus-related inputs are linear functions of the stimulus contrast multiplied by different weights to E and to PV, $h^\alpha = h^E, h^{PV}, 0, 0$. Eq. (S6) 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 and h_c^α. The entries of the matrix A and the vector y are the functions of the recorded firing rates at six contrast values. For each cell type, the matrix A has 6 rows corresponding to each of the six contrast values. The number of columns of the matrix

A is equal to the number of unknown connectivity and input constants for the cell type α. 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, for each cell type α, we will have:

$$\begin{bmatrix} r^E_{c_1} & -r^P_{c_1} & -r^S_{c_1} & -r^V_{c_1} & 1 & c_1 \\ r^E_{c_2} & -r^P_{c_2} & -r^S_{c_2} & -r^V_{c_2} & 1 & c_2 \\ r^E_{c_3} & -r^P_{c_3} & -r^S_{c_3} & -r^V_{c_3} & 1 & c_3 \\ r^E_{c_4} & -r^P_{c_4} & -r^S_{c_4} & -r^V_{c_4} & 1 & c_4 \\ r^E_{c_5} & -r^P_{c_5} & -r^S_{c_5} & -r^V_{c_5} & 1 & c_5 \\ r^E_{c_6} & -r^P_{c_6} & -r^S_{c_6} & -r^V_{c_6} & 1 & c_6 \end{bmatrix} \begin{bmatrix} \omega_{\alpha E} \\ \omega_{\alpha P} \\ \omega_{\alpha S} \\ \omega_{\alpha V} \\ h^\alpha_b \\ h^\alpha_c \end{bmatrix} = \begin{bmatrix} \sqrt{r^\alpha_{c_1}} \\ \sqrt{r^\alpha_{c_2}} \\ \sqrt{r^\alpha_{c_3}} \\ \sqrt{r^\alpha_{c_4}} \\ \sqrt{r^\alpha_{c_5}} \\ \sqrt{r^\alpha_{c_6}} \end{bmatrix} \tag{S7}$$

In the work presented here we expanded the above matrix A to solve all 24 equations (6 $\times$ number of cell types). Values of parameters that approximately solve the Eq. (S6) can be found by computing the non-negative least squares (NNLS) (Chen and Plemmons, 2009) solution. The NNLS solution of Eq. (S6) constructed from mean firing rates, gives *one* set of connectivity and input parameters x. To obtain distributions of connectivity and input parameters instead, we created surrogate contrast responses sets by sampling from a multivariate Gaussian distribution with mean $r^\alpha_{c_i}$ and standard error of the mean $s^\alpha_{c_i}$. For each input configuration, we sampled 2500000 seeds to create these surrogate contrast response curves. For each sample contrast response k, NNLS gave one connectivity and input parameter set. Using each parameter set and the steady-state equations in (S6) we computed the fit $\hat{r}^\alpha(k)$ of the kth sample contrast response. Keeping the stable solutions (negative eigenvalues, all time constants were chosen to be equal to 1), the likelihood of that parameter set k

$$\mathscr{L}_k = \prod_{c_i,\alpha} \frac{1}{\sqrt{2\pi} s^\alpha_{c_i}} \exp\left\{ -\frac{(\hat{r}^\alpha_{c_i}(k) - r^\alpha_{c_i})^2}{2 s^{\alpha 2}_{c_i}} \right\} \tag{S8}$$

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. 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 Figure 1**a** are the top 200 of the 700 models that later were used as prior *seeds*.

3.3 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

$$\chi_{\alpha\beta} = \frac{dr_\alpha}{dh_\beta} = (\mathbf{f}'^{-1} - \omega)^{-1}_{\alpha\beta} \qquad \mathbf{f}'_{\alpha\beta} = \delta_{\alpha\beta} f'_\alpha \tag{S9}$$

where f'_α is the gain of population α at the considered steady state, $\mathbf{f}'$ is the 4-dimensional diagonal matrix with elements f'_α, $\delta_{\alpha\beta}$ is a Kroenecker delta which is 1 only if $\alpha = \beta$ and zero otherwise, and I is the 4-dimensional identify matrix. Defining the diagonal matrix of time constants $T_{\alpha\beta} = \delta_{\alpha\beta}\tau_\alpha$, Eq. (S9) can be written as a function of the Jacobian $J = T^{-1}(-I + \mathbf{f}'\omega)$, and in particular of its negative $-J = T^{-1}(I - \mathbf{f}'\omega)$, as

$$\chi \mathbf{f}'^{-1} T = (-J)^{-1} \qquad \rightarrow \qquad \frac{\tau_\beta}{f'_\beta}\chi_{\alpha\beta} = \frac{(-1)^{\alpha+\beta}}{\det(-J)} M_{\beta\alpha} \tag{S10}$$

where we have used the general formula for the inverse of a matrix, and $M_{\beta\alpha}$ is the corresponding minor of $-J$, that is, the determinant of $-J$ with its βth row and α^{th} column deleted. In particular, the diagonal entries of χ are

$$\frac{\tau_\alpha}{f'_\alpha}\chi_{\alpha\alpha} = \frac{1}{\det(-J)} M_{\alpha\alpha} \tag{S11}$$

We let J_α be the Jacobian of the sub-circuit without cell type α. Then $M_{\alpha\alpha} = \det(-J_\alpha)$. Thus we find that:

$$\frac{\tau_\alpha}{f'_\alpha}\chi_{\alpha\alpha} = \frac{1}{\det(-J)} \det(-J_\alpha) \tag{S12}$$

We assume the fixed point is stable, which means all of the eigenvalues of J have negative real part, so all the eigenvalues of $-J$ have positive real part and $\det(-J) > 0$. Similarly, if the subcircuit without α is stable, $\det(-J_\alpha) > 0$, while if $\det(-J_\alpha) < 0$, that subcircuit is unstable. Thus, if cell type α has a paradoxical response, meaning that $\chi_{\alpha\alpha} < 0$, then $\det(-J_\alpha) < 0$, meaning that at the fixed point, the subcircuit without α is unstable.

3.3.1 EI networks Evaluating Eq. (S12) in the EI case we obtain the result from (Tsodyks et al., 1997)

$$\chi_{II} \propto 1 - f'_E \omega_{EE} \tag{S13}$$

which makes the parameter independent prediction that when recurrent excitation strong (i.e. its 1D Jacobian is positive), the response of inhibition is paradoxical, $\chi_{II} < 0$.

3.3.2 SOM paradoxical response implies unstable E-PV sub-circuit (VIP projects only to SOM) In the particular case in which VIP projects only to SOM, the Eq. S12 reduces to

$$\frac{\tau_S}{f'_S}\chi_{SS} = \frac{1}{\det J} \det J_{EP} \tag{S14}$$

In EI networks, the network is stable if and only if the trace of the Jacobian is negative and the determinant of the Jacobian is positive. As in this networks the E time constant can always be chosen sufficiently large to satisfy the conditions on the trace, the stability will be given by the determinant alone. Therefore, measuring the paradoxical response of SOM translates in E-PV being unstable, and observing a non-paradoxical response of SOM means that E-PV is stable given a suitable time constant.

3.4 Disinhibition 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, *Disinhibition symmetries*.

$$\chi_{\mathrm{EV}} = - f'_{\mathrm{V}} \omega^{\mathrm{SV}} \chi_{\mathrm{ES}} \tag{S15}$$
$$\chi_{\mathrm{PV}} = - f'_{\mathrm{V}} \omega^{\mathrm{SV}} \chi_{\mathrm{PS}} \tag{S16}$$
$$\chi_{\mathrm{SV}} = - f'_{\mathrm{V}} \omega^{\mathrm{SV}} \chi_{\mathrm{SS}} \tag{S17}$$
$$\chi_{\mathrm{VV}} = - f'_{\mathrm{V}} \omega^{\mathrm{SV}} * \chi_{\mathrm{V}S} + f'_{\mathrm{V}} \tag{S18}$$

$$\chi_{\mathrm{VS}} = f'_{\mathrm{V}} (\omega^{\mathrm{VE}} \chi_{\mathrm{ES}} - \omega^{\mathrm{VP}} \chi_{\mathrm{PS}} - \omega^{\mathrm{VS}} \chi_{\mathrm{SS}}) \tag{S19}$$

$$\tag{S20}$$

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

$$\chi_{\alpha\beta} = \frac{1}{D} k^{\alpha\beta} \qquad \frac{1}{D} = \frac{\det\left(T^{-1}\right)}{\det(-J)} \tag{S21}$$

Where $\det(-J)$ is the determinant of the negative Jacobian of the full system, defined above Eq. (S10). 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

$$k^{\mathrm{ES}} = - f'_{\mathrm{E}} f'_{\mathrm{S}} (\omega^{\mathrm{ES}} (1 + f'_{\mathrm{P}} \omega^{\mathrm{PP}}) - f'_{\mathrm{P}} \omega^{\mathrm{EP}} \omega^{\mathrm{PS}}) \tag{S22}$$
$$k^{\mathrm{EV}} = f'_{\mathrm{E}} f'_{\mathrm{S}} f'_{\mathrm{V}} \omega^{\mathrm{SV}} (\omega^{\mathrm{ES}} (1 + f'_{\mathrm{P}} \omega^{\mathrm{PP}}) - f'_{\mathrm{P}} \omega^{\mathrm{EP}} \omega^{\mathrm{PS}}) \tag{S23}$$

$$k^{\mathrm{PS}} = - f'_{\mathrm{P}} f'_{\mathrm{S}} (\omega^{\mathrm{PS}} (1 - f'_{\mathrm{E}} \omega^{\mathrm{EE}}) + f'_{\mathrm{E}} \omega^{\mathrm{ES}} \omega^{\mathrm{PE}}) \tag{S24}$$
$$k^{\mathrm{PV}} = f'_{\mathrm{P}} f'_{\mathrm{S}} f'_{\mathrm{V}} \omega^{\mathrm{SV}} (\omega^{\mathrm{PS}} (1 - f'_{\mathrm{E}} \omega^{\mathrm{EE}}) + f'_{\mathrm{E}} \omega^{\mathrm{ES}} \omega^{\mathrm{PE}}) \tag{S25}$$

$$k^{\mathrm{SS}} = f'_{\mathrm{S}} ((1 - f'_{\mathrm{E}} \omega^{\mathrm{EE}})(1 + f'_{\mathrm{P}} \omega^{\mathrm{PP}}) + f'_{\mathrm{E}} f'_{\mathrm{P}} \omega^{\mathrm{EP}} \omega^{\mathrm{PE}}) \tag{S26}$$
$$k^{\mathrm{SV}} = - \omega^{\mathrm{SV}} f'_{\mathrm{V}} f'_{\mathrm{S}} ((1 - f'_{\mathrm{E}} \omega^{\mathrm{EE}})(1 + f'_{\mathrm{P}} \omega^{\mathrm{PP}}) + f'_{\mathrm{E}} f'_{\mathrm{P}} \omega^{\mathrm{EP}} \omega^{\mathrm{PE}}) \tag{S27}$$

$$k^{\mathrm{VS}} = - f'_{\mathrm{S}} f'_{\mathrm{V}} \left(f'_{\mathrm{E}} \left(\omega^{\mathrm{ES}} \omega^{\mathrm{VE}} - \frac{f'_{\mathrm{P}} \left| \omega^{0} \right|}{\omega^{\mathrm{SV}}} \right) + \omega^{\mathrm{VS}} (1 - f'_{\mathrm{E}} \omega^{\mathrm{EE}}) + f'_{\mathrm{P}} (\omega^{\mathrm{PP}} \omega^{\mathrm{VS}} - \omega^{\mathrm{PS}} \omega^{\mathrm{VP}}) \right) \tag{S28}$$
$$k^{\mathrm{VV}} = f'_{\mathrm{V}} (f'_{\mathrm{E}} f'_{\mathrm{P}} f'_{\mathrm{S}} \omega^{\mathrm{SE}} (\omega^{\mathrm{ES}} \omega^{\mathrm{PP}} - \omega^{\mathrm{EP}} \omega^{\mathrm{PS}}) + (1 - f'_{\mathrm{E}} \omega^{\mathrm{EE}})(1 + f'_{\mathrm{P}} \omega^{\mathrm{PP}}) + f'_{\mathrm{E}} f'_{\mathrm{P}} \omega^{\mathrm{EP}} \omega^{\mathrm{PE}} + f'_{\mathrm{E}} f'_{\mathrm{S}} \omega^{\mathrm{ES}} \omega^{\mathrm{SE}}) \tag{S29}$$

3.5 Disinhibition symmetries underlying behavioral modulations of activity.

We find that, in the absence of visual stimulation, locomotory inputs target only SOM and VIP. In that case, for the change in E cell activity to vanish, its needed that the change in E activity due to the locomotory input to SOM (ΔE_S) and the changes in E due to the locomotory inputs to VIP (ΔE_V) cancel one another. Using the notation above and introducing the size of the locomotory inputs to S and V as I_S and I_V respectively, the locomotory inputs that would lead to such cancellation should respect:

$$\Delta E = \Delta E_S + \Delta E_V = \chi_{\mathrm{ES}}\delta S + \chi_{\mathrm{EV}}\delta V = \chi_{\mathrm{ES}} I_S - f'_{\mathrm{V}}\omega^{\mathrm{SV}}\chi_{\mathrm{ES}} I_V = 0 \rightarrow \frac{I_S}{I_V} = f'_{\mathrm{V}}\omega^{\mathrm{SV}} \tag{S30}$$

3.6 Equivalent class for low-dimensional circuits .

To understand how the conclusions derived here would be modified by considering firing rates instead of deconvolved calcium imaging data we follow (Khan and Hofer, 2018), 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

$$r^{\mathrm{new}} = A^{\xi} r \tag{S31}$$

$$W^{\mathrm{new}} = AWA^{-\xi} \tag{S32}$$

$$h^{\mathrm{new}} = Ah \tag{S33}$$

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

$$J^{\mathrm{new}} = A^{\xi} J A^{-\xi} \qquad R^{\mathrm{new}} = A^{\xi} R A^{-1} \tag{S34}$$

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 *Disinhibition symmetries* equations will be re-scaled.

4 High-dimensional circuit models

4.1 Definition

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 (Tremblay et al., 2016). The steady-state activity r_i^{α} of the unit i in the population α is given by:

$$r_i^\alpha = f(z_i^\alpha) \qquad z_i^\alpha = \left(\sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta} r_j^\beta + h_i^\alpha \right) \tag{S35}$$

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:

$$\langle w_{ij}^{\alpha\beta} \rangle = w^{\alpha\beta}/N \qquad \langle (w_{ij}^{\alpha\beta})^2 \rangle - \langle w_{ij}^{\alpha\beta} \rangle^2 = \sigma^{\alpha\beta^2}/N \tag{S36}$$

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. (S35) can be re written as a function of the input to each cell :

$$z_i^\alpha = \sum_\beta^n \sum_j^{N^\beta} w_{ij}^{\alpha\beta} f\left(z_j^\beta\right) + h_i^\alpha \tag{S37}$$

4.2 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 (2015). 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:

$$m^\alpha = \langle f(z_i^\alpha) \rangle \tag{S38}$$

$$v^\alpha = \langle f(z_i^\alpha)^2 \rangle \tag{S39}$$

$$q^\alpha = N^\alpha/N \tag{S40}$$

By taking the mean and the variance of Eq. (S37) 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

$$u^\alpha = \sum_\beta w^{\alpha\beta} q^\beta m^\beta + h^\alpha \tag{S41}$$

$$\Delta^\alpha = \sum_\beta (\sigma^{\alpha\beta})^2 q^\beta v^\beta + (\lambda^\alpha)^2 \tag{S42}$$

where

$$m^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z) e^{-z^2/2} dz \tag{S43}$$

$$v^\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^\alpha + \sqrt{\Delta^\alpha} z)^2 e^{-z^2/2} dz \tag{S44}$$

We observe that if there is no disorder, Eqs. (S41) and (S43) reduce to the model from Eq. (S4) with $\omega^{\alpha\beta} = w^{\alpha\beta} q^{\beta}$ and $m^{\alpha} = r^{\alpha}$.

4.3 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

$$\frac{dr_i^{\alpha}}{dL} = f'(u^{\alpha} + \sqrt{\Delta^{\alpha}} z_i) \left(\frac{du^{\alpha}}{dL} + \frac{1}{2\sqrt{\Delta^{\alpha}}} \frac{d\Delta^{\alpha}}{dL} z_i \right) \tag{S45}$$

Taking the average and using Eq. (S43), we find that the mean of the response distribution to laser perturbation is given by the change in the mean activity of the population:

$$\begin{aligned} \langle \frac{dr_i^{\alpha}}{dL} \rangle =& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(u^{\alpha} + \sqrt{\Delta^{\alpha}} z) \left(\frac{du^{\alpha}}{dL} + \frac{1}{2\sqrt{\Delta^{\alpha}}} \frac{d\Delta^{\alpha}}{dL} z \right) e^{-z^2/2} dz \\ =& \frac{d}{dL} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u^{\alpha} + \sqrt{\Delta^{\alpha}} z) e^{-z^2/2} dz \right) \\ =& \frac{dm^{\alpha}}{dL} \end{aligned} \tag{S46}$$

This equation relates how the mean of the distribution of responses to perturbation relates to the response of the mean activity in the fully nonlinear system.

4.4 Data Fitting

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 means and variances of the connections and the inputs have to be found simultaneously for the large scale model to fit the data.

To fit the system defined by Eqs. (S41, S43), 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} (Roxin et al., 2011):

$$P^{\alpha}(r) = \frac{1}{\sqrt{2\pi} r^{1-1/\xi} \xi \sqrt{\Delta^{\alpha}}} e^{-\frac{(r^{1/\xi} - u^{\alpha})^2}{2\Delta^{\alpha}}} \Theta(r) + \frac{1}{2} \left(1 - \mathrm{erf}\left(\frac{u^{\alpha}}{\sqrt{2\Delta^{\alpha}}} \right) \right) \delta(r) \tag{S47}$$

$$P^{\alpha}(r) = P^{+}(r)\Theta(r) + P^{0}\delta(r) \tag{S48}$$

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. S47 with $\xi = 2$ and proceed as follows (Fig. 1**a**, see also Fig. S1): For each cell type α and each contrast c, we fit the analytical distribution of rates from Eq. (S47) to the distribution of experimentally recorded activity. We denote the fit distribution by $P_{\mathrm{d}}^{\alpha}(r, \mu_{\mathrm{d}}^{\alpha}(c), \Delta_{\mathrm{d}}^{\alpha}(c))$. 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 mean external input to the population α has the form $h^{\alpha} = h_b^{\alpha} + h_c^{\alpha} c$, and has a standard deviation λ^{α} which is independent of contrast. All the cell-type receive baseline inputs but only E and PV have contrast dependent inputs. The model has in the end 34 parameters (16 mean weights and 8 low-rank variances of the weights, 4 mean baseline inputs, 2 mean weights for the stimulus-related inputs to E and PV, and 4 input variances, independent of contrast).

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 Approximate Bayesian Computation, also known as ABC search (Sisson et al., 2018), 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 to be Gaussian distributions with mean 0.5 and variance 0.3. The only dependence on contrast is through the inputs mean, the variance of the inputs is independent of contrast. This was enforced to capture the dynamical spreading out of the activity with contrast (see Figure 1) and to not enforce it with an input that already is broader with contrast. Nevertheless, it is possible that the inputs received by layer 2/3 cells from layer 4 are already broader at higher 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 squared norm of the matrix of the Kullback-Leibler divergences between these two distributions:

$$E = \sum_{c} \sum_{\alpha} D(P_{\mathrm{d}}^{\alpha}(c) || P_{\mathrm{mf}}^{\alpha}(c))^2 \tag{S49}$$

where D (dropping temporarily the dependence on the contrast for ease of notation) is given by:

$$D(P_{\mathrm{d}}^{\alpha} || P_{\mathrm{mf}}^{\alpha}) = \int_{-\infty}^{\infty} P_{\mathrm{d}}^{\alpha}(r) \log \frac{P_{\mathrm{d}}^{\alpha}(r)}{P_{\mathrm{mf}}^{\alpha}(r)} dr = P_{\mathrm{d}}^{0,\alpha} \log \frac{P_{\mathrm{d}}^{0,\alpha}}{P_{\mathrm{mf}}^{0,\alpha}} + \int_{0^+}^{\infty} P_{\mathrm{d}}^{+,\alpha}(r) \log \frac{P_{\mathrm{d}}^{+,\alpha}(r)}{P_{\mathrm{mf}}^{+,\alpha}(r)} dr = I_A^{\alpha} + I_B^{\alpha} \tag{S50}$$

with

$$\mathrm{I}_A^{\alpha} = \frac{1}{2} \operatorname{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right) \log \left(\frac{\operatorname{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)}{\operatorname{erf}\left(\frac{\mu_{\mathrm{mf}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{mf}}^{\alpha}}}\right)} \right) \tag{S51}$$

$$\mathrm{I}_B^{\alpha} = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)\right) \log \frac{\sqrt{\Delta_{\mathrm{mf}}^{\alpha}}}{\sqrt{\Delta_{\mathrm{d}}^{\alpha}}} + \frac{1}{4\Delta_{\mathrm{mf}}^{\alpha}} \Bigg(\left(\Delta_{\mathrm{d}}^{\alpha} - \Delta_{\mathrm{mf}}^{\alpha} + \left(\mu_{\mathrm{d}}^{\alpha} - \mu_{\mathrm{mf}}^{\alpha}\right)^2\right)\left(1 + \operatorname{erf}\left(\frac{\mu_{\mathrm{d}}^{\alpha}}{\sqrt{2\Delta_{\mathrm{d}}^{\alpha}}}\right)\right) + \tag{S52}$$

$$\sqrt{\frac{2}{\pi \Delta_{\mathrm{d}}^{\alpha}}} e^{-\frac{\mu_{\mathrm{d}}^{\alpha 2}}{2\Delta_{\mathrm{d}}^{\alpha}}} \left(\Delta_{\mathrm{d}}^{\alpha}(\mu_{\mathrm{d}}^{\alpha} - 2\mu_{\mathrm{mf}}^{\alpha}) + \Delta_{\mathrm{mf}}^{\alpha} \mu_{\mathrm{d}}^{\alpha}\right) \Bigg) \tag{S53}$$

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 either take the first 300 or those with KL divergence smaller than 0.5. This defines a family of high-dimensional models (Fig. 1) 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.

4.5 Equivalent class for high-dimensional circuits .

Analogously as shown for the LD case above, we are interested in understanding how to map the models derived here to those we could have derived had we had access to firing rates recordings instead of deconvolved calcium imaging data. We assume that calcium activity $\frac{\Delta F}{F}$ and firing rates can be related via a linear relationship, as shown in (Khan and Hofer, 2018). Let us say we want to change the mean activity of each cell type α by a factor x^α

$$m^\alpha_{\text{new}} = x^\alpha r^\alpha \tag{S54}$$

Then from Eqs. (S43, S41) we have

$$w^{\alpha\beta}_{\text{new}} = \frac{w^{\alpha\beta} f^{-1}(x^\alpha)}{x^\beta} \tag{S55}$$

$$h^\alpha_{\text{new}} = h^\alpha f^{-1}(x^\alpha) \tag{S56}$$

$$\sigma^{\alpha\beta\,2}_{\text{new}} = \sigma^{\alpha\beta\,2}\left(\frac{f^{-1}(x^\alpha)}{x^\beta}\right)^2 \tag{S57}$$

$$\sigma^{\alpha\;\;2}_{\text{new}} = \sigma^{\alpha 2}\left(f^{-1}(x^\alpha)\right)^2 \tag{S58}$$

$$\tag{S59}$$

5 Analytical approach to linear response of disordered networks

5.1 Set up

We call the steady state solution of Eq. (S35) ${}^*r^\alpha_i$ and the steady state input ${}^*z^\alpha_i$. The time evolution of the response to a perturbation δh^α_i, can be described by the dynamics of δr^α_i:

$$\tau^\alpha_i \delta \dot{r}^\alpha_i = -\delta r^\alpha_i + f'^\alpha_i \cdot \left(\sum_\beta^n \sum_j^{N^\beta} w^{\alpha\beta}_{ij} \delta r^\beta_j + \delta h^\alpha_i\right) \qquad \text{where} \qquad f'^\alpha_i = f'^\alpha({}^*z^\alpha_i) = f'\left(\sum_\beta^n \sum_j^{N^\beta} w^{\alpha\beta}_{ij} {}^*r^\beta_j + h^\alpha_i\right) \tag{S60}$$

Switching from now onwards to matrix notation, we define

$$F_{ij}^{\alpha\beta} = \delta_{ij}\delta_{\alpha\beta} f_i'^{\alpha} \tag{S61}$$

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:

$$\delta\vec{r} = (F^{-1} - W)^{-1}\vec{\delta h} \qquad \vec{\delta r} = R\vec{\delta h} \tag{S62}$$

which defines the high-dimensional linear response matrix $R = (F^{-1} - W)^{-1}$. In order to be able to compute the block-wise variance of the response matrix analytically, we need to constrain the cell-type specific variance to be low rank. The block-wise variance of W (defined in Eq. S36) is then written as $(\sigma^{\alpha\beta})^2/N = \nu^{\alpha}\kappa^{\beta}/N$. We write W as the sum of a homogeneous component W_0, with $W_{0ij}^{\alpha\beta} = \frac{1}{N}w^{\alpha\beta}$ and a random component $\Pi_L\Psi\Pi_R$, where Ψ 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:

$$W = W_0 - \Pi_L\Psi\Pi_R \qquad \text{with} \qquad \Pi_L = \delta_{ij}\sqrt{\kappa_{\alpha_j}} \qquad \Pi_R = \delta_{ij}\sqrt{\nu_j} \tag{S63}$$

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. That approximation, which we call the Homogeneous fixed point approximation or HFA, assumes that the linearized system can be written as :

$$\tau_i^{\alpha}\delta\dot{r}_i^{\alpha} = -\delta r_i^{\alpha} + f'^{\alpha}\cdot\left(\sum_{\beta}^{n}\sum_{j}^{N^{\beta}} w_{ij}^{\alpha\beta}\delta r_j^{\beta} + \delta h_i^{\alpha}\right) \qquad f'^{\alpha}_i = f'^{\alpha} = f'\left(\sum_{\beta}^{n} w^{\alpha\beta}q^{\beta *}r^{\beta} + h^{\alpha}\right) \tag{S64}$$

In practice, we solve the non-disordered system to compute f'^{α} and look at the response matrix of that system. We observe that the HFA will approximate the response of the nonlinear circuit to small perturbations but wont approximate the activity itself.

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

Using results from (Ahmadian et al., 2015) we find that in the special case of the HFA, where $f'^{\alpha}_i = f'^{\alpha}$ as described above, the mean linear response matrix over different instantiations of the disorder is the linear response of the non- disordered case:

$$\langle (F_{HFA}^{-1} - W)^{-1}\rangle_J = \langle (F_{HFA}^{-1} - W_0 + \Pi_L J \Pi_R)^{-1}\rangle_J = (F_{HFA}^{-1} - W_0)^{-1} = R^0 \tag{S65}$$

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:

$$\text{the } \textit{measuring} \text{ matrix} \qquad B = \delta b \delta b^{\mathsf{T}} \tag{S66}$$

$$\text{the } \textit{optogenetic perturbation} \text{ matrix} \qquad \Sigma = \delta h \delta h^{\mathsf{T}} \tag{S67}$$

we can write the second moment of that measured signal s (Ahmadian et al., 2015):

$$\langle s^2 \rangle = \langle (\vec{\delta r}^{\mathsf{T}} \delta b)^2 \rangle = \langle \vec{\delta r}^{\mathsf{T}} B \vec{\delta r} \rangle = \mathscr{F} + \Delta \mathscr{F} \tag{S68}$$

where

$$\mathscr{F} = Tr(BR^0 \Sigma R^{0\mathsf{T}}) \qquad \Delta \mathscr{F} = \frac{1}{N} \frac{Tr\big(BR^0 \Pi_L \Pi_L R^{0\mathsf{T}}\big) Tr\big(\Pi_R \Pi_R R^0 \Sigma R^{0\mathsf{T}}\big)}{1 - \frac{1}{N} Tr\big(\Pi_R \Pi_R R^0 \Pi_L \Pi_L R^{0\mathsf{T}}\big)} \tag{S69}$$

Where we used the definitions in Eq. (S63). 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 $e_i = \{0, ..., 1, ..., 0\}$

$$\langle \delta r_i \rangle_J = e_i R^0 \vec{\delta h} \qquad \text{where} \qquad e_i = \{0, ..., 1, ..., 0\} \tag{S70}$$

Eq. (S70) 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. (S65)) and the variance Λ_i given by

$$\Lambda_i^2 = \langle \delta r_i^2 \rangle_J - \langle \delta r_i \rangle_J^2 = \langle \delta r_i^2 \rangle_J - \mathscr{F} = \Delta \mathscr{F} \tag{S71}$$

Equations (S65), (S68) and (S69) are general formulas for how to compute the mean and the variance of the linear response distributions as a function of the optogenetic perturbation Σ and the observation matrix B. In the following sections we will explicitly compute the mean response matrix R^0 for both full and low rank connectivity and the variance in different optogenetic perturbation configurations.

5.4 Computation of the response matrix R^0 without disorder

For computing R^0 (given by Eq. (S65)) we write the block-structured matrix W_0 as a function of the LD system connectivity $\omega^{\alpha\beta} = w^{\alpha\beta} q^{\beta}$. We define the $N \times n$ matrices V and U, with columns given by N dimensional vectors

$$v^{\alpha} = \delta_{i \in \alpha} \qquad v^{(\alpha)} = (\underbrace{0, \ldots, 0}_{\sum_{l=1}^{\alpha-1} N_l}, \underbrace{1, \ldots, 1}_{N_\alpha}, \underbrace{0, \ldots, 0}_{N - \sum_{l=1}^{\alpha} N_l}) \tag{S72}$$

$$u^{\alpha} = \frac{1}{N^{\alpha}} \delta_{i \in \alpha} \qquad u^{(\alpha)} \frac{1}{N^{\alpha}} = v^{(\alpha)} \tag{S73}$$

We can then write:

$$W_0 = V\omega U^{\intercal} \tag{S74}$$

Where $w^{\alpha\beta}$ and q^{α} were introduced in Eqs (S36) and (S40), respectively. To obtain R^0 defined in Eq. (S65) 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.

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

$$R^0 = \left(F^{-1} - W_0\right)^{-1} = \left(F^{-1} - V\omega U^{\intercal}\right)^{-1} = F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^{\intercal} F \tag{S75}$$

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

$$R^0_{ij} = \delta_{ij} f'_{\alpha_i} - \frac{f'_{\alpha_i} f'_{\alpha_j}}{N_{\alpha_j}} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i \alpha_j} \tag{S76}$$

$\mathbf{f}'$ was defined in Eq. (S9). 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

$$S_{\alpha_i \alpha_j} = \delta_{\alpha_i \alpha_j} f'_{\alpha_j} - f'_{\alpha_i} f'_{\alpha_j} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i \alpha_j} \tag{S77}$$

Substituting Eq. (S77) into (S76), we obtain an expression for the linear response which will be useful in later sections:

$$R^0_{ij} = \delta_{ij} f'_{\alpha_i} + \frac{S_{\alpha_i \alpha_j}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i \alpha_j} f'_{\alpha_j}}{N_{\alpha_j}} \tag{S78}$$

We point out that the Eq. (S77) is independent of N, and is finite in the limit of large N.

5.4.1 Case of rank-one ω In the case of rank-one ω , we cannot invert ω in Eq. (S75) . Instead we write:

$$W_0 = \frac{vu^{\intercal}}{N} \qquad v = \{1, ..., 1, ..., 1\} \tag{S79}$$

$$u = \{\underbrace{w_1, \cdots, w_1}_{N_1}, \underbrace{w_2, \cdots, w_2}_{N_2}, \cdots, \underbrace{w_n, \cdots, w_n}_{N_n}\} \tag{S80}$$

Using Sherman–Morrison formula we find that

$$R^0 = (F^{-1} - W_0)^{-1} = F + \frac{1}{N} \frac{F v u^T F}{1 - \frac{u^T F v}{N}} \tag{S81}$$

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

$$R^0_{ij} = \delta_{ij} f'_{\alpha_j} + \frac{1}{N} \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} = \delta_{ij} f'_{\alpha_j} + \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S82}$$

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

$$S_{\alpha_i \alpha_j} = \delta_{\alpha_i \alpha_j} f'_{\alpha_j} + N_{\alpha_j} \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S83}$$

We note that the above expression is also finite in the large N limit. Using (S82) and (S83) we conclude that the linear response R^0_{ij} satisfies Eq. (S78) also in this case.

5.5 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).

If ω full rank (R^0_{ij} is given by (S76)), 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

$$\mu_i = \sum_j R_{ij} \delta h_j = f'_{\alpha_i} \delta_{i \in \mathscr{P}_{\alpha_i}} - f'_{\alpha_i} \sum_{\alpha_j} \gamma_{\alpha_j} f'_{\alpha_j} \left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i \alpha_j} \tag{S84}$$

The expression in Eq. (S84) 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

$$\mu^{IN}_{i\in\mathscr{P}_{\alpha_i}} = f'_{\alpha_i} - f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} f'_{\alpha_j}\left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S85}$$

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

$$\mu^{OUT}_{i\notin\mathscr{P}_{\alpha_i}} = -f'_{\alpha_i}\sum_{\alpha_j}\gamma_{\alpha_j} f'_{\alpha_j}\left[\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\right]_{\alpha_i\alpha_j} \tag{S86}$$

We note that these expressions critically depend on the sign of $\left(\mathbf{f}' - \omega^{-1}\right)^{-1}$. To capture Eq. (S85) and Eq. (S86) as a single equation we define a matrix

$$\begin{aligned}\chi^{\boldsymbol{\gamma}} =& \mathbf{f}'\delta_p - \mathbf{f}'\left(\mathbf{f}' - \omega^{-1}\right)^{-1}\mathbf{f}'\boldsymbol{\gamma} && \text{(S87)}\\ =& \mathbf{f}'\delta_p + \mathbf{f}'\omega\left(\mathbf{f}'^{-1} - \omega\right)^{-1}\boldsymbol{\gamma} && \text{(S88)}\\ =& \mathbf{f}'\delta_p + \mathbf{f}'\omega\chi\boldsymbol{\gamma} && \text{(S89)}\end{aligned}$$

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 (S84) as

$$\mu_i = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} + f'_{\alpha_i}\gamma_{\alpha_i}[\omega\chi]_{\alpha_i\alpha_i} = f'_{\alpha_i}\delta_{i\in\mathscr{P}_{\alpha_i}} + \gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i} - f'_{\alpha_i}) \tag{S90}$$

If the response is paradoxical in the LD 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

$$0 < \frac{f'_{\alpha_i}}{f'_{\alpha_i} - \chi_{\alpha_i\alpha_i}} < \gamma^C_{\alpha_i} < 1 \tag{S91}$$

In Fig. 5, 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.

5.5.1 Case of rank-one ω

$$R_{ij} = \delta_{ij} f'_{\alpha_j} - \frac{1}{N} \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} = \delta_{ij} f'_{\alpha_j} + \frac{f'_{\alpha_i} f'_{\alpha_j} w_{\alpha_j}}{ND} \tag{S92}$$

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

$$\mu_i^{IN} = f'_{\alpha_i} + f'_{\alpha_i} \sum_{\alpha_j} \gamma_{\alpha_j} q_{\alpha_j} \frac{f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} \tag{S93}$$

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

$$\mu_i^{OUT} = f'_{\alpha_i} \sum_{\alpha_j} \gamma_{\alpha_j} q_{\alpha_j} \frac{f'_{\alpha_j} w_{\alpha_j}}{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}} \tag{S94}$$

5.5.1.1 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. (S94) is always negative), but the response of those neurons that were stimulated will only be paradoxical when $\mu_i^{IN} < 0$

$$\frac{1 - \sum_{\alpha_k} q_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{q_{\alpha_j} f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S95}$$

First let us 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 (S95) is now

$$\frac{n - \sum_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S96}$$

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.

5.5.1.2 Comparison with Sadeh et al. (2017) If now we would normalize the weights in Eq. (S79) as in (Sadeh et al., 2017), meaning that $w_j/N \to w_j/(N/n)$ (See Eq. 2 of their paper), the above equation would be

$$\frac{1 - \sum_{\alpha_k} f'_{\alpha_k} w_{\alpha_k}}{f'_{\alpha_j} |w_{\alpha_j}|} < \gamma^C_{\alpha_j} \tag{S97}$$

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

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

From Eq. (S69) 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 (S78). We will first find the general expression and then evaluate for particular cases: For that we write (S69) as

$$\Lambda^2 = \frac{1}{N}\frac{Tr\big(BR^0\Pi_L\Pi_L R^{0\intercal}\big)Tr\big(\Pi_R\Pi_R R^0\Sigma R^{0\intercal}\big)}{1-\frac{1}{N}Tr\big(\Pi_R\Pi_R R^0\Pi_L\Pi_L R^{0\intercal}\big)} = \frac{M.O}{1-D} \tag{S98}$$

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:

$$\Sigma_{jk} = \sum_\eta\sum_{\eta'}\delta_{\alpha_j\eta}\delta_{\alpha_k\eta'}\delta_{k\in\mathscr{P}_{\alpha_k}}\delta_{j\in\mathscr{P}_{\alpha_j}} \tag{S99}$$

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

To compute the expression of the variance when we need to make three different terms defined above in Eq. (S98): The measuring term M, the optogenetic perturbation term O and the Denominator term D

5.6.1 The measuring term M

Using the definitions in Eq.(S63), we write

$$R^0\Pi_L = \delta_{ij}f'_{\alpha_j}\sqrt{\kappa_{\alpha_j}} + \frac{S_{\alpha_i\alpha_j}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j}f'_{\alpha_j}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}} \tag{S100}$$

and

$$\Pi_R R^0 = \delta_{ij}f'_{\alpha_j}\sqrt{\nu_{\alpha_i}} + \frac{S_{\alpha_i\alpha_j}\sqrt{\nu_{\alpha_i}}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j}f'_{\alpha_j}\sqrt{\nu_{\alpha_i}}}{N_{\alpha_j}} \tag{S101}$$

such that

$$\Pi_R R^0\Pi_L = \delta_{ij}f'_{\alpha_j}\sqrt{\nu_{\alpha_i}}\sqrt{\kappa_{\alpha_j}} + \frac{S_{\alpha_i\alpha_j}\sqrt{\nu_{\alpha_i}}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j}f'_{\alpha_j}\sqrt{\nu_{\alpha_i}}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}} \tag{S102}$$

Therefore, the term we are interested in obtaining will be

$$
\begin{aligned}
M &= Tr\big(BR^0\Pi_L\Pi_L R^{0\intercal}\big) = \sum_k\sum_i \delta b_k \delta b_i [R^0\Pi_L(R^0\Pi_L)^T]_{ik} \\
&= \sum_k\sum_i \delta b_k \delta b_i \sum_j \left(\delta_{ij} f'_{\alpha_i}\sqrt{\kappa_{\alpha_j}} + \frac{S_{\alpha_i\alpha_j}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j} f'_{\alpha_j}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}}\right)\left(\delta_{kj} f'_{\alpha_j}\sqrt{\kappa_{\alpha_j}} + \frac{S_{\alpha_k\alpha_j}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}} - \frac{\delta_{\alpha_k\alpha_j} f'_{\alpha_j}\sqrt{\kappa_{\alpha_j}}}{N_{\alpha_j}}\right) \\
&= \sum_k\sum_i \delta b_k \delta b_i \big(\delta_{ik} f'^2_{\alpha_i}\kappa_{\alpha_i} + \sum_{\alpha_j}\frac{\kappa_{\alpha_j}}{N_{\alpha_j}} S_{\alpha_i\alpha_j}S_{\alpha_k\alpha_j} - \delta_{\alpha_i\alpha_k}\frac{f'^2_{\alpha_i}\kappa_{\alpha_i}}{N_{\alpha_i}}\big)
\end{aligned} \tag{S103}
$$

Which cannot be further reduced without specifying the specific shape of δb. In the articular case in which we are observing a single neuron, then $\delta b = e_l = \{0,\cdots,0,1,0,\cdots,0\}$ at some l. In that case $\delta b_i = \delta_{il}$ and

$$
M = f'^2_{\alpha_l}\kappa_{\alpha_l} + \sum_{\alpha_j}\frac{\kappa_{\alpha_j}}{N_{\alpha_j}} S^2_{\alpha_l\alpha_j} - \frac{f'^2_{\alpha_l}\kappa_{\alpha_l}}{N_{\alpha_l}} \tag{S104}
$$

5.6.2 The optogenetic perturbation term

$$
\begin{aligned}
O &= \sum_i [\Pi_R\Pi_R R^0\Sigma R^{0\intercal}]_{ii} = \sum_i [\Pi_R R\Sigma(\Pi_R R)^T]_{ii} \\
&= \sum_{il}\delta_{il}\sum_{jk}\left(\delta_{ij} f'_{\alpha_i}\sqrt{\nu_{\alpha_i}} + \frac{S_{\alpha_i\alpha_j}\sqrt{\nu_{\alpha_i}}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j} f'_{\alpha_j}\sqrt{\nu_{\alpha_i}}}{N_{\alpha_j}}\right)\sum_\eta\sum_{\eta'}\delta_{\alpha_j\eta}\delta_{\alpha_k\eta'}\delta_{k\in\mathscr{P}_{\alpha_k}}\delta_{j\in\mathscr{P}_{\alpha_j}} \\
&\qquad\qquad \left(\delta_{lk} f'_{\alpha_l}\sqrt{\nu_{\alpha_l}} + \frac{S_{\alpha_l\alpha_k}\sqrt{\nu_{\alpha_l}}}{N_{\alpha_k}} - \frac{\delta_{\alpha_l\alpha_k}\beta_{\alpha_k}\sqrt{\nu_{\alpha_l}}}{N_{\alpha_k}}\right) \\
&= \left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta(1-\gamma_\eta)N_\eta + \sum_{\eta'}\nu_{\eta'}\left(\sum_\eta S_{\eta',\eta}\gamma_\eta\right)^2 N_{\eta'}\right)
\end{aligned} \tag{S105}
$$

5.6.3 The denominator

$$
\begin{aligned}
D &= \frac{1}{N}\sum_{ik}\delta_{ik}[\Pi_R R^0\Pi_L\Pi_L^\intercal R^{0\intercal}\Pi_R^\intercal]_{ik} \\
&= \frac{1}{N}\sum_{ik}\delta_{ik}[\Pi_R R\Pi_L(\Pi_R R\Pi_L)^T]_{ik} \\
&= \frac{1}{N}\sum_{ik}\delta_{ik}\sum_j\sqrt{\nu_{\alpha_i}}\sqrt{\kappa_{\alpha_j}}\left(\delta_{ij} f'_{\alpha_i} + \frac{S_{\alpha_i\alpha_j}}{N_{\alpha_j}} - \frac{\delta_{\alpha_i\alpha_j} f'_{\alpha_j}}{N_{\alpha_j}}\right)\sqrt{\nu_{\alpha_k}}\sqrt{\kappa_{\alpha_j}}\left(\delta_{kj} f'_{\alpha_j} + \frac{S_{\alpha_k\alpha_j}}{N_{\alpha_j}} - \frac{\delta_{\alpha_k\alpha_j} f'_{\alpha_j}}{N_{\alpha_j}}\right) \qquad \text{(S106)} \\
&= \frac{1}{N}\sum_{ik}\delta_{ik}\sqrt{\nu_{\alpha_i}\nu_{\alpha_k}}\big(\delta_{ik} f'^2_{\alpha_i}\kappa_{\alpha_i} + \sum_{\alpha_j}\frac{\kappa_{\alpha_j}}{N_{\alpha_j}} S_{\alpha_i\alpha_j}S_{\alpha_k\alpha_j} - \delta_{\alpha_i\alpha_k}\frac{f'^2_{\alpha_i}\kappa_{\alpha_i}}{N_{\alpha_i}}\big) \\
&= \frac{1}{N}\sum_{\alpha_i}\nu_{\alpha_i}\big(f'^2_{\alpha_i}\kappa_{\alpha_i}N_{\alpha_i} + N_{\alpha_i}\sum_{\alpha_j}\frac{\kappa_{\alpha_j}}{N_{\alpha_j}} S^2_{\alpha_i\alpha_j} - f'^2_{\alpha_i}\kappa_{\alpha_i}\big) \qquad \text{(S107)}
\end{aligned}
$$

5.6.4 The final expression 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 ($q_\eta = N_\eta/N$):

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

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

$$\Lambda^2_{\alpha_l} = \frac{f'^2_{\alpha_l}\kappa_{\alpha_l}\left(\sum_\eta \nu_\eta f'^2_\eta \gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'} \nu_{\eta'} \left(\sum_\eta S_{\eta',\eta}\gamma_\eta \right)^2 q_{\eta'} \right)}{1 - \sum_\eta \nu_\eta q_\eta f'^2_\eta \kappa_\eta} \tag{S109}$$

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. (S109) 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*, which is different from the *fractional paradoxical* effect, which is an effect due to the mean changes.

5.6.4.1 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 post-synaptic cell type, i.e. when $\kappa_\alpha = \nu_\alpha = \sigma$ we obtain a simpler expression for the variance of the populations:

$$\Lambda^2_{\alpha_l} = \sigma^2 \frac{\left(f'^2_{\alpha_l} + \sum_\eta \frac{S^2_{\alpha\eta}}{N_\eta} - \frac{f'^2_{\alpha_l}}{N_{\alpha_l}} \right)\left(\sum_\eta f'^2_\eta \gamma_\eta(1-\gamma_\eta)q_\eta + \sum_{\eta'} \left(\sum_\eta S_{\eta',\eta}\gamma_\eta \right)^2 q_{\eta'} \right)}{1 - \sigma^2 \sum_\eta \left(q_\eta f'^2_\eta + q_\eta \sum_{\eta'} S^2_{\eta,\eta'} \frac{1}{N_{\eta'}} - \frac{1}{N} f'^2_\eta \right)} \tag{S110}$$

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

In the case of partial stimulation, we will have a total distribution that is a mixture of Gaussians with means

$$\rho_{\alpha_i}^{\mathrm{IN}} = \frac{1}{\sqrt{2\pi}\Lambda_{\alpha_i}} \exp\left\{ -\frac{(x-\mu_{\alpha_i}^{\mathrm{IN}})^2}{2\Lambda_{\alpha_i}^2} \right\} \tag{S111}$$

$$\rho_{\alpha_i}^{\mathrm{OUT}} = \frac{1}{\sqrt{2\pi}\Lambda_{\alpha_i}} \exp\left\{ -\frac{(x-\mu_{\alpha_i}^{\mathrm{OUT}})^2}{2\Lambda_{\alpha_i}^2} \right\} \tag{S112}$$

So the total distribution of responses is

$$\rho_{\alpha_i} = \gamma_{\alpha_i}\rho_{\alpha_i}^{\mathrm{IN}} + (1-\gamma_{\alpha_i})\rho_{\alpha_i}^{\mathrm{OUT}} \tag{S113}$$

Where $\mu^{\mathrm{IN}} = R^{IN}_{j\in\mathscr{P}_{\alpha_j}}$ and $\mu^{\mathrm{OUT}} = R^{OUT}_{j\in\mathscr{P}_{\alpha_j}}$ given by Eqs (S93, S94) for low rank ω or by (S85,S86) for invertible ω, and a variance given by Eq. (S108)

5.8 Simple description of the fractional paradoxical effect

The fractional paradoxical effect can be intuitively understood in the system without disorder (the EI case with low-rank ω was studied by (Sadeh et al., 2017)). In this case, the distribution of responses will be bimodal, with two delta functions at the values given by Eq. (S90). The density then will be given by the limit of vanishing variance of Eq. (S113)

$$\rho_{\alpha_i}(x) = (1-\gamma_{\alpha_i})\delta(x-\gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i}-f'_{\alpha_i})) + \gamma_{\alpha_i}\delta(x-f'_{\alpha_i}-\gamma_{\alpha_i}(\chi_{\alpha_i\alpha_i}-f'_{\alpha_i})) \tag{S114}$$

If the unit α_i is paradoxical in the LD 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. (S91), for values of γ_{α_i} smaller than $\gamma^C_{\alpha_i}$, the mean of the perturbed population will remain positive. In this range, increasing the fraction of perturbed cells, will result in a decrease of the mass of negative responses ${}^{0}_{-\infty}\rho_{\alpha_i}(x)dx$ like $(1-\delta)$. In the non-disordered case, as soon as $\gamma_{\alpha_i} > \gamma^C_{\alpha_i}$, the mass of negative responses is unity. Given that when working with the homogeneous approximation, the response of the non -disordered system is the mean of the response of the disordered system, the intuitions here apply to the mean of the disordered case.

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

Here we show that the mean response of the perturbed fraction of the 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 one perturbed and one non-perturbed populations. We know that mapping a high-dimensional non-disordered network to a LD 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

$$\chi^5 = [f_5'^{-1} - \omega_5]^{-1} = -\begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}\gamma_P & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}\gamma_P - 1/f_P' & \omega^{\mathrm{PP}}(1-\gamma_P) & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}\gamma_P & \omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P' & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}\gamma_P & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}\gamma_P & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix}^{-1} \tag{S115}$$

The term we are interested in is the PP term, given by:

$$\chi^5_{PP} = [f_5'^{-1} - \omega_5]^{-1}_{pp} = \frac{1}{\det\left[\mathbf{f'}_5{}^{-1} - \omega_5\right]} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{PE}} & \omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P' & \omega^{\mathrm{PS}} & \omega^{\mathrm{PV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S116}$$

$$\chi^5_{PP} = \frac{1}{\det\left[\mathbf{f'}_5{}^{-1} - \omega_5\right]} \Big(-\omega^{\mathrm{PE}} \det \begin{bmatrix} \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S117}$$

$$+ (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P') \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{ES}} & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SS}} - 1/f_s' & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VS}} & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S118}$$

$$- \omega^{\mathrm{PS}} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{EV}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SV}} \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VV}} - 1/f_V' \end{bmatrix} \tag{S119}$$

$$+ \omega^{\mathrm{PV}} \det \begin{bmatrix} \omega^{\mathrm{EE}} - 1/f_E' & \omega^{\mathrm{EP}}(1-\gamma_P) & \omega^{\mathrm{ES}} \\ \omega^{\mathrm{SE}} & \omega^{\mathrm{SP}}(1-\gamma_P) & \omega^{\mathrm{SS}} - 1/f_s' \\ \omega^{\mathrm{VE}} & \omega^{\mathrm{VP}}(1-\gamma_P) & \omega^{\mathrm{VS}} \end{bmatrix} \qquad dotted \tag{S120}$$

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

$$\chi^5_{PP} = \frac{1}{\det\left[\mathbf{f'}_5{}^{-1} - \omega_5\right]} \left(\omega^{\mathrm{PE}} M_{PE}(1-\gamma_P) - (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P')M_{PP} + \omega^{\mathrm{PS}} M_{PS}(1-\gamma_P) - \omega^{\mathrm{PV}} M_{PV}(1-\gamma_P)\right) \tag{S121}$$

Where $M_{\alpha\beta}$ are the minors of the original 4D matrix $(\mathbf{f'}^{-1} - \omega)$. Using that $\chi_{\alpha\beta} = \frac{1}{\det\left(\mathbf{f'}^{-1} - \omega\right)}(-1)^{\alpha\beta} M_{\beta\alpha}$, and that $\det\left(\mathbf{f'}^{-1} - \omega\right) = f_p' \det\left[\mathbf{f'}_5{}^{-1} - \omega_5\right]$

$$\begin{aligned} \chi^5_{PP} = & - f_p' \left(\omega^{\mathrm{PE}} \chi_{EP}(1-\gamma_P) + (\omega^{\mathrm{PP}}(1-\gamma_P) - 1/f_P')\chi_{PP} + \omega^{\mathrm{PS}} \chi_{SP}(1-\gamma_P) + \omega^{\mathrm{PV}} \chi_{VP}(1-\gamma_P)\right) \\ & - f_p' \left([\omega\chi]_{PP}(1-\gamma_P) - 1/f_P' \chi_{PP}\right) \end{aligned} \tag{S122}$$

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

$$\chi^5_{PP} = f'_p\big(\gamma_P[\omega\chi]_{PP} + 1\big) \tag{S123}$$

Given that the mean response of the perturbed population in a high-dimensional system, given by Eq. (S84) (and also Eq. S87) is $\chi^{\boldsymbol{\gamma}}_{PP} = f'_P(\gamma_P[\omega\chi]_{PP} + 1\big)$, we obtain that

$$\chi^5_{PP} = \chi^{\boldsymbol{\gamma}}_{PP} \tag{S124}$$

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

5.10 Response distribution to partial and non-homogeneous perturbations.

We now consider the case in which each population receives a partial, but the intensity of the perturbation is different for each perturbed neuron, mimicking disorder in the ChR2 expression. More specifically we need to recompute the expressions in equations (S68 , S69) 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. (S99), will be in this case:

$$\Sigma_{jk} = \sum_\eta \sum_{\eta'} \delta_{\alpha_j \eta} \delta_{\alpha_k \eta'} \delta_{k \in \mathscr{P}_{\alpha_k}} \delta_{j \in \mathscr{P}_{\alpha_j}} (D^{\alpha_i}_i)(D^{\alpha_j}_j) \tag{S125}$$

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 (S68 , S69). Specifically:

$$\delta r_i = \sum_j R^0_{ij} \delta h_j + \Lambda_{\alpha_i} \xi_i \tag{S126}$$

$$= M_{\alpha_i} + \Lambda_{\alpha_i} \xi_i \tag{S127}$$

with

$$M_{\alpha_i} = f'_{\alpha_i} D_i - f'_{\alpha_i} \sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1} \right)^{-1} \right]_{\alpha_i \eta} d_\eta \tag{S128}$$

where $\xi_i \sim \mathcal{N}(0,1)$ and Λ_{α_i} is the generalization of Eq. (S108) to disordered perturbations, obtained by replacing Eq. (S125) into (S98) (we note that the only term that needs to be re-computed is the term M).

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

In the large N limit, this equation reduces to

$$\Lambda^2_{\alpha_i} = \frac{{f'}^2_{\alpha_i} \kappa_{\alpha_i} \left(\sum_\eta \nu_\eta {f'_\eta}^2 \gamma_\eta \left((1-\gamma_\eta) d^2_\eta + g^2_\eta \right) q_\eta + \sum_{\eta'} \nu_{\eta'} \left(\sum_\eta S_{\eta',\eta} d_\eta \gamma_\eta \right)^2 q_{\eta'} \right)}{1 - \sum_\eta \nu_\eta q_\eta {f'}^2_\eta \kappa_\eta} \tag{S130}$$

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

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

$$\mu^{OUT}_{\alpha_i} = -f'_{\alpha_i} \sum_\eta \gamma_\eta f'_\eta \left[\left(\mathbf{f}' - \omega^{-1} \right)^{-1} \right]_{\alpha_i \eta} d_\eta \tag{S132}$$

$$\Lambda^{2,IN}_{\alpha_i} = {f'_{\alpha_i}}^2 g^2_{\alpha_i} + \Lambda^2_{\alpha_i} \tag{S133}$$

$$\Lambda^{2,OUT}_{\alpha_i} = \Lambda^2_{\alpha_i} \tag{S134}$$

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

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

$$\rho^{\mathrm{OUT}}_{\alpha_i} = \frac{1}{\sqrt{2\pi} \Lambda^{\mathrm{OUT}}_{\alpha_i}} \exp\left\{ -\frac{(x - \mu^{\mathrm{OUT}}_{\alpha_i})^2}{2 \Lambda^{\mathrm{OUT}2}_{\alpha_i}} \right\} \tag{S136}$$

So the total distribution of responses is

$$\rho_{\alpha_i} = \gamma_{\alpha_i} \rho^{\mathrm{IN}}_{\alpha_i} + (1 - \gamma_{\alpha_i}) \rho^{\mathrm{OUT}}_{\alpha_i} \tag{S137}$$

5.11 Link to the low-dimensional system linear response

The activity of the LD 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 LD system. If 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

$$\chi_{\alpha_i,\alpha_j} = \frac{1}{N_{\alpha_i}} \sum_{i\in\alpha_i} \big(\sum_{j\in\alpha_j} R^0{}_{ij} \big) = U^{\intercal}_{\alpha_i} R^0 V_{\alpha_j} \tag{S138}$$

Inserting (S75) into the above expression we obtain:

$$\chi = U^{\intercal} R V = U^{\intercal}(F - FV\left(\mathbf{f}' - \omega^{-1}\right)^{-1} U^{\intercal} F)V \tag{S139}$$

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

$$= \left(\mathbf{f}'^{-1} - \omega\right)^{-1} \tag{S141}$$

Which is the definition of χ as in Eq. (S9). We also need to show that the variance vanishes for large N. Writing $B_{ij} = \frac{1}{N^2_{\eta'}}\delta_{\alpha_i\alpha_j}\delta_{\alpha_i\eta'}$ ($B = \delta b \delta b^T$), and inserting it and Eq. (S92) into Eq. (S69) we find that :

$$\Lambda_{\eta\eta'} = \frac{\sigma^2}{N} \frac{\left(\sum_{\alpha_i} S^2_{\alpha_i\eta'} N_{\alpha_i}\right)\left(\sum_{\alpha_i} \frac{S^2_{\alpha_i\eta'}}{N_{\alpha_i}}\right)}{1 - \frac{\sigma^2}{N}\left(\sum_{\alpha_i} N_{\alpha_i} f'^{\,2}_{\alpha_i} + \sum_{\alpha_i,\alpha_j} S^2_{\alpha_i,\alpha_j} \frac{N_{\alpha_i}}{N_{\alpha_j}} - \sum_{\alpha_i} f'^{\,2}_{\alpha_i}\right)} \tag{S142}$$

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.

6 Eigenvalue Spectrum of the Jacobian in the HFA

In (Tao, 2013), it is shown that given a matrix with i.i.d. entries to which is added a deterministic matrix with rank $\mathscr{O}(1)$ and operator norm $\mathscr{O}(1)$, 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 outlier eigenvalues, are placed asymptotically in the same location as the eigenvalues of the low rank added matrix.

We now show that the Jacobian of our system under the HFA satisfies that condition. The scaling chosen for the low rank part of the connectivity matrix in the HFA guarantees that its operator norm is $\mathscr{O}(1)$. Using the definition of F and of $\mathbf{f}'$ from Eq. (S61) , the Jacobian under the HFA is written as

$$\mathscr{J} = -I + FW = I + F(W^0 + \Pi_L J \Pi_R) = -I + F\Pi_L J \Pi_R + FW^0 \tag{S143}$$

Let us for a moment write the non-disordered component of the connectivity as

$$W^0_{ij} = \frac{w_{\alpha_i\alpha_j}}{s(N)}\delta_{i\in\alpha_i}\delta_{i\in\alpha_j} \tag{S144}$$

where $w_{\alpha\beta}$ is as before and s(N) is the scaling, which up to now we took to be $s(N) = N$. If we take U and V as defined in Eq.S72 and define the n-dimensional diagonal matrix L with elements $L_{\alpha\alpha} = q^\alpha = N^\alpha/N$ and I_n as the n-dimensional identity, then we have the following relations :

$$V^TV = LNI_n \qquad U^TU = \frac{L^{-1}}{N}I_n \qquad V^T = NLU^T \quad \text{and} \quad V^TU = I_n \tag{S145}$$

We want to compute the operator norm of the low-rank part $\mathscr{J}_0 = FW^0$, given by $||\mathscr{J}_0||^2 = \lambda_1(\mathscr{J}_0\mathscr{J}_0^T)$ where λ_1 is the max eigenvalue. We notice that $\mathscr{J}_0 = FW^0 = V\mathbf{f}'\frac{w}{s(N)}V^T \to \frac{N}{s(N)}V\mathbf{f}'wL_nU^T$. We notice that $\omega = wL_n$ is the connectivity of the non-disordered LD system, and that therefore if $s(N) = N$ then the nonzero eigenvalues of J_0 are those of the LD system. In order to compute the eigenvalues of $J_0J_0^T$ we notice:

$$\begin{aligned} J_0J_0^T =& (V\mathbf{f}'\frac{w}{s(N)}L_nNU^T)(UL_nN\frac{w^T}{s(N)}\mathbf{f}'V^T) && \text{(S146)}\\ =& \frac{N}{s(N)^2}V(\mathbf{f}'wL_nw^T\mathbf{f}')V^T && \text{(S147)}\\ =& \frac{N^2}{s(N)^2}V(\mathbf{f}'wL_nw^T\mathbf{f}'L_n)U^T && \text{(S148)} \end{aligned}$$

Where we used Eq S145 twice. This matrix then, has (N-n) zero eigenvalues and n eigenvalues that are those of $\frac{N^2}{s(N)^2}(\mathbf{f}'wL_n)(w^T\mathbf{f}'L_n)$. We notice that only if $s(N) = N$ these eigenvalues will be $\mathscr{O}(1)$. In the simplest case in which $F\Pi_LJ\Pi_R$ is i.i.d., taking $s(N) = N$ would be the only requirement to prove that the HFA has outlier eigenvalues if the LD system. For the case studied here, in which the disordered part of the connectivity is not *i.i.d.* but has a block-structured variance, there is not current proof that the deterministic outliers are those of the low-rank perturbation, but we nevertheless observe good agreement numerically.

7 Numerics

All numerical simulations and analyses were done with custom code in Python.

7.1 Network simulations

All network simulations were done with an Euler scheme with d=0.05 ms. The networks had N=3000 neurons with the number of E, PV, SOM and VIP cells being the 0.8 %,0.25 %,0.125 %, and 0.125 % of the total, respectively. We observe that due to the scaling in the model, none of these simulations depend on network size. Simulations usually ran for half a second (500 ms) each. All the simulations were done on models obtained as described in Sec. 4.4, with the additional requirement that they are stable, i.e. that the Jacobian of the network had all negative eigenvalues.

7.2 Linear response

The numerical calculation of the linear response was done by simulating the network, defining the matrix F as in Eq. S61 and numerically evaluating Eq. S62. Usually, 25 network instantiations were used to compute the distributions shown in the main text, given the relative small number of cells in the interneuronal populations.

7.3 Inference of locomotion inputs

In order to infer the locomotion inputs, fitted the distribution of changes in activity induced by locomotion per cell type and stimulus condition with a Gaussian distribution. This gave us 4 mean values and 4 variance values per contrast condition. This would correspond to the left hand side of Eq. S128, and S130 respectively. Now the only unknowns in those equations are $d_\eta(c)$ and $g_\eta(c)$ the (contrast dependent) locomotory inputs per cell-type η. These functions were chosen to be $d_\eta(c) = a\eta + b\eta * C(c)$ where $C(c) = \sqrt{c}$ in the main text (Fig. 5) and $C(c) = c$ or $C(c) = 1$ (linear with, or independent of, contrast) in the supplement (Fig. S5). By numerically solving for all the 8 constraints to occur for each contrast, we find the locomotion dependent inputs.

Supplementary Figures

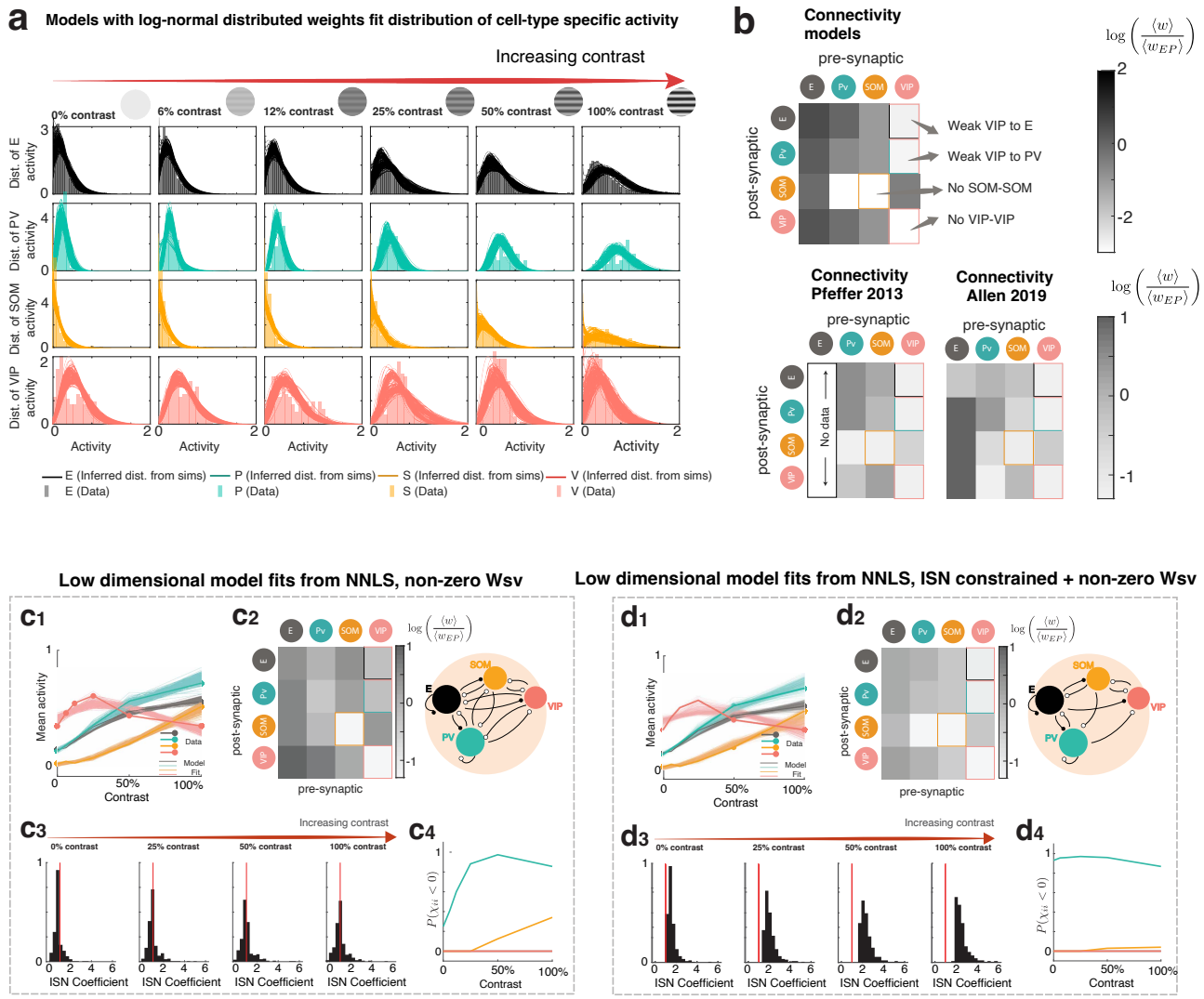

Figure S1 **a) Large-scale models with log-normally distributed weights.** The mean field theory used here is valid for any weight distribution whose tails decay as fast or faster than a Gaussian (Jonathan Touboul, personal communication). In particular any truncated distribution will obey this constraint. Here, we instead show that if we take the parameters found in the family of models and use non-constrained log-normal weight distributions, the models are still in good agreement with the data. Shown here are the extracted envelope (i.e. fitting the simulation like the data to extract the envelope given by Eq. S47) of simulations with log-normal weights for all models whose mean field fit has a total error smaller than 0.5. **b) Direct comparison with experimentally reported synaptic connectivity.** Top: Mean of the distributions of weights shown in 1, normalized by the synaptic weight from PV to E, for comparison with the available experimental data (background grayscale in panel 1**e**). Bottom Left: Synaptic weight connectivity as obtained in (Pfeffer et al., 2013). Bottom 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 (Pfeffer et al., 2013). These two matrices are shown here for comparison. There is currently no agreement on the strength of the connection from PV to VIP.

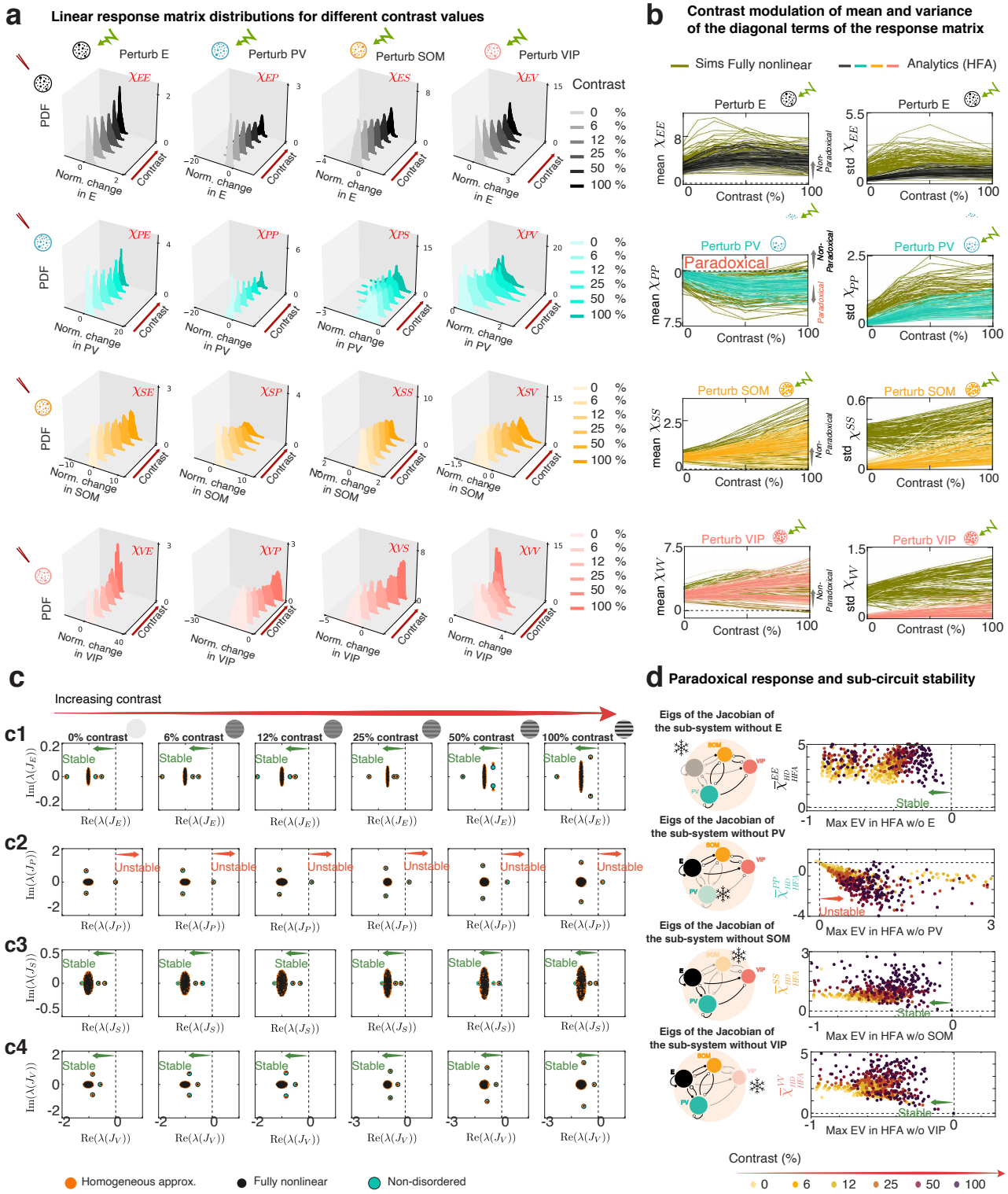
a
Linear response matrix distributions for different contrast values
Perturb E
Perturb PV
Perturb SOM
Perturb VIP
Contrast
0 %
6 %
12 %
25 %
50 %
100 %
PDF
Norm. change in E
Norm. change in PV
Norm. change in SOM
Norm. change in VIP
b
Contrast modulation of mean and variance of the diagonal terms of the response matrix
Sims Fully nonlinear
Analytics (HFA)
mean
std
Contrast (%)
Paradoxical
Non-Paradoxical
c
Increasing contrast
c1
c2
c3
c4
0% contrast
6% contrast
12% contrast
25% contrast
50% contrast
100% contrast
Stable
Unstable
Homogeneous approx.
Fully nonlinear
Non-disordered
d
Paradoxical response and sub-circuit stability
Eigs of the Jacobian of the sub-system without E
Eigs of the Jacobian of the sub-system without PV
Eigs of the Jacobian of the sub-system without SOM
Eigs of the Jacobian of the sub-system without VIP
Max EV in HFA w/o E
Max EV in HFA w/o PV
Max EV in HFA w/o SOM
Max EV in HFA w/o VIP
Contrast (%)

Figure S2 **a) Response distributions to full population perturbations for different values of contrast in large-scale models.** Distribution of linear responses to homogeneous full-population perturbations for different values of the contrast for all those models chosen in Fig. 1. **b)** Left panels: 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). Right panels: Standard deviation of the response. Note the paradoxical response of PV at all contrasts for almost all models 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 (Miller and Palmigiano, 2020). Nevertheless, none of the data-compatible models obtained had an excitatory paradoxical mean response.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. **c) Linking stability and response to full population perturbations in large-scale models.** $\mathbf{c}_1$)Eigenvalues of the Jacobian of the sub-circuit without the entire E population (i.e. J_E) as a function of contrast. The response of pyramidal cells is never paradoxical and the sub-circuit without it is always stable, both in the system using the HFA approximation (orange), the non-disordered system (i.e. the LD system, turquoise), and the fully-non linear system without approximations (black). $\mathbf{c}_2$**)** Idem c_1) for the sub-circuit without PV (J_P). The sub-system without PV is always unstable, with only one eigenvalue positive, guaranteeing a paradoxical response (see also Miller and Palmigiano, 2020). $\mathbf{c}_3$**)** for the sub-circuit without SOM (J_S), although this sub-circuit approaches instability for large contrast in this particular example it remains non-paradoxical. $\mathbf{c}_4$**)** for the sub-circuit without VIP (J_V). **d)** 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.

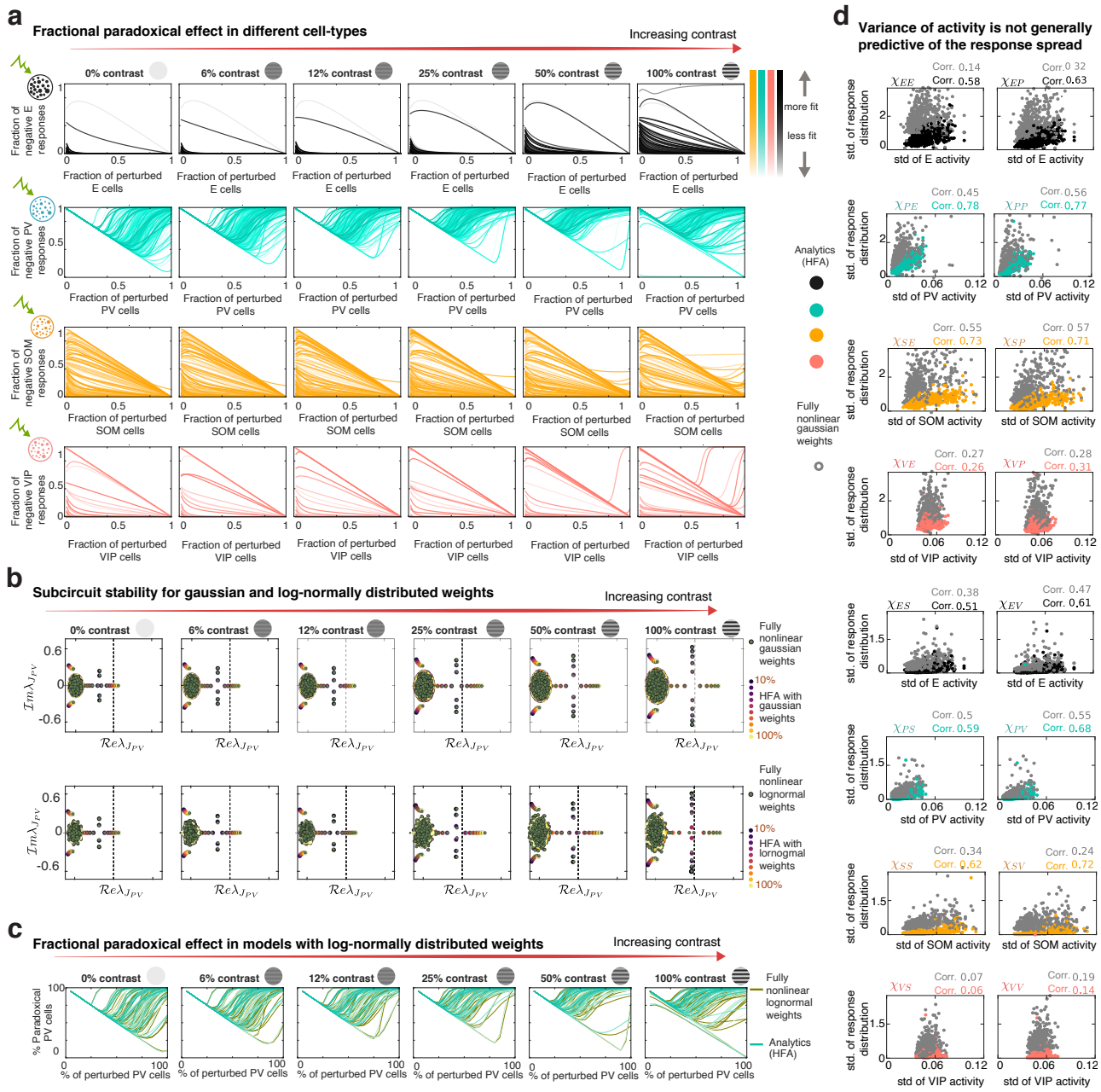

Figure S3 **a) Fractional paradoxical effect in the family of large-scale models with the HFA.** Dependence of the fraction of negative responses of the E (top), PV (middle top), SOM (middle bottom) and VIP (bottom) cells as a function of the fraction of stimulated cells in the same population. Shaded color indicates fitness. Notice how few models have a VIP fractional paradoxical effect in VIP at high contrast. **b) Linking stability and response to partial perturbations in large-scale models.** (Top) Eigenvalue distribution of the sub-circuit obtained by 'freezing' or removing different percentages of PV neurons from it, for an example model with Gaussian distributed weights. Different colors indicate different fractions of removed PV neurons in the HFA network (dark blue without 10% of PV neurons, yellow without the entire PV population). Dots in green are the corresponding eigenvalues of the sub-circuit in the fully nonlinear case. The eigenvalue distribution in the HFA and the fully nonlinear case match perfectly. (Bottom) for the same example network but with log-normally distributed weights. **c) Fractional paradoxical response for networks with log-normally distributed weights.** The models show a fractional paradoxical effect which is in good agreement to the effect described in networks with HFA with Gaussian weights, for which the theory of response to perturbations is exact. **d) Spread of the activity distribution is a poor predictor of the spread of optogenetic responses.** Standard deviation of the responses of each cell type to perturbations to each cell type as a function of the standard deviation of the activity, for all values of contrast. We see that the correlation between the spread of these distributions is only mild, both under the analytical approximation (black E, turquoise PV, orange SOM and pink VIP), and for the fully nonlinear system in grey. The spread of the distribution of activity is therefore not a strong predictor of the distribution of responses

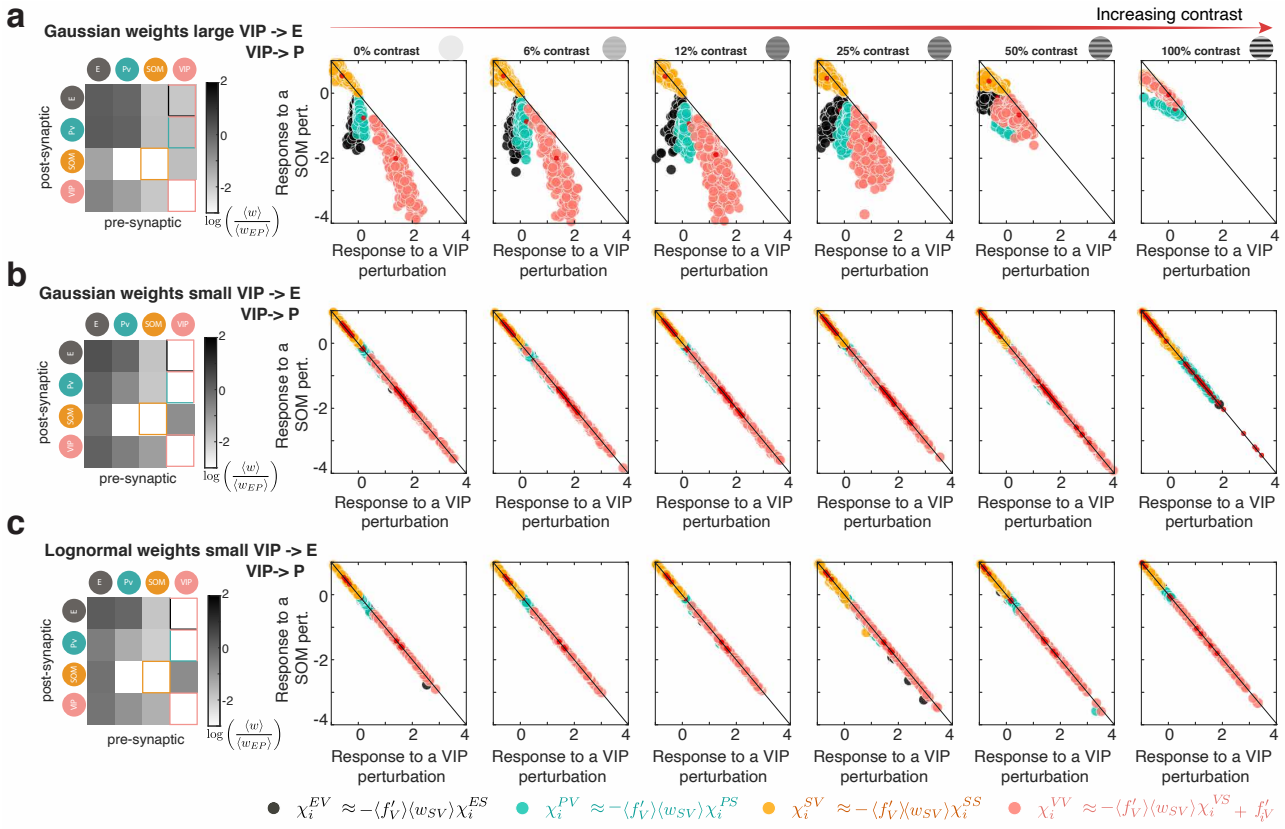

Figure S4 Disinhibition symmetries in the response only depends on the VIP projections and not the shape of the weight distribution. **a)** Left: Mean connectivity of a model in which the mean projections from VIP to E (w_{EV}) and to PV (w_{PV}) are large. Right: Response of each cell to a full SOM perturbation vs the response to a full VIP population, In this case, the LD hidden symmetries are not respected (red points , as in Fig. 3, are outside the negative diagonal). Furthermore, the entire distribution of responses to a VIP and a SOM perturbation does not obey a contrast-independent relation. Still, there is a clear correlation between the response to a SOM and a VIP perturbation for each cell. **b)** Model with KL error smaller than 0.5 and $w_{EV} > -0.1$. Left: mean connectivity across models in this configuration. Right: All these models show clear HD symmetries. **c)** Top 10% of models with KL error smaller than 0.5 and $w_{EV} > -0.1$, simulations done with a log-normal distribution of weights. Left: mean connectivity across models in this configuration. Right: All these models, even with heavy tailed weight distributions, show clear HD symmetries.

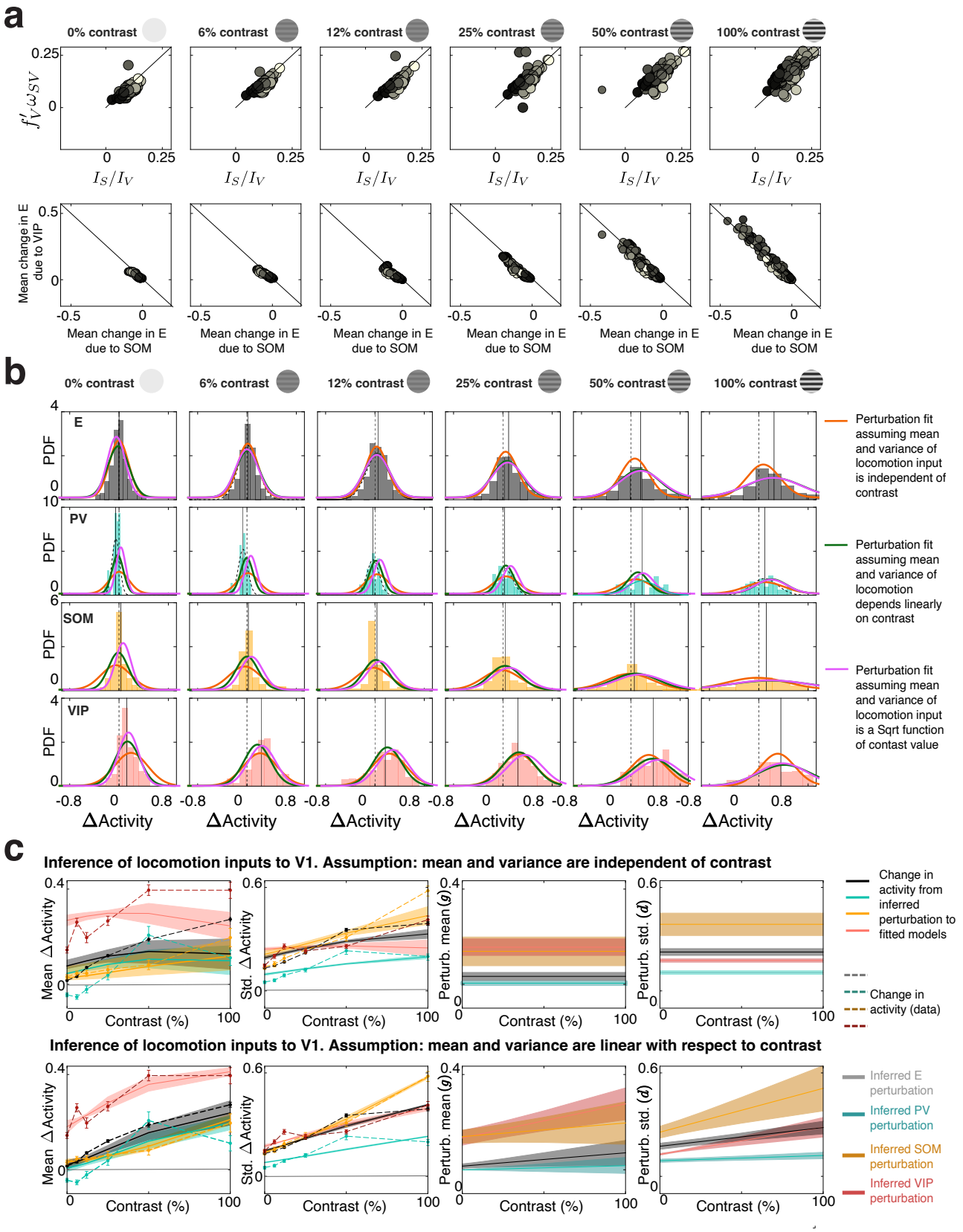
a
0% contrast
6% contrast
12% contrast
25% contrast
50% contrast
100% contrast
$f'_V \omega_{SV}$
I_S/I_V
Mean change in E due to VIP
Mean change in E due to SOM
b
E
PV
SOM
VIP
PDF
ΔActivity
Perturbation fit assuming mean and variance of locomotion input is independent of contrast
Perturbation fit assuming mean and variance of locomotion depends linearly on contrast
Perturbation fit assuming mean and variance of locomotion input is a Sqrt function of contast value
c
Inference of locomotion inputs to V1. Assumption: mean and variance are independent of contrast
Mean ΔActivity
Std. ΔActivity
Perturb. mean (g)
Perturb. std. (d)
Contrast (%)
Change in activity from inferred perturbation to fitted models
Change in activity (data)
Inference of locomotion inputs to V1. Assumption: mean and variance are linear with respect to contrast
Inferred E perturbation
Inferred PV perturbation
Inferred SOM perturbation
Inferred VIP perturbation

Figure S5 Strong and wide inputs to VIP and SOM lead to network responses that mimic the effect of locomotion. **a)** Top: Scatter plot showing that the disinhibition symmetries underlie behavioral modulations of activity. The theory predicts that, if the ratio of the inputs to SOM and VIP are proportional to the mean VIP gain (under the HFA), then the net effect of those perturbations in E will indeed cancel (see Eq. S30). Here we show that when we infer the locomotory inputs to SOM and VIP, they are proportional to the mean VIP gain times the mean strength of the connection from VIP to SOM (different for each model). This is the case even when the stimulus contrast is large and the mean change in E due to locomotion no longer vanishes (see below). Bottom: Change in E activity due locomotory inputs to SOM Vs due to locomotory inputs to VIP. At the maximum contrast, models largely lie on the diagonal and only few models show disinhibition (i.e inputs lead to an increase in E activity). **b)** 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. Dashed lines indicate best Gaussian fits. Solid lines are fits from the explicit expressions (see Eqs. S126 and S129) under three different assumptions i) mean and variance of the inputs to each population are independent on contrast (orange, corresponding to panel b) ii) mean and variance of the inputs to each population depend linearly on contrast (green, corresponding to panel c) and ii) mean and variance of the inputs depend as the square root of the contrast (magenta, same curves as in main text. **c** Top: 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) and full lines are the mean of the distribution of fits as described in **a**, for the family of models that fit the stationary data, under the assumption that locomotory inputs do not depend on contrast. The mean (g) and the standard deviation (d) of the inferred perturbation are shown in the right panels, and independent of stimulus properties. Bottom: Same as top but requiring the locomotory inputs to depend linearly on the contrast.