Inhibition stabilization is a widespread property of cortical networks

A. Sanzeni*,[1,2] B. Akitake*,[1] H. C. Goldbach,[1] C. E. Leedy,[1] N. Brunel,[2,3] and M. H. Histed[1]

[1] *NIMH Intramural Program, National Institutes of Health, Bethesda MD 20892, USA*

[2] *Department of Neurobiology, Duke University, Durham, NC 27710, USA*

[3] *Department of Physics, Duke University, Durham, NC 27710, USA*

I. ABSTRACT

Inhibitory-stabilized network (ISN) models of the cerebral cortex predict that local (recurrent) synaptic connectivity plays a vital role in shaping cortical activity. Yet it has been experimentally unclear if cortical recurrent connectivity is strong enough to make ISNs good descriptions of cortical function. Here we test several ISN predictions, including the counterintuitive (paradoxical) suppression of inhibitory firing when the inhibitory network is stimulated. We pair, in awake mice, optogenetic stimulation of all inhibitory subtypes with in vivo pharmacology to identify inhibitory cells. We observe responses consistent with an ISN in the upper layers of visual, somatosensory, and motor cortex. Stimulating parvalbumin (PV)-positive inhibitory neurons produces a population paradoxical effect only with transgenic, not viral, opsin expression. This effect is explained in a model where viral expression targets a subset of PV cells, showing inhibitory cell responses to stimulation can be highly dependent on the number of stimulated cells.

II. INTRODUCTION

Extensive recurrent connectivity between nearby neurons is an ubiquitous feature of the cerebral cortex [1; 2; 3; 4]. Theoretical work has shown that strength of recurrent coupling has a major impact on several computational properties of networks of excitatory (E) and inhibitory (I) neurons, including the speed of the response to external stimuli [5; 6], the ability of a network to sustain persistent activity [7], the capacity and robustness of memory storage [8], and how various input modes are amplified [9].

Strong excitatory recurrent coupling, however, can lead to unstable dynamics, unless stabilized by inhibition. When recurrent connectivity is weak, excitatory cells can show stable firing rates independent of the activity of inhibitory cells. But in networks with strong recurrent connections, excitatory-to-excitatory (E-E) connections amplify responses so that the excitatory network is unstable if the firing rates of inhibitory neurons are kept fixed. Stable excitatory network operation across a range of firing rates can be

restored if inhibitory recurrent connections are sufficiently strong, allowing inhibition to track and balance excitation [5; 6; 7; 10; 11]. Such network models, requiring inhibition for stability, are called inhibitory-stabilized networks (ISNs). Examples of ISNs are balanced network models [5; 6; 12] in which both excitatory and inhibitory inputs are extremely large, but balance each other, and the stabilized supralinear network (SSN) [13; 14] which can show transitions from non-ISN to ISN behavior as the inputs to the network are increased.

A prominent prediction of ISN models is *paradoxical inhibitory suppression*[10] — a counterintuitive decrease of inhibitory firing rate as inhibitory cells are stimulated (Fig. 1b). This paradoxical effect is a result of the strong synaptic coupling between E and I cells. In an ISN, the initial increase in inhibition due to I stimulation is amplified by the strong E-E coupling, leading to a decrease in excitatory input to I cells that exceeds the added external drive to I cells (see Fig. S12). However, this paradoxical effect in ISN models is only observed in a finite range of inputs. For large inputs to the I population, the E population is completely suppressed, and hence can no longer provide a sufficient decrease in E inputs to counterbalance the inputs to I neurons.

Experimental evidence for inhibition stabilization in the cortex has been mixed. Ozeki et al [15] found indirect evidence for inhibition stabilization through measurements of surround suppression in cat V1. In the last decade, optogenetic perturbation studies have allowed experimentalists to directly test the 'paradoxical effect' prediction. Both paradoxical and non-paradoxical effects have been reported [16; 17; 18; 19]. However, many of these studies used opsins expressed in specific classes of interneurons, such as PV- or somatostatin(SOM)-positive neurons. As we explore further below, when only a subset of inhibitory cells is stimulated, non-paradoxical effects can result, even in an ISN. Thus one explanation for non-paradoxical effects in these studies could be that only a fraction of cells is stimulated. On the other hand, when multiple inhibitory populations are involved, paradoxical effects can be generated without inhibition stabilization via a disinhibition mechanism (see also discussion in [20]).

Therefore, we set out to examine whether cortical networks operate as ISNs (Fig. 1) by combining optogenetic stimulation of inhibitory cells, extracellular recordings in awake animals, and theoretical analyses to test a larger set of ISN predictions than has been previously achieved. We used a mouse line in which all I cells respond to optogenetic stimulation to directly test the ISN paradoxical effect prediction. In a subset of experiments, we used pharmacology to weaken recurrent connections, as one criterion for classification of I cells. (Because I cells express the opsin, they respond to optogenetic stimulation even when recurrent connection strength is reduced.) We find that measured responses in visual, somatosensory, and motor cortex under different experimental conditions are in quantitative agreement with a network model operating in the ISN regime. Finally, stimulating PV neurons in V1 in experiments using viral or transgenic expression strategies, we find differences in paradoxical suppression that can be explained by a model in which the fraction of stimulated PV cells is varied. Our results provide multiple lines of evidence for the ISN scenario, and also experimentally show why past evidence for cortical ISN behavior has been mixed: inhibitory cells' responses to stimulation can dramatically change depending on the fraction of stimulated I neurons.

III. RESULTS

Mouse primary visual cortex is inhibition stabilized

We first describe experiments where all inhibitory neurons were optogenetically stimulated. We expressed an excitatory opsin (Channelrhodopsin-2; ChR2) in all inhibitory neurons using a transgenic mouse line (VGAT-ChR2 [21]). We delivered blue light to the surface of the cortex, while recording activity extracellularly with multi-site silicon recording arrays (Fig. 1c). Mice were given drops of reward approximately once per minute (Methods) to keep them awake and alert. Because neurons in the superficial cortical layers showed the largest responses to the stimulation light (Supp. Fig. S3), we restricted our analyses to units within 400um of the cortical surface (Methods), which primarily includes neurons in the superficial layers

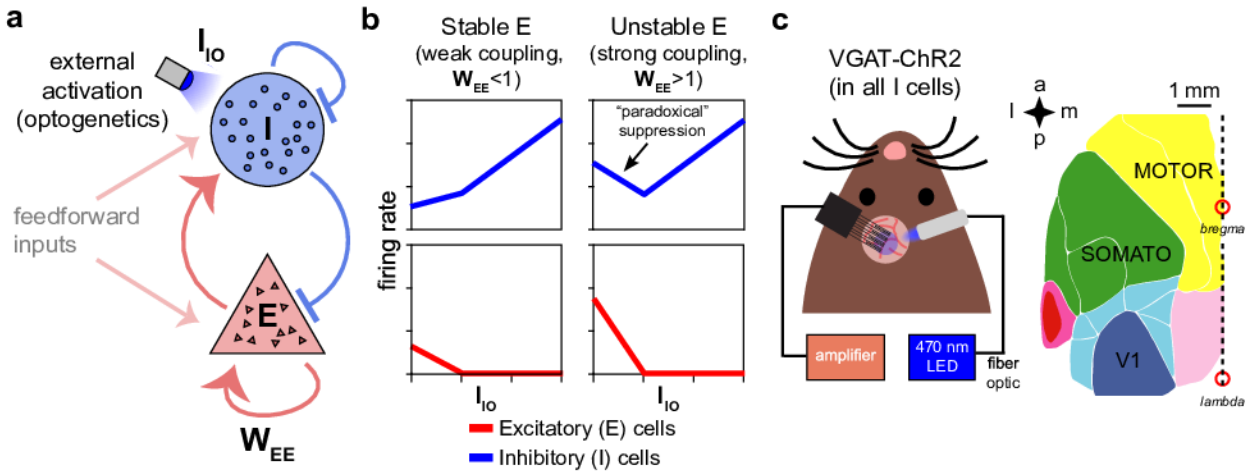

FIG. 1 **Model predictions of E and I responses to I stimulation.** (**a**) Schematic of model, showing connections between excitatory (E) and inhibitory (I) neuron populations. (W_{EE}, the average strength of E-E connections, is the key parameter controlling non-ISN vs ISN operation, see model discussion in Results.) (**b**) Predictions for average neural responses with weak recurrent coupling (left) and strong coupling (right), when inhibitory cells are externally stimulated (I_{IO}). (**c**) Schematic of experiment, extracellular recordings in visual (V1), primary somatosensory, and motor/premotor cortices (see Supp. Fig. S1 for electrode locations) while optogenetically stimulating all inhibitory cells at the recording site in awake VGAT-ChR2-YFP animals.

(2/3 and upper layer 4 [22]). We sorted units into single- (separated from noise and other units) and multi-units (lower SNR or containing several apparent units) by waveform (Methods; using a more restrictive threshold for unit inclusion does not affect our conclusions, Supp. Fig. S4), and we excluded the multi-units from further analyses.

The majority of recorded units were suppressed by light (146/167 single units, 87%; Fig. 2b-d; 4 animals, 7 recording days). Given that approximately 80% of cortical neurons are excitatory [23; 24], and that extracellular recordings can show several sorts of biases in sampling cortical neurons, [25; 26], this measured percentage could be consistent with either the presence or absence of paradoxical suppression in inhibitory neurons. Therefore, we classified recorded units as inhibitory or excitatory using *in vivo* pharmacology. (Our results are confirmed by other classification methods: identifying inhibitory neurons using their response at high laser intensity or by waveform width give similar results, see below and Fig. S7.)

We first recorded neurons' responses to optogenetic stimulation, and then applied blockers of excitatory and inhibitory synapses (CNQX, AP5, bicuculline; which block AMPA, kainate, NMDA, and GABA-A synapses; Methods). When recurrent synapses are substantially suppressed, inhibitory cells (which express ChR2) should increase their firing rate to optogenetic input, while excitatory cells' firing should not increase. Indeed, by measuring responses to I stimulation in the presence of E and I blockers, we were able to classify units into two groups (Fig. 2e). Units in the first class (putative excitatory) were silent for strong stimulation and units in the second class (putative inhibitory) increased their firing rate in response to stimulation. To quantitatively classify cells, we measured the change in activity induced by light when the blockers had been applied (Fig. 2f-g); cells were labeled inhibitory (33%; 56/167 units) if the change compared with baseline produced at maximum laser intensity was positive (according to Welch's t-test, $p < 0.01$), and excitatory otherwise. Choosing different classification thresholds did not affect our results (Supp Fig. S6), examining a subset of cells with the most stable waveforms over time did not affect our results (Fig. S5), and, finally, examining a subset of I units where waveforms had high signal-to-noise ratio also did not qualitatively change the results (Supp. Fig. S4). Therefore, we used these response differences in the presence of blockers to identify recorded inhibitory units.

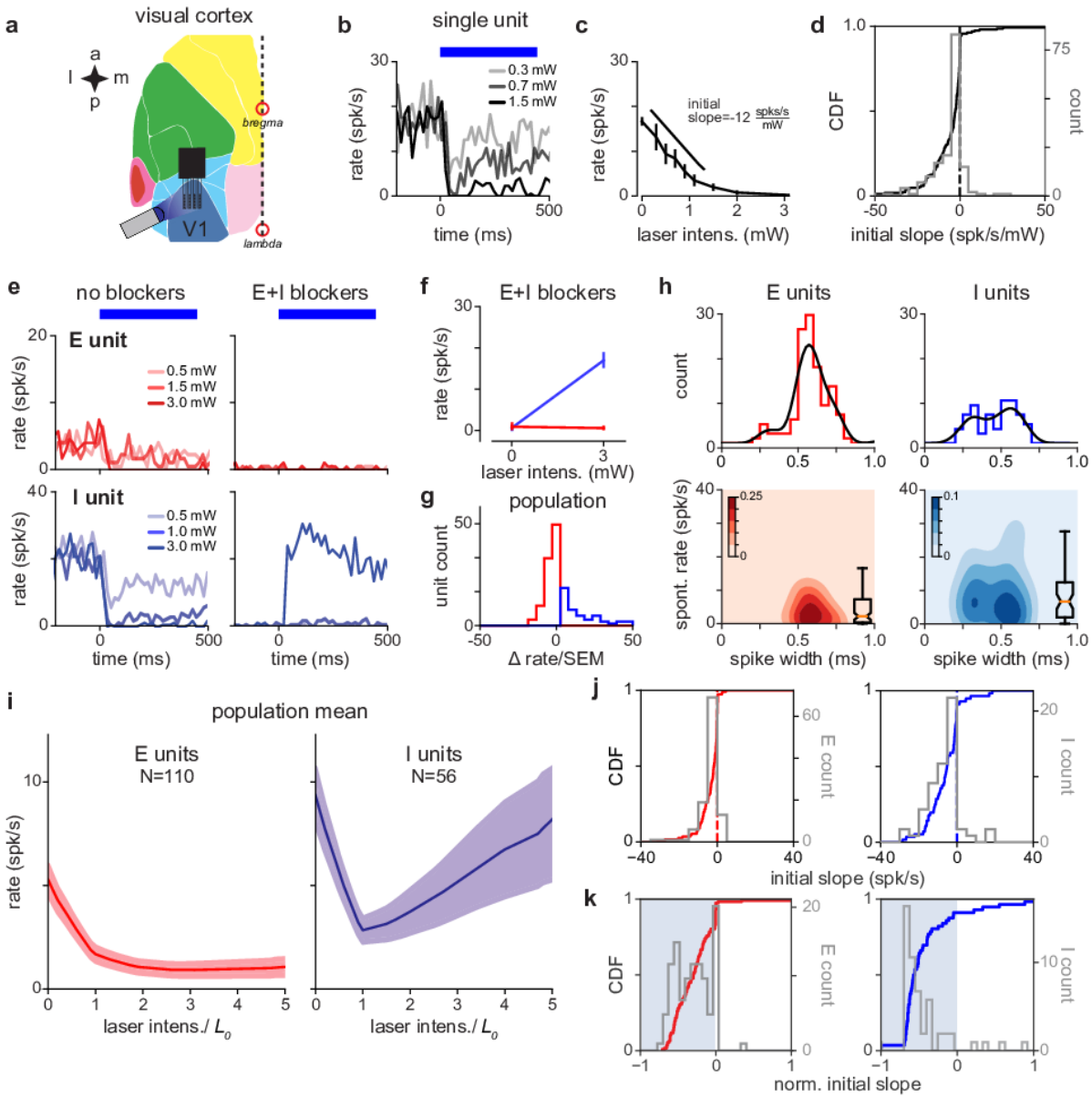

FIG. 2 **V1 response is consistent with inhibition stabilization.** (**a**) Schematic of experiment, showing recordings/stimulation in V1. a: anterior, p: posterior, l: lateral, m: medial. V1 recording locations and intrinsic signal imaging to define V1 shown in Supp. Fig. (S1). (**b**) Example recorded unit; greater laser power produces greater firing rate suppression. Blue bar: duration of constant-power laser pulse. (**c**) Average firing rate of example recorded unit (rate is steady-state firing rate, over last 300 ms of laser stimulus). Suppression is quantified via initial slope of firing rate vs laser intensity (Methods). (**d**) Distribution of initial slopes for all recorded units. (**e**) Example cell responses without (left) and with (right) blockers. Top: Cell classified as excitatory; does not respond to laser in presence of blockers (CNQX, APV, bicuculline; Methods). Bottom: Cell classified as inhibitory; increases firing rate to laser in presence of blockers. (**f**) Firing rates for units in (e) in presence of blockers; I cell (blue) shows a large positive change. (**g**) Distribution of firing rate changes to laser stimulus with blockers. Red: E-classified units; blue: I-classified units. (**h**) Spontaneous firing rates and waveform width (time from waveform first local minimum to first local maximum) for units classified as E and I via the method in g. (**i**) Average population response for units classified as E and I. I cells show a prominent paradoxical effect. (To account for variations in tissue optical properties across recording days, we combined data by fitting a piecewise-linear function to the average inhibitory response for each day to find a change in slope (L_0: always found to be a minimum between 0.3 and 2.7 mW) and rescaling the laser intensity so $L_0 = 1$.) (**j**) Distribution of initial slopes for units classified as E (red, left) and I (blue, right). (**k**) Distribution of normalized initial slopes ($\frac{\text{init. slope}}{\text{baseline rate}+0.1\text{spk/s}}$), to more clearly show small slopes, for units classified as E and I. Several E cells with very low firing rates in the no-blocker case have near-zero normalized slopes.

Pharmacologically-identified inhibitory and excitatory units showed differences in waveform width and spontaneous firing rate (Fig. 2h), though these two factors were insufficient alone to completely classify inhibitory cells (i.e. insufficient to predict whether units would increase firing rate in the presence of blockers). Identified excitatory cells had a wider waveform (0.58±0.01ms) and lower firing rate (6.0±0.8 spk/s) than inhibitory cells (width 0.46±0.02 ms. rate 10.7±1.7 spk/s). However, inhibitory neurons' width was broadly distributed, supporting the idea that that some inhibitory classes can have broader waveforms than other inhibitory classes [27; 28].

Once recorded units were classified as excitatory or inhibitory, we could ask whether the same units before adding blockers (i.e. with the animal in the awake state, with recurrent connections intact) showed paradoxical inhibitory effects. Indeed, inhibitory neurons showed strong suppression when stimulated (Fig. 2i-j). This paradoxical inhibitory suppression is one piece of evidence that an ISN model is a good description of the upper layers of primary visual cortex. (Below we show that response dynamics and responses at high laser intensity are additional evidence for an ISN.)

These experiments were done without sensory stimulation, as animals viewed an unchanging neutral gray screen. Because visual cortical neurons' firing rates increase with increasing contrast across a range of overall luminance levels, we expected that we would see little difference between data collected with a neutral gray screen and data collected in the dark. Confirming that expectation, in one experimental session we measured inhibitory responses to stimulation while animals were in the dark (with the blue optogenetic light shielded with light-blocking materials), and found paradoxical suppression (N=6/6 inhibitory units with initial negative slope, median initial slope -6.9 ±3.5 spk/s; p=0.046, median < 0, Mann-Whitney test).

In an ISN, paradoxical inhibitory suppression arises because exciting inhibitory cells in turn suppresses excitatory cells, which withdraw excitation from the stimulated inhibitory cells. Then, because more excitatory input from recurrent sources is withdrawn from I cells than excitatory input from optogenetic stimulation is added, the steady-state response of I cells is suppression. For this mechanism to hold, inhibitory cells must transiently increase their firing, even if only a small number of inhibitory spikes are generated, before the paradoxical suppression [10; 15]. Therefore, we looked for transient increases in inhibitory firing with short latencies after stimulation, and found a small increase in inhibitory firing rate immediately after stimulation, before the larger paradoxical suppression (Fig. 3). The transient was brief (FWHM 7.2 ms), and so the actual number of extra spikes fired in the transient was small: only 0.03 extra spikes were fired on average per inhibitory-classified unit per stimulation pulse, though because we used 100 pulse repetitions, the majority of classified inhibitory units showed a detectable initial transient (Fig. S2).

Another prediction is that I cells should increase their firing rates at high laser intensities. In a recurrent network operating as an ISN, stimulating I cells first decreases I firing rates, but stronger stimulation eventually suppresses E cells until they can decrease no further. Then, increasing stimulation of I cells produces increases in I firing rates, as the network moves into a non-ISN regime (Fig. 1C). Indeed, we observed this in our E and I identified populations (Fig. 2H).

While our extracellular recording approach has advantages, such as allowing precise measurements of spike rate dynamics, and recording a large number of E and I neurons in the same experiment, one concern might be that misclassification of E waveforms into I units could mask I cell responses. However, confirming that our recording procedures accurately measured inhibitory waveforms, (1) paradoxical suppression is maintained when we make our classification thresholds more stringent (restricting analysis to units with highest waveform signal-to-noise ratio, Supp. Fig. S4; varying classification threshold in the presence of blockers; Supp. Fig. S6), (2) units classified by spike width, or by response at high laser power, also showed clear paradoxical effects (Supp. Fig. S7), and (3) inhibitory, but not excitatory, response dynamics showed an initial increase followed by suppression (Fig. 3), the response patterns expected for those two classes of neurons in an ISN, with magnitudes difficult to explain by waveform mixing (Supp. Fig. S2), Also, as shown below, in parvalbumin-positive neuron stimulation experiments using the same waveform-sorting procedures, we observed no average paradoxical effects, arguing that paradoxical effects in VGAT-ChR2 animals do not arise due to waveform sorting, but instead due to the properties of the cortical inhibitory network.

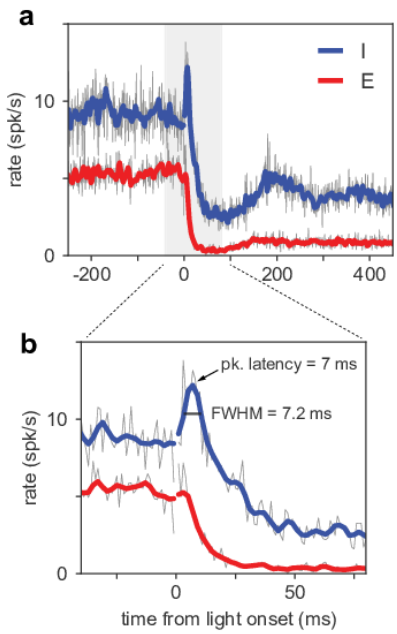

FIG. 3 **Inhibitory neurons show a small increase in firing before paradoxical suppression.** (**a**) Average timecourse of neural responses from units classified as E (N=110) and I (N=56). Light pulse is constant-intensity and lasts for 800 ms (power: $2.6L_0$, see Supp. Fig. S2). Data from VGAT-ChR2 animals. Gray lines: rates calculated in 1 ms bins. Heavy lines: 1 ms-bin data smoothed with a LOWESS filter. Inhibitory neurons show a brief increase in firing before suppression. (**b**) Same data as (a), enlarged to show the initial transient (time range here is indicated in (a) by gray shaded region). Inhibitory initial positive transient has peak amplitude 3.6 spk/s above baseline (rate over 12 ms window after pulse is greater than a matched-duration interval before pulse, $p < 0.01$, KS test), latency to peak 7 ms, full width at half maximum 7.2 ms, and inhibitory rate crosses baseline into suppression at 13.1 ms.

Taken together, these results show that mouse V1 responses are consistent with those of a strongly coupled network whose activity is stabilized by inhibition.

Model-based inference of network coupling strength and stability

We next used these data to infer network connection parameters in a standard two-population model, that describes the dynamics of population-averaged firing rates of excitatory and inhibitory neurons (see Eqn. 6; Methods). This parameter inference is possible because we obtained E and I response measurements in three pharmacological conditions (no blockers; with E blockers; with E and I blockers; Fig. 4a), and with data from these three conditions, there are more experimental observations than model parameters (Methods: Model degrees of freedom). The model uses a rectified-linear single-neuron transfer function, and using a nonlinear transfer function gives similar results (Supp. Fig. S11).

First, we find that the model, despite being overconstrained by the data, is a good description of V1 responses (Fig. 4a). The model also makes a number of predictions that are verified in the data. It predicts that the I firing rate slope for high laser intensities should be the same in the no-blocker and E-blocker conditions, as seen in the data (Fig. 4a, blue: left vs. middle columns). The model makes three other predictions: that the reversal point of inhibitory cells (L_0, the point at which initial negative inhibitory slopes become positive), in all three experimental conditions, should match the point where excitatory cells change their slope to become nearly silent. These effects are also seen in the data (e.g. clearly visible in the no-blocker case, Fig. 4a, left column).

One prediction of balanced-state models with strong recurrent coupling [5] is the total excitatory and

inhibitory currents should each be large, though they cancel to produce a small net input. To check this, we computed from the model the currents flowing into excitatory cells (Fig. 4b). The excitatory and inhibitory currents are each approximately ten times larger than threshold (medians are 10.9 and -8.63, respectively), and their sum is small compared to the magnitudes of E and I currents (median net current/threshold = 2.2). The model also allows us to infer the value of the key parameter for inhibition stabilization, W_{EE} (the excitatory self-amplification). The excitatory network is unstable, requiring inhibition for stabilization, and yielding an ISN, when $W_{EE} > 1$. As expected when paradoxical effects are present, the fitted value is greater than 1 (Fig. 4c: median 4.7, mode 2.5). This value is not much larger than 1, suggesting recurrent coupling is not as strong as what the balanced-state model would predict [5; 6; 7] but is rather more consistent with a moderately-coupled network [13; 14; 29].

Finally, we computed the parameter ranges over which the network is stable, as a function of the time constants of E and I firing rate dynamics, τ_E and τ_I. We find, by computing the eigenvalues of the connection matrix (Methods) that the network is stable for τ_I/τ_E in the interval $[0, 5.3]$, and the observed dynamics are consistent with $\tau_I/\tau_E = 4.4$ and $\tau_E = 7.8$ms (Fig. 4e); that is, with excitatory and inhibitory synaptic time constants of the same order. The values of $\tau_{(E,I)}$ inferred from the response dynamics are within the stable region, providing further support that the ISN model is a good description of the effects we measure. (Note that the time constants of the opsin may increase these estimates, so the values of τ_E and τ_I, while small, are only upper bounds on the true dynamic response of the network to an instantaneous conductance change.) The ratio between E and I time constants is close to the point at which the network becomes unstable through an oscillatory instability (Fig. 4e). The fact that the inferred network model describes the data and is stable with $W_{EE} > 1$ supports the experimental observation that the underlying cortical network is inhibition-stabilized.

Inhibition stabilization is present in other cortical areas

The data above show that the superficial layers of mouse V1 are inhibition stabilized. Another open question is how general the ISN regime is across cortical areas. We repeated our procedure to look for signatures of inhibitory stabilization in two other cortical areas (Fig. 5): somatosensory (body-related primary somatosensory, largely medial to barrel cortex) and motor/premotor cortex. (See Supp. Fig. S1 for locations of all recording sites.) As in V1, we recorded units extracellularly, restricted analysis to the superficial layers of the cortex, and examined responses to optogenetic stimulation of all I cells (as above, using the VGAT-ChR2 line).

In these experiments, we classified neurons as inhibitory based on their activity at high laser intensity. In our V1 experiments, paradoxical effects are clear whether neurons are classified in this way, or with the pharmacological method (Supp. Fig. S7). This high-laser-power classification is likely more stringent (i.e. may reject some I cells if they are suppressed by other I cells, due e.g. to heterogeneity of recurrent connections) than pharmacological classification. Supporting that this classification may be more stringent, some pharmacologically-classified inhibitory cells show paradoxical effects but little increase at high power (Supp. Fig. S9).

In both the recorded areas (Fig. 5), inhibitory neurons showed paradoxical effects, supporting that these cortical areas do also operate as an ISN. Similar to the V1 data, and as expected in an ISN, both areas showed a transition to a non-paradoxical response at large laser intensity. Compared to the somatosensory data, the motor/premotor recordings show a smaller average suppression (Fig. 5b) and more I units with non-paradoxical effects (rate increases with stimulation; Fig. 5d). However, in both areas, means and medians are significantly negative (see legend for statistical tests). Thus, the superficial layers of both sensory cortical areas we examined (visual and somatosensory), and one non-sensory area (motor/premotor) showed responses consistent with strong coupling and ISN operation.

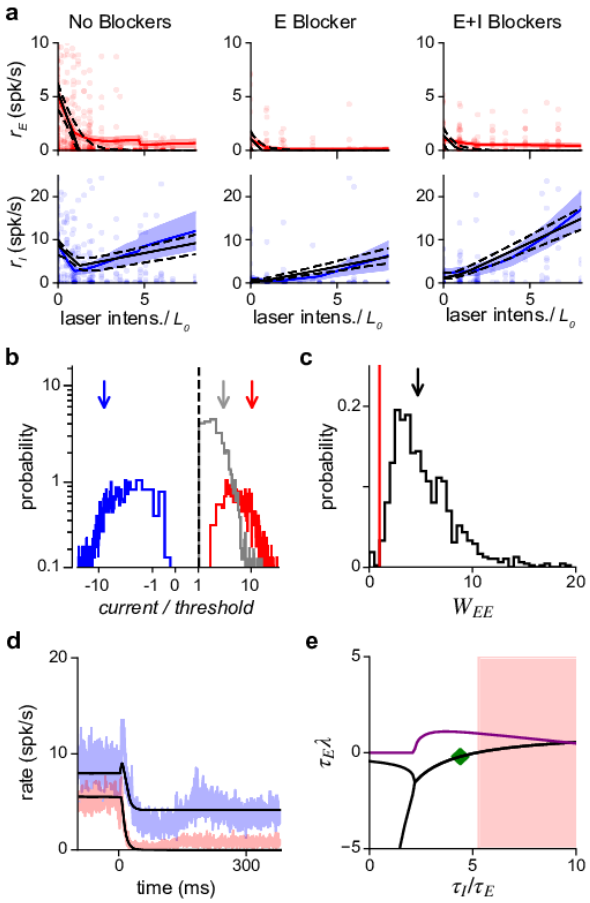

FIG. 4 **Population response is consistent with a network with moderately strong coupling** (**a**) Average excitatory (red) and inhibitory (blue) response measured in three conditions: without synaptic blockers, with only the E blocker present, and with both E and I blockers present. Model (Eq. 6) that best fits data: black continuous line; dashed: 1 s.d. obtained by bootstrap; solid red/blue lines: E/I data means; shaded region: ±1 SEM. Because we applied blockers sequentially (generating separate E and E+I blocker measurements) the number of independent observations were increased, allowing the inference of all model parameters (Supplementary Methods; S10). Small steps in left panels (no blockers) arise because a subset of experimental days used a maximum laser intensity of $5 \cdot L_0$. (**b**) Excitatory (red), inhibitory (blue), and net (gray) current influx into excitatory cells predicted by the model. Arrows: medians (over bootstrap repetitions); E: 10.9, I: -8.63; modes: E: 2.5, I: -2.7. (**c**) Distribution of W_{EE} values compatible with the data; the red line represents the transition point between the ISN ($W_{EE} > 1$) and the non-ISN ($W_{EE} < 1$) regime. Median (arrow) 4.7; mode 2.5. (**d**) Estimation of time constants of E and I populations. Black line shows the dynamics resulting from fitting the data (blue: I population; red: E population; shaded region ±1 SEM) with the model for the same laser power shown in 3. The full model provides a good approximation to the dynamics even though it is constrained to simultaneously fit the time constants and the responses at different intensities (A). (**e**) Stability analysis. Real (black) and imaginary (purple) parts of the eigenvalues of the Jacobian matrix as a function of the ratio τ_I/τ_E. Imaginary part (purple) greater than zero signifies the network can show damped oscillations when being driven to a new stationary point. Note that these damped oscillations are not seen in (d) because of rectification. When the real part (black) is greater than zero, the network is unstable (shaded red area); best-fit $\tau_I/\tau_E = 4.4$ms, $\tau_E = 7.8$ ms.

Stimulation of different numbers of PV cells can explain observed differences in paradoxical suppression with viral or transgenic opsin expression.

Up to this point, we have studied paradoxical suppression by stimulating an opsin expressed in all inhibitory cells via the VGAT-ChR2 mouse line. A remaining question is whether stimulation of a subclass of inhibitory neurons also yields paradoxical suppression, as in principle, even in an ISN stimulation of any single subclass of I cells need not produce paradoxical suppression (see Methods and [17; 20]). We studied this by stimulating parvalbumin-positive (PV) neurons, which provide strong inhibitory input to other cells.

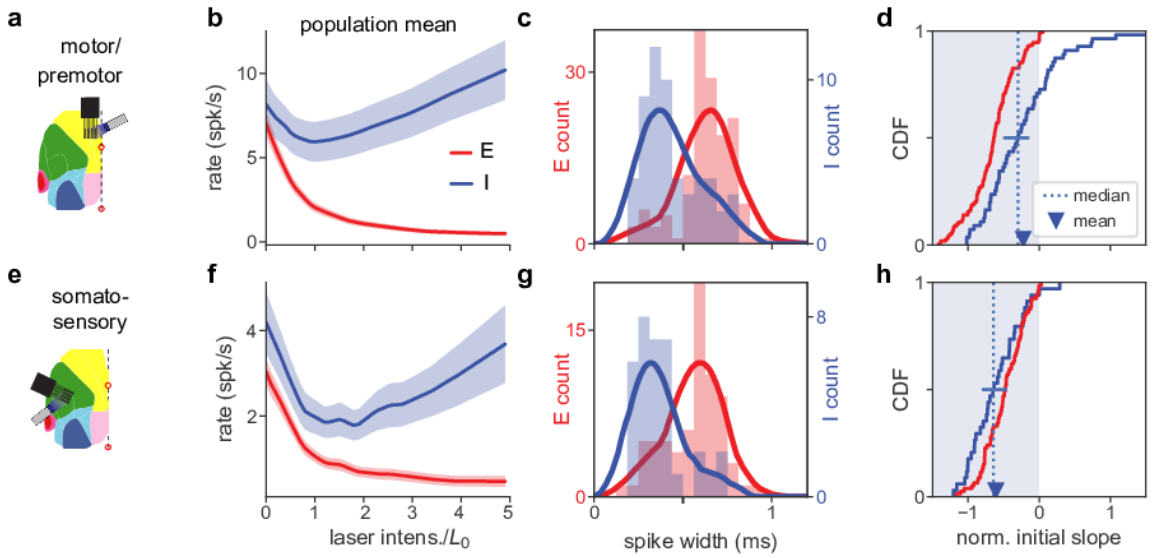

FIG. 5 **Inhibition stabilization across cortical areas. (a)** Motor (primary/secondary) cortex recordings (see Supp. Fig. S1 for recording locations). **(b)** Motor cortex population firing rates for E and I units. Initial mean response of inhibitory cells is negative, showing paradoxical suppression. Mean rate is significantly reduced ($p < 10^{-4}$, paired t-test, rate at 0 vs rate at L_0). **(c)** Spike width distributions for E and I units. Units are classified as E or I here by response at high laser power (Methods), independently of spike width, which nonetheless varies with E or I unit identity. **(d)** Normalized initial slope distributions for all units. E: red. Blue: I. Both mean and median of initial slopes are negative (paradoxical). Mean I slope is negative ($p < 0.01$, t-test). Horizontal bar at I median shows 95% confidence interval calculated by bootstrap. **(e-h)** Same as (a-d), but for recordings from somatosensory cortex. In (g), highest red bar is truncated for visual clarity (value is 22). Mean I rate (f) is sigificantly reduced ($p < 10^{-7}$, paired t-test). Mean I slope (h) is negative ($p < 10^{-10}$, t-test). Horizontal bar in (h) shows 95% CI around median slope.

PV basket cells are the most numerous class of cortical inhibitory cells and make strong synapses near the somata of excitatory cells (for review, see [24]), and PV stimulation effectively suppresses network firing rates (e.g. [30; 31]). We first used viral methods to express opsins in PV neurons, injecting a Cre-dependent adeno-associated virus (AAV) encoding an excitatory opsin (Chronos, [32]), into a transgenic mouse line (PV-Cre [33]).

In these experiments, the network neurons can be divided into three populations (Fig. 6a): (1) excitatory, (2) Chronos-expressing PV inhibitory (PV-Chronos), and (3) remaining inhibitory neurons: non-PV (e.g. somatostatin-positive, etc.) or non-Chronos-expressing PV neurons. We identified PV-Chronos cells by measuring whether cells' firing is increased at high laser intensity (statistically significant increase at maximum laser power, as in Fig. 5; see Methods). In V1 VGAT-ChR2 experiments, classification by response at high laser power produces qualitatively similar measures of paradoxical suppression as pharmacology-based classification, Supp. Fig. S7). Non PV-Chronos cells (without increase in firing at high laser intensity), are likely either E cells, non-PV, or non-expressing cells. Supporting the idea that our classification approach identifies PV-Chronos cells (note that while not all PV-positive cells are basket cells [34], many PV-positive neurons are basket interneurons and are fast-spiking cells with narrow waveforms), the majority of classified I cells have narrow waveforms (Fig. 6f).

Stimulating the PV cells produced no significant paradoxical effect on average (Fig. 6b, blue). From a theoretical point of view, this average I response is an important measure, as with the standard two-population ISN model, e.g. [10], it is the average response (averaged over I cells) that is paradoxically suppressed when a network is inhibition-stabilized and strongly coupled. A second important measure, e.g. for experimentalists wishing to identify inhibitory neurons, is how many individual I cells show paradoxical effects. Examining individual units showed that PV-Chronos units often showed no paradoxical effect:

initial slope could be positive and thus non-paradoxical (e.g. Fig. 6d; summarized in Fig. 6f-g, see legend for statistical tests.).

One reason why PV-Chronos stimulation might produce no paradoxical effect on average is that viral expression could target only a subset of PV cells. This could occur by the virus infecting only a subset of neurons [35], or by the virus yielding different levels of opsin expression in different neurons so that only a subset are strongly stimulated.

We first examined the effect of I cell subset stimulation in a model (the rate model of Fig. 4, using network parameters inferred there). Stimulating a subset of I neurons in the model (60%) reduced the magnitude of the average paradoxical effect (Fig. 6c, analytical derivation in Methods), and described the data (Fig. 6b) well.

The model predicts that increasing the fraction of stimulated cells will produce stronger paradoxical effects. Therefore, to determine experimentally if this reduced paradoxical effect could be due to stimulating a subset of PV neurons, we used a transgenic approach to express a different excitatory opsin in PV cells (PV-Cre;ReaChR transgenic mice; since we study steady-state firing rates (Methods) we do not expect differences in the speed of opsin responses to affect these measurements.) In these mice, the opsin expresses in most or all PV cells [36]. We found that, with transgenic expression, the average paradoxical effect was present (Fig. 6e). Further, the fraction of I cells that showed paradoxical effects was significantly larger in PV-ReaChR transgenic animals than in PV-Chronos animals (Fig. 6g), and was similar to the VGAT-ChR2 data in V1, somatosensory, and motor/premotor cortex (Fig. 6).

Taken together, these data and our analysis support the idea that V1 (and other areas) of the mouse cortex are strongly coupled and operate as an ISN, yet whether an average paradoxical effect is seen depends on the number of inhibitory cells stimulated. The heterogeneous responses we saw in PV-Chronos experiments (about half of I cells paradoxical, half non-paradoxical, Fig. 6d,f,g) also may in part explain why paradoxical inhibitory suppression (and thus ISN operation) has not been more widely reported in optogenetic PV-stimulation experiments using viral expression (also see Discussion).

No transition out of inhibition stabilization is seen during anesthesia, at lower network firing rates

Theoretical studies have pointed out that a network could switch from non-ISN to ISN when external inputs are increased [13] If the cortical network did transition out of the ISN state at lower activity levels, this would be computationally important, as the way inputs sum can vary with whether the network operates as an ISN or not [13; 14]. Our results above (Fig. 2,3) show that superficial layers of mouse V1 operate as an ISN at rest (i.e. without sensory stimulation), and also show the transition from ISN to non-ISN below the level of spontaneous activity, as network activity is decreased by optogenetic stimulation of I neurons. (The transition point is where the slope of I rates vs stimulation intensity switches from being negative to positive.) We next aimed to determine whether reduction of cortical activity levels produced by light anesthesia would also result in a transition.

Under light isoflurane anesthesia, we found that paradoxical inhibitory suppression is maintained (Fig. 7a). At 0.25% isoflurane (a low level, as surgical levels are often 1.0% and above), spontaneous firing rates of excitatory neurons are reduced (Fig. 7c; mean 6.4 spk/s reduced to 3.5 spk/s, $p < 0.02$, Wilcoxon signed-rank test; inhibitory neurons' rates show a negative trend but are not statistically different, 14.5 spk/s awake to 11.6 spk/s anesth., $p < 0.10$, Wilcoxon). Thus, the network changes induced by anesthesia do not cause the network to transition out of the ISN state. At this low level of anesthesia, we did not observe prominent up and down state slow oscillations. In one experiment, we used a higher level of anesthesia (0.5% isoflurane, Fig. 7b), yielding even lower firing rates but still preserving the paradoxical effect. Further confirming the robustness of coupling to changes with anesthesia, the distribution of response slopes is roughly unchanged (Fig. 7c), suggesting the network is far from a transition into a non-ISN state. Anesthesia thus preserves the paradoxical inhibitory response, leaving the network still an ISN.

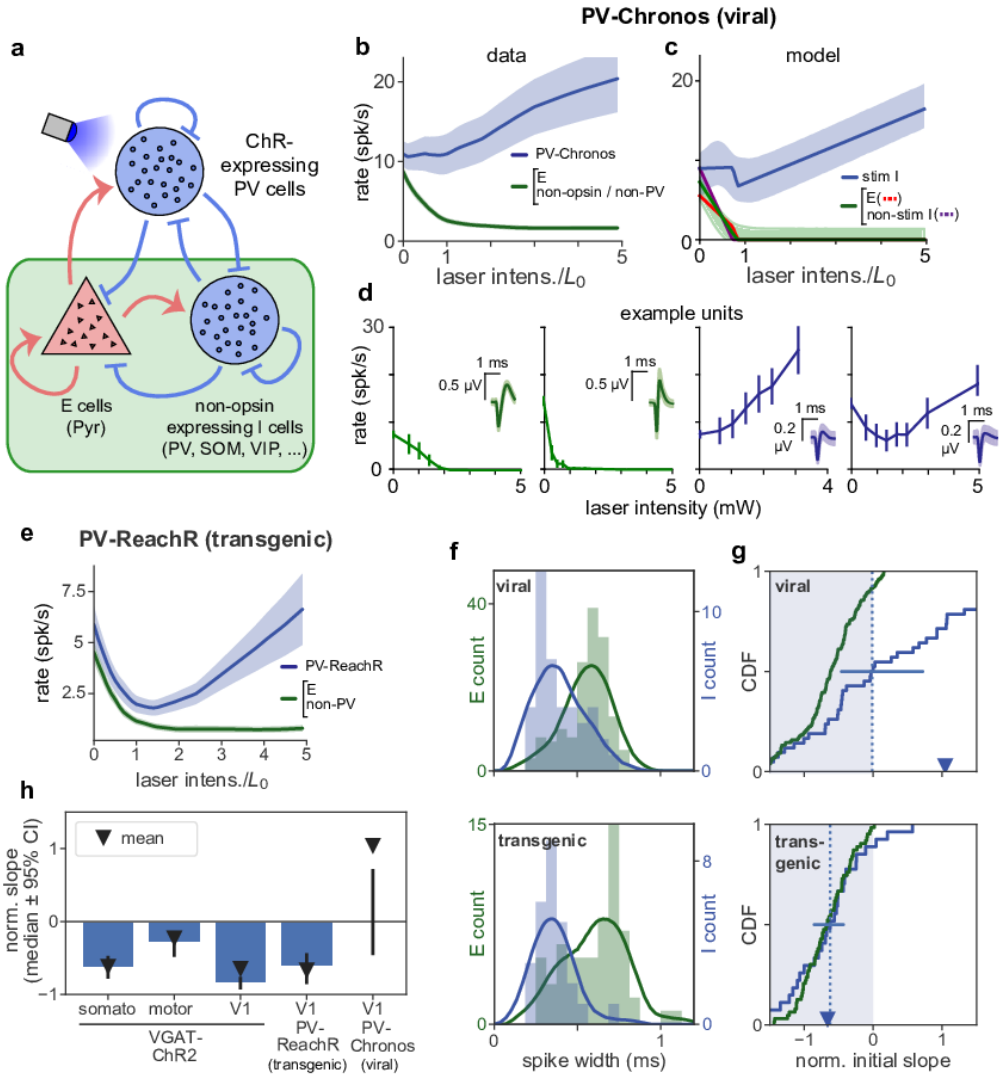

FIG. 6 **Optogenetic stimulation of parvalbumin-positive inhibitory neurons suggests paradoxical effects depend on number of cells stimulated.** (**a**) Schematic of cell populations: opsin-expressing PV cells, non-stimulated I cells (non-PV I cells: SOM, VIP, . . . , and non-opsin PV cells), and pyramidal (E) cells. Data were collected in two experiments: 1, Viral expression of opsin (Chronos) in PV neurons (AAV-Chronos;PV-Cre), and 2, transgenic expression of opsin (ReaChR) in most or all PV neurons (lox-stop-lox-ReachR;PV-Cre). (**b**) Population responses for viral expression of Chronos in PV neurons. Non-PV-Chronos (E, non-Chronos, or non-PV) cells: green (N=152). PV-Chronos (blue) cells (N=42, 21%), identified by responses at high laser power (see text; validation against pharmacology-based classification for V1 data: Supp. Fig. S7). Compared to when all I cells are stimulated (VGAT-ChR2 line, Fig. 2), weaker paradoxical suppression is seen (blue line; mean rate is not significantly suppressed: $p > 0.05$, paired t-test between rate at 0, rate at $1 \cdot L_0$) as initial response slope is near zero. (**c**) A model with a subset of I cells stimulated (60%) can recapitulate the data in (b). Shaded region: ±1 s.d. via data bootstrap as in Fig. 4. With this three-population model, there are now three possible phases of the I response (blue), demarcated by two transition points where the excitatory and non-stimulated inhibitory cells reach their minimum. We use network parameters as inferred in Fig. 4 (also see Methods), and with these parameters the non-stimulated I cells reach their minimum at lower laser intensity than the E cells. (**d**) Example PV-Chronos units. Some narrow spiking units (blue, rightmost two panels) show non-paradoxical initial increases, and some show paradoxical initial suppression. (**e**) Population responses for transgenic expression of ReaChR in PV neurons. Unlike the PV-Chronos data (b), the identified I cells in these experiments show paradoxical suppression (mean firing rate is significantly suppressed: $p < 10^{-6}$, paired t-test). (**f,g**) Distribution of spike widths (f) and initial slopes (g) for experiments using viral PV-Chronos (top) and transgenic PV-ReaChR (bottom). Same conventions as in Fig. 5c-d,g-h, except here colors are as shown above in panel (a). Though I units are classified by response at high laser power, differences in spike width are visible in both datasets. Viral mean and median slopes are zero or positive (tests for negative mean/median not sig.; t-test, Mann-Whitney U; 22/42 (53%) negative I slopes); transgenic mean and medians are negative (tests for negative mean/median: $p < 10^{-7}$, $p < 10^{-7}$; 24/27 (89%) negative I slopes). (**h**) Summary of mean and median I cell initial slopes for these data and data from V1, somatosensory, and motor cortex.

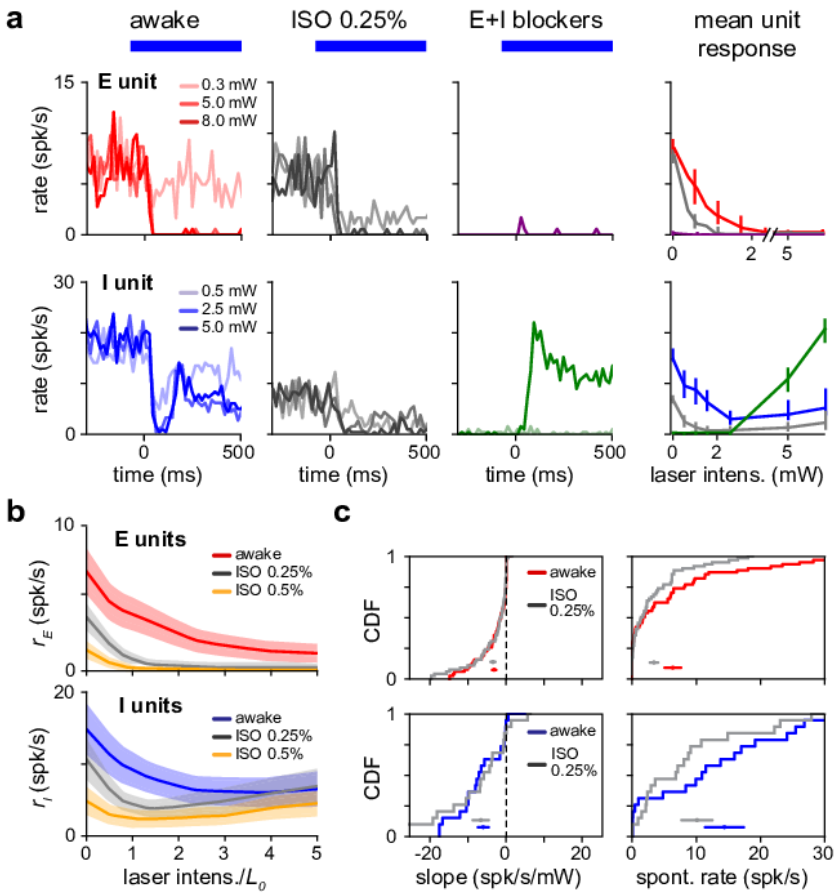

FIG. 7 **Paradoxical response is preserved with lower network activity due to anesthesia** (**a**) Single excitatory (top) and inhibitory (bottom) unit response in the awake state (no anesthesia, no synaptic blockers), under anesthesia with no blockers (isoflurane, 0.25%) and with synaptic blockers (CNQX, APV, bicuculline; Methods) and no anesthesia. (**b**) Population average response with (gray) and without (blue;red) anesthesia. Spontaneous firing rates are reduced for both the E (top) and I (bottom) populations below the awake firing rates, yet inhibitory paradoxical suppression is preserved with anesthesia (lower panel: gray line shows negative initial slope). As laser power increases, weaker increase is seen in awake inhibitory response (blue) compared to Fig. 2I; this could be due to variability across cells/experimental sessions, perhaps due to tissue properties affecting laser effectiveness, but this does not affect paradoxical suppression. (**c**) Distribution of initial slope and spontaneous rate before and after anesthesia. Initial slope is largely preserved while excitatory and inhibitory spontaneous rates are suppressed. Colored horizontal bars within each panel show meanSEM for each distribution. Initial slopes are plotted here non-normalized (units spk/s/mW) to show the slopes are quantitatively similar across firing rate changes; normalized slopes are shown in Supp. Fig. S8, and their means and medians remain negative across anesthesia state.

IV. DISCUSSION

These data show a signature of inhibitory stabilization — the paradoxical suppression of inhibitory cells to optogenetic excitation — in several different areas of mouse cortex (V1, S1 and motor cortex). To resolve whether paradoxical effects were present, we explored systematically a wide range of stimulation strengths, used a pharmacological approach to identify inhibitory interneurons, and compared viral expression to transgenic lines in which all inhibitory interneurons express opsins. We found that for transgenic opsin expression, the large majority of inhibitory interneurons decrease their rates in response to optogenetic stimulation with low intensities, and this result was independent of the way inhibitory interneurons were identified (through pharmacology; through their response to optogenetic stimulation at high intensities; or through spike width). Furthermore, we verified another ISN prediction, that inhibitory neurons transiently

increase their firing rates before they are suppressed by optogenetic stimulation. We find that a simple two population firing rate model operating in the ISN regime captures quantitatively our data, and fitting this model to the data allowed us to infer the coupling strengths between the two populations composing the network. Finally, we showed that inhibitory stabilization is present in several cortical areas, persists under anesthesia (a state in which the activity of both excitatory and inhibitory neurons is reduced), and that stimulation of PV neurons also can produce paradoxical suppression.

Past experiments have examined visual surround suppression to determine if visual cortex is inhibition stabilized. *In vivo* intracellular recordings [15] support the idea that cat V1 operates as an ISN with sensory stimulation, as the firing rate suppression resulting from visual stimulation of receptive field surrounds is associated with a decrease in both excitatory and inhibitory current [15]. However, while such current decreases are consistent with paradoxical suppression of inhibitory cells, they can also be consistent with a disinhibitory mechanism (e.g. surround stimulation increases SOM cell firing rates, which drive PV cell decreases) with no requirement for an ISN. Indeed, in the mouse, surround suppression does not produce a significant decrease in inhibitory current [37], and SOM cell firing rate increases are observed with surround stimulation [38]). A more straightforward way to distinguish ISN and non-ISN explanations is to measure paradoxical inhibitory suppression directly through optogenetic stimulation.

In the last decade, multiple experiments have performed optogenetic stimulation of cortical interneurons, but these studies have yielded mixed results on inhibition stabilization. In layer 2/3 of mouse V1, optogenetic excitation of PV cells with viral expression was shown to generate non-paradoxical modulation of inhibitory neurons [16]. Moore et al. [19] reported that stimulation of PV cells, also using viral transfection, in mouse auditory cortex produced paradoxical effects in some PV cells and not others, and they report evidence for disinhibitory, non-ISN, mechanisms. Also in A1, Kato et al. [18] found that suppressing inhibitory cells with viral ArchT produces paradoxical effects in intracellular currents, but because intracellular currents reflect summed activity from multiple cell types, these data can also be consistent with non-ISN mechanisms. In addition, many experiments have used excitation of inhibitory cells [e.g. [30]] to suppress excitatory activity, without reporting paradoxical inhibitory effects. On the other hand, paradoxical effects for PV and all-inhibitory stimulation have been reported in a recent study [39] in somatosensory and motor cortex, using cell-attached recording. Several features of our data suggest why paradoxical inhibitory suppression has not been previously widely reported. First, it appears a significant fraction of inhibitory cells must be stimulated to produce large paradoxical effects. Stimulating PV cells with viral methods (which can yield subsets of cells with weaker or no expression; see also [17] for a similar explanation for non-paradoxical effects with viral expression), produced in our hands mixed effects on single cells and weaker mean paradoxical effects (Fig. 6). Second, especially in experiments using extracellular recording, it can be difficult to definitively identify inhibitory cells. Some inhibitory cells' waveforms are broad (Fig. 2), and in an ISN, excitatory and inhibitory cells' firing rate changes are often similar (inhibitory neurons may only show increased activity for a few milliseconds before paradoxical suppression begins). Third, for strong enough drive to inhibitory neurons, ISNs transition to a non-ISN state. In this non-ISN state with low excitatory activity, tracking of excitatory fluctuations is not needed to stabilize the network, and some past experiments may have used stimulation large enough for parts of the network to enter into this second phase of activity, hiding the signature of inhibition stabilization. The differences between PV responses using viral and transgenic approaches we found experimentally can be reproduced in a model in which the fraction of stimulated cells is smaller in the viral case, and provide a potential explanation of why prior studies have reported mixed results. Moreover, we observe here for the first time ISN-predicted response dynamics, providing additional, strong evidence for ISN operation.

In an ISN, strong bidirectional E-I synaptic coupling allows inhibitory cells to stabilize an unstable excitatory network, by tracking and responding to changes in local excitation. While paradoxical effects in a single I subclass need not be related to overall network stability (Fig. 6e), PV neurons may well be the principal providers of stabilizing input. PV neurons receive input from many diverse local excitatory cells [40], show responses that can average local excitatory responses [41; 42], and target virtually all nearby E cells with strong peri-somatic synapses [43; 44]. In principle, other cell types of the diverse inhibitory

classes in the cortex [24] could contribute to stabilization, though evidence suggests specific circuit effects (e.g. gain control) may be controlled by other classes of inhibitory cells (VIP, SOM). Some of those predicted circuit effects would require some classes of I cells to vary independently from excitation and thus deviate from a role in stabilization. An interaction between modeling and experimental approaches seems likely to be necessary to determine how multiple inhibitory classes function within an ISN [20].

Most models of cortical networks predict that such networks should be inhibition-stabilized. In particular, the 'balanced network' model predicts that both excitatory and inhibitory inputs to single neurons should be very large, but approximately cancel each other, leading to a balance between excitation and inhibition. This scenario accounts in a parsimonious way for multiple ubiquitous properties of activity in cortex, such as the highly irregular nature of neuronal firing, and the broad distributions of firing rates across neurons. The parameters of our simple model that best fit the data produce large (compared to threshold) excitatory and inhibitory inputs of approximately similar magnitudes, similar to the 'balanced network' scenario. Another influential model of cortical operation, the supralinear stabilized network (SSN) is inhibition-stabilized for strong enough inputs, but is not ISN for sufficiently weak inputs, and these two regimes have been linked to a transition between supralinear and sub-linear modes of summation of inputs. Therefore, the SSN predicts that at sufficiently low activity levels, the network should no longer be an ISN. Here, we present two pieces of evidence that the network in the awake spontaneous state is far from this transition into low-activity non-ISN behavior: first, excitatory rates have to be substantially suppressed before inhibitory responses switch from paradoxical suppression to increasing their firing rate (e.g. Fig.2e), and second, under light anesthesia, we find network activity is lowered but ISN behavior is preserved. Therefore, our data argue that the weak coupling regime present in the SSN is not a regime mouse V1 enters during normal processing during awake behavior.

Overall, the data reported here strongly suggest that inhibition-stabilization is a ubiquitous property of cortical networks. This data is consistent with a network that is in the strong coupling regime, but is not extremely strongly coupled as some balanced network models would predict [5; 7]. It is tempting to speculate that, while strong coupling allows the network to perform non-trivial computations on its inputs, increasing coupling further might not be optimal for several reasons. First, in the strong coupling limit, only linear computations are available to a balanced network (unless synaptic non-linearities are included, [45]), while networks with moderate coupling can combine inputs in a non-linear fashion [13]. Second, maintaining strong connections might be metabolically expensive, so that a moderately coupled network might represent an ideal compromise for cortical computations.

ACKNOWLEDGMENTS

Supported by the NIMH Intramural Research Program and by NIH BRAIN U01 NS108683 (to M.H. and N.B.). We thank K. Miller and J. Reynolds for discussions and the NIMH Instrumentation Core for technical support. This work used the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

V. METHODS

A. Experimental methods

1. Animals.

All procedures were conducted in accordance with the guidelines of the National Institutes of Health. Seven VGAT-ChR2, three PV-Cre, and two PV-Cre;ReaChR mice were used (JAX stock #s 014548, 008069, and 024846; 2 females and 10 males; singly housed on a reverse light/dark cycle).

2. Cranial window implants.

Mice were implanted with a titanium headpost and a transparent window (optical glass, 0.8mm thickness, 3 or 5 mm diameter) in the left cerebral hemisphere. The window provided access to the primary visual cortex (V1) and auditory, somatosensory, or motor cortices (Aud, S1/S2, and M1/M2), for imaging and silicon neural probe recordings with optogenetics. Mice were given dexamethasone (3.2 mg/kg, i.p.), 2 hours before surgery. Animals were anesthetized during surgery with isoflurane (1.0%-4% in 100% O2). Using aseptic technique, a headpost was affixed using C&B Metabond (Parkell), and a 3 or 5 mm diameter craniotomy made.

3. Hemodynamic intrinsic imaging.

To determine the location of V1, we delivered small visual stimuli to animals at different retinotopic positions and measured changes in the absorption of 532 nm light resulting from cortical hemodynamic responses [46]. We evenly illuminated the brain with a 532 nm fiber coupled LED (M530F2, Thorlabs) passed through a 530 nm-center bandpass filter (M530F2, Thorlabs). Images were collected on a stereo microscope (Discovery V8, Zeiss) through a green long-pass emission filter (500 nm cutoff) using a 1x objective (PlanApoS 1.0x, Zeiss) onto a Retiga R3 camera (Q Imaging, captured at 2 Hz with 4x4 binning). For retinotopic mapping we presented upward-drifting square wave gratings (2 Hz, 0.1 cycles/degree) masked with a circular window (10°diameter) for 5 s with 10 s of mean luminance preceding each trial. Stimuli were presented in random order at four positions in the right monocular field of view. The response to a stimulus was calculated as the fractional change in intensity between the average of the 10 frames immediately preceding the stimulus (as baseline) and a 6-10 frame window 1 s after stimulus onset (as response) to match the timecourse of the hemodynamic response [47; 48].

4. Viral injections.

We used viral injections to express Chronos [32] in PV-Cre [33] animals. Mice were anesthetized (isoflurane 1-1.5%) and the cranial window implant removed. We used a stereotaxic injection system (QSI, Stoelting Inc.) to deliver 500 nL of a 10:1 mixture of AAV1-hSyn-FLEX-Chronos-GFP (UNC Vector Core Stock) and 100 μM sulforhodamine (SR101, Invitrogen, for visualization) at a depth of 200-400 m. We made 3 to 4 injections spaced 0.5-1.0 mm apart to cover the visual cortex. After the injections a new cranial window was affixed. Viral expression was monitored over the course of days by imaging GFP fluorescence, and allowed to develop for greater than 4 weeks before electrophysiological recordings.

5. Electrophysiological recording with optogenetic activation of inhibitory neurons.

For recording experiments, we first affixed a 3D printed ring to the cranial window to retain fluid. With the ring in place, we removed the cranial window and flushed the craniotomy site with sterile normal saline to remove debris. Kwik-Sil silicone adhesive (World Precision Instruments) was used to seal the craniotomy between recording days. Using a stereomicroscope on an articulating arm we positioned the end of an optical fiber (600 m diameter, Doric Lenses) fitted with a light-tight coupler to an optical cannula (400 m diameter, Thorlabs CFMLC14L02) over our target cortical area, with a slight 10-30° angle from vertical to provide space for the electrodes. We used fiber coupled LED light sources (M470F3 (Chronos and ChR2) or M625F2 (ReaChR), Thorlabs) to deliver illumination with peak at 470 nm or 625 nm to the brain. We calibrated total intensity at the entrance of the cannula using a power meter and photodiode (meter model 1918-R, photodiode model 918D-SL-OD3R, Newport Corp.). The cannula distance from the dura was adjusted to provide a light spot with a full width at half maximum intensity of 0.8-1.2 mm (as measured

with a small digital macro documentation microscope, Opti-TekScope); thus, our reported spot irradiance in mW/mm^2 is within a factor of two of the reported power (mW). To compare effects across days, we normalized laser intensity based on inhibitory response. We targeted a multisite silicon probe electrode (NeuroNexus; 32-site model 4x8-100-200-177; 4 shank, 8 sites/shank; sites were electrochemically coated with PEDOT:PSS [poly(3,4-ethylenedioxythiophene): poly(styrenesulfonate)] [49]) to the center of the LED spot using a micromanipulator (MPC-200, Sutter Instruments). With the fiber optic cannula and electrode in place, we removed the saline buffer using a sterile absorbent triangle (Electron Microscopy Sciences, Inc.) and allowed the dura to dry for 5 minutes. After insertion, we waited 30-60 minutes without moving the probes to reduce slow drift and provide more stable recordings. We isolated single and multiunit threshold crossings (3 times RMS noise) by amplifying the site signals filtered between 750 Hz and 7.5 Khz (Cerebus, Blackrock microsystems). During recordings animals were awake and passively viewing a gray screen. To keep animals awake and alert, animals were water scheduled [50], and a 1ul water reward was randomly provided on 5% of the stimulus trials. Optogenetic stimuli were square pulses with 2 ms linear ramps at start and end to reduce recording artifacts. Pulses were on for 600 ms and off for 1000 ms at a range of power levels (0.3-10 mW), presented in random order, with 100 repetitions per power level.

6. Pharmacological blocking of excitatory and inhibitory synapses.

To classify cells using pharmacological weakening of cortical synapses, we divided the experiment into three phases, each 30-45 minutes in duration: first no blockers, then excitatory blockers, then excitatory plus inhibitory blockers. For each section, we delivered the same optogenetic stimulation protocol (above). To apply the pharmacological agents, we made a hole near the recording electrodes in the agarose on top of the brain, removed the normal saline covering the agarose and hole, and replaced it with a solution containing the agent in saline. After 15 minutes, the initial application of blocking solution was removed via aspiration and refreshed. We waited a total of 20 mins for the pharmacological agent(s) to take effect before recording. Excitatory synaptic blockers (2 mM CNQX, 6 mM APV) affected AMPA, kainate, and NMDA synapses. Inhibitory blockers (1 mM bicuculline) affected GABA-A synapses. Source: Sigma-Aldrich (#C239, #A8054, and #14343).

7. Histology and electrophysiology probe tracking

To determine the location and depth of recording electrodes we coated the shanks with 1,1'-Dioctadecyl-3,3,3',3'-tetramethylindocarbocyanine perchlorate (DiI) (Thermo Fisher #D282, 50 mg/mL solution in ethanol), prior to insertion into the brain. We dipped the electrode tips 5 times into the DiI solution and allowed the coating to dry for 30 seconds between immersions. Fluorescent electrode tracks were then visualized in coronal sections of fixed brain tissues. Mice were anesthetized with isoflurane and injected intraperitoneally with pentobarbital sodium (150 mg/kg). They were then perfused transcardially with 25mL cold (4C) PBS followed by 25mL cold 4% paraformaldehyde. Brains were extracted and fixed in 4% paraformaldehyde for 6-12 hours and then cryoprotected in a 30% (w/v) sucrose solution in PBS until they sank. Brain tissues were sectioned at 50 m on a freezing microtome (Leica), mounted on glass slides, and coverslipped with mounting media containing DAPI (Fluoromount-G with DAPI, Electron Microscopy Sciences). Slides were imaged using an Olympus slide-scanner (Olympus BX61VS, Japan).

8. Anesthesia during electrophysiology.

To test the effects of lowering overall activity on the stability of the cortical network we fitted animals with an isoflurane inhalation mask system (model V-1, VetEquip Inc.) and provided anesthesia (0.25-0.5% isoflurane in 100% O2) to put the animal into a lightly anesthetized but awake state during recordings. At

the lower 0.25% isoflurane concentration, recordings in V1 displayed no synchronized oscillatory Up/Down state activities and the animals eyes were open throughout. The anesthesia experiments were divided into four sections: awake animal, anesthesia with no blockers, anesthesia with excitatory blockers, anesthesia with inhibitory blockers. Anesthesia was delivered for 20 minutes before starting recording.

9. Spike sorting

Spike waveforms were sorted after the experiment using OfflineSorter (Plexon, Inc.). Single units were identified as waveform clusters that showed clear and stable separation from noise and other clusters, unimodal width distributions, and inter-spike interval histograms consistent with cortical neuron absolute and relative refractory periods. Multiunits were clusters that were distinct from noise but did not meet one or more of those criteria, and thus these multiunits likely group together a small number of single neurons. Signal-to-noise ratio (SNR) [51; 52] of single unit waveforms (median ±std): in visual cortex, 3.6±1.7 (N=166), see Supplementary Fig. 4; in other datasets: motor, 3.7 ±1.2 (N=103); somatosensory, 3.7 ±1.7; in V1 PV data (Fig. 6) 2.91 ±1.3. Only single-units were analyzed; multiunits were discarded. We also repeated our single-unit analyses using both more- and less- stringent criteria to qualify a unit as a single unit, and found no qualitative differences in the results (Supplementary Fig. 4). Supporting the idea that single units did not group together multiple units, we found no significant correlation between SNR and baseline firing rate in any of the four datasets (all $p > 0.2$ by linear regression).

B. Mathematical methods

1. Model equations

To quantitatively analyze the network response, we used a two population rate model [53], in which the average firing rates of excitatory (E) and inhibitory (I) populations (r_E and r_I, respectively) evolve according to

$$\begin{cases} \tau_E \frac{dr_E}{dt} &= -r_E + \phi_E \left(W_{EE}\ r_E - W_{EI}\ r_I + I_{EX} \right) \\ \tau_I \frac{dr_I}{dt} &= -r_I + \phi_I \left(W_{IE}\ r_E - W_{II}\ r_I + I_{IX} + \lambda L \right) \end{cases} \tag{1}$$

where ϕ_A, I_{AX} and τ_X are the static transfer function (f-I curve), external input and time constant of population A (=E,I), respectively, while W_{AB} is the strength of connections from population B to A. The optogenetic stimulation is described by the parameters λ and L which represent the efficacy and the intensity of the stimulation light. Results shown in the main text (Fig. 4) have been obtained with a rectified-linear transfer function

$$\phi_A(x) = a_A \left[x - x_{0A} \right]_+ , \tag{2}$$

which is zero for input x smaller than the threshold x_{0A}, and increases linearly, with a gain a_A, otherwise. This transfer function is the simplest one that can describe the data, which shows an approximately piecewise linear dependence of firing rates on stimulation intensity (see Fig. 4). We also fit the data using a different transfer function

$$\phi_A(x) = b_A \log \left\{ 1 + \exp \left[\frac{a_A}{b_A} \left(x - x_{0A} \right) \right] \right\} , \tag{3}$$

that smoothes the threshold non-linearity of the rectified linear function, using an additional parameter b_A that controls the width of the exponential region around threshold. This function reduces to the rectified-linear transfer function when $b_A \to 0$. We find that the nonlinear transfer function provides a minor

improvement in describing the data and does not significantly affect the values of the inferred parameters (Supp. Fig. .S11).

In the experiments, recordings of the response are done in three separate phases: (1) A 'normal' phase in which recurrent interactions are intact; (2) A phase with E synaptic blockers (see Experimental methods); and (3) A phase with both E and I synaptic blockers (see Experimental methods). The addition of blockers in the model is described with two parameters, $\epsilon_{E,I} \in [0,1]$, representing the decrease in strength of excitatory and inhibitory synapses. After the addition of excitatory blockers, connectivity is modified as

$$W_{EE},\, I_{EX},\, W_{IE},\, I_{IX} \to \epsilon_E\, W_{EE},\, \epsilon_E\, I_{EX},\, \epsilon_E\, W_{IE},\, \epsilon_E\, I_{IX}\,, \tag{4}$$

and with inhibitory blockers, connectivity is modified as

$$W_{EI},\, W_{II} \to \epsilon_I\, W_{EI},\, \epsilon_I\, W_{II} \tag{5}$$

Combining Eqs.(1), (4) and (5) with the transfer function of Eq. (2), the model uses 15 parameters (2 time constants, 2 thresholds, 2 gains, 4 connectivity strengths, 2 external inputs, 1 stimulation efficacy, 2 blocker efficacies) to describe simultaneously the three phases of the experiment.

To compare the model response with the data, we find the equilibrium solution of Eq. (1), shown in Eq. (6). Since τ_E and τ_I do not affect the equilibrium solution, the number of relevant independent variables in the model reduces to 13. Moreover, the values of a_E and a_I can be reabsorbed in the definitions of W, I and λ. It follows that all the parameters can be inferred up to a proportionality constant and, without loss of generality, we can fix $a_E = a_I = 1$. This reduces the number of independent parameters to 11.

Note that this number of parameters is smaller than the number of parameters needed for piecewise linear fits of the data (average firing rates vs light intensity for both E and I neurons). Such piecewise linear fits of the data need 15 parameters: 3 phases times 5 parameters per phase (3 for I neurons - 2 per linear region, minus one for the continuity constraint; and 2 for E neurons, for the single linear region at low intensities). Thus, the model does not have enough parameters to fit successfully the data, unless the model structure accurately describes the cortical responses. Indeed, Fig. 4 shows that the model gives a good description of the data. The additional 4 degrees of freedom present in the data lead to four parameter free model predictions described in the main text.

The fixed point solutions of Eq. (1) (without blockers) are

$$\begin{cases} r_E = \frac{(W_{II}+1)(I_{EX}-x_{0,E})-W_{EI}(I_{IX}+\lambda L-x_{0I})}{W_{EI}W_{IE}-(W_{II}+1)(W_{EE}-1)} \\ r_I = \frac{W_{IE}(I_{EX}-x_{0,E})-(W_{EE}-1)(I_{IX}+\lambda L-x_{0I})}{W_{EI}W_{IE}-(W_{II}+1)(W_{EE}-1)} \end{cases} \quad \text{and} \quad \begin{cases} r_E = 0 \\ r_I = \frac{I_{IX}+\lambda L-x_{0,I}}{W_{II}+1}\,. \end{cases} \tag{6}$$

For every value of the laser intensity, the excitatory and inhibitory nullclines are defined as the functions $r_I = r_I(r_E)$ which solve the first and the second line of Eq. (6), respectively. Every intersection between the two nullclines is a solution of Eq. (6) and gives a stationary state of the network dynamics.

Eq. (6) has two dimensions that can be inferred only up to a multiplicative constant:

$$\begin{cases} W_{EE}-1,\, W_{EI},\, I_{EX},\, x_{0,E} \to \gamma_E(W_{EE}-1),\, \gamma_E W_{EI},\, \gamma_E I_{EX},\, \gamma_E x_{0,E}\,, \\ W_{IE},\, W_{II}+1,\, I_{IX},\, x_{0,I},\, \lambda \to \gamma_I W_{IE},\, \gamma_I(W_{II}+1),\, \gamma_I I_{IX},\, \gamma_I x_{0,I},\, \gamma_I \lambda\,. \end{cases} \tag{7}$$

Because of these invariances, using data from a single experimental phase, only ratios between parameters can be inferred. This is no longer true when the three phases are considered; in this case all the 11 relevant model parameters can be found.

2. Parameter inference

Given the dataset of excitatory and inhibitory responses, the best set of parameters describing the data is found via global optimization as follows:

1. Select random initial parameters in the interval [0,10].

2. Find the optimal set of parameters through a least-squares optimization (python function "curve_fit", variables are constrained to be positive and, for $\epsilon_{E,I}$, in the interval [0,1]) of the difference between observed rates and model predictions given by Eq. (6).

3. Repeate the procedure 10^4 times, and select the solution with minimal error as the optimal parameter set describing the data.

The optimization procedure is applied simultaneously on data from all phases and recording sessions in V1. For the first phase (awake), data from all recording sessions from VGAT-ChR2 animals in V1 (9 days, 4 animals) were pooled together for fitting. For the second and third phases (E blockers, E+I blockers), in order to describe the variability from day to day of the efficacy of the blockers, we use a different set of $\epsilon_{E,I}$ for each recording session (6 days, 2 animals). For these two phases, 3 out of the 9 recording sessions are excluded from the analysis because blockers were added after anesthetizing the animals. The optimal parameters found by this approach are:

$$\begin{gathered} W_{EE} = 2.56,\, W_{EI} = 1.77,\, I_{EX} = 8.51\, spk/s,\, x_{0E} = 1.19\, spk/s, \\ W_{IE} = 8.54,\, W_{II} = 7.11,\, I_{IX} = 34.16\, spk/s,\, x_{0I} = 8.65\, spk/s,\, \lambda = 6.3\, spk/s \end{gathered} \tag{8}$$

These parameters are used in Eq. (6) to generate the best model description of the data in the first phase of the experiment (Fig. 4A, first column, black line). For the second and third phases, predictions also depend on the efficacy of synaptic blockers, which are different in the different recording sessions (see Fig. S10). Data and model predictions for all sessions are shown in Fig. S10; in the main text, we showed averages of the two computed across sessions for each phase (Fig. 4A, second and third columns, black line).

We use a bootstrap approach, combined with the global optimization described above, to estimate the precision with which model parameters can be inferred from the data. The optimal (minimum-error) estimates obtained from 10^4 random resamplings (random with replacement) are shown in Supp. Fig. S10. Despite the non-uniqueness of the solutions, the bootstrap shows (dashed lines in Fig. 4: 1 s.d. via boostrap) that model rates are clustered around the data in all three phases of the experiment.

3. Stability of the solution

To analyze the stability of the system, we compute the eigenvalues of the dynamical matrix defined by Eq.(1). For the linear transfer function we find

$$\lambda_{\pm} = \frac{a_E W_{EE}}{\tau_E}\left[\left(\frac{a_E W_{EE}-1}{a_E W_{EI}} - r\frac{a_I W_{II}+1}{a_I W_{IE}}\right) \pm \sqrt{\left(\frac{a_E W_{EE}-1}{a_E W_{EI}} + r\frac{a_I W_{II}+1}{a_I W_{IE}}\right)^2 - 4r}\right] \tag{9}$$

with $r = \frac{\tau_E a_I W_{IE}}{\tau_I a_E W_{EI}}$ This ratio cannot be inferred from the data since $\tau_{E,I}$ do not appear in the static solution. We thus determined τ_E and τ_I (shown in Fig. 4e) from the response dynamics (Fig. 4d).

4. With multiple inhibitory subclasses, ISN operation need not imply paradoxical suppression

To understand how the network response to PV stimulations depends on its structure, here we analyze a three population model (Fig. 6a) with one excitatory and two inhibitory populations, referred hereafter as E (pyramidal cells), P (ChRonos-expressing PV cells), and I (non-expressing PV , SOM, VIP...). The network dynamics is described by

$$\begin{cases} \tau_E \frac{dr_E}{dt} &= -r_E + \phi_E\left(W_{EE} r_E - W_{EI} r_I - W_{EP} r_P + I_{EX}\right)\,, \\ \tau_I \frac{dr_I}{dt} &= -r_I + \phi_I\left(W_{IE} r_E - W_{II} r_I - W_{IP} r_P + I_{IX}\right)\,, \\ \tau_P \frac{dr_P}{dt} &= -r_P + \phi_P\left(W_{PE} r_E - W_{PI} r_I - W_{PP} r_P + I_{PX} + \lambda L\right)\,. \end{cases} \tag{10}$$

The model predictions in Fig. 6 have been obtained from Eq. 10 using

$$\begin{aligned} W_{EE} &= W_{EE}^{VGAT}\,, \quad W_{EI} = (1-x)W_{II}^{VGAT}\,, \quad W_{EP} = xW_{II}^{VGAT}\,, \quad I_{EX} = I_{EX}^{VGAT}\,, \\ W_{IE} &= W_{IE}^{VGAT}\,, \quad W_{II} = (1-x)W_{II}^{VGAT}\,, \quad W_{IP} = xW_{II}^{VGAT}\,, \quad I_{IX} = I_{EX}^{VGAT}\,, \\ W_{PE} &= W_{IE}^{VGAT}\,, \quad W_{PI} = (1-x)W_{II}^{VGAT}\,, \quad W_{PP} = xW_{II}^{VGAT}\,, \quad I_{PX} = I_{EX}^{VGAT}\,, \end{aligned} \tag{11}$$

where $x \in (0,1)$ represents the fraction of ChRonos-expressing PV cells, W_{AB}^{VGAT} and I_A^{VGAT} represent the parameters of Eq. (8), obtained using the V1 data from the VGAT-ChR2 mouse line.

Using the rectified linear transfer function of Eq. (2), and assuming $a_A = 1$ to simplify expressions, the equilibrium response to stimulation of the P population is given by

$$r_P(L) - r_P(0) = \frac{(1-W_{EE})(W_{II}+1) + W_{EI}W_{IE}}{N}\lambda L \tag{12}$$

with

$$\begin{aligned} N = (1-W_{EE})\left[(W_{II}+1)(W_{PP}+1) - W_{IP}W_{PI}\right] \\ - W_{EI}\left[W_{IP}W_{PE} - W_{IE}(W_{PP}+1)\right] + W_{EP}\left[(W_{II}+1)W_{PE} - W_{IE}W_{PI}\right] \end{aligned}$$

Eq. (12) shows that, both for $W_{EE} < 1$ and $W_{EE} > 1$, depending on the network connectivity, the ratio on the r.h.s. can be either positive or negative. It follows that, in a network with two inhibitory populations, an unstable excitatory subnetwork does not imply paradoxical suppression when one of the populations is stimulated. This can be seen more easily in a simplified model in which connectivity depends only on the identity of the presynaptic unit. In particular, using $W_{EE} = W_{IE} = W_{PE} = W$, $W_{EI} = W_{II} = W_{PI} = k_I W$, and $W_{EP} = W_{IP} = W_{PP} = k_P W$, Eq. (12) becomes

$$r_P(L) - r_P(0) = \frac{1 + (k_I - 1)W}{1 + (k_P + k_I - 1)W} \tag{13}$$

In the large W limit, in which the excitatory population is unstable without inhibition, the network shows the paradoxical effect only if $k_I < 1$ and $k_P > 1 - k_I$. (Note that, for the network to be stable in this condition, the matrix of the coefficients of Eq. (10), linearized around the fixed point, must have eigenvalues with a negative real part, a requirement which also involves the magnitude of τ_E, τ_I, and τ_P.)

SUPPLEMENTARY FIGURES

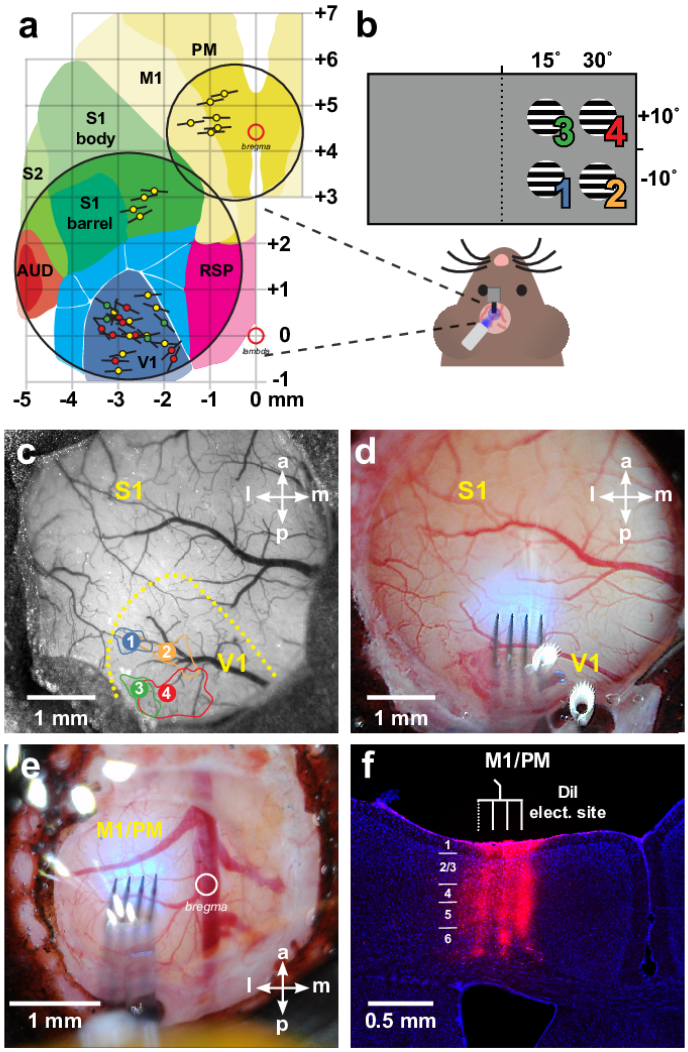

FIG. S1 **Locations of recording sites.** (**A**) Schematic of all physiology probe placements in the visual, somatosensory, and motor cortices (black lines give approximate extent of shanks on the 4-shank probes; colored circles signify experiment types: VGAT-ChR2: yellow; viral PV-Cre;AAV-FLEX-ChRonos: red; transgenic PV-Cre;ReaChR: green). V1: primary visual (Allen notation: VISp); RSP: retrosplenial; AUD: auditory areas; S1: primary somatosensory (SSp); S2: secondary somatosensory (SSs); M1: primary motor (MOp), PM: premotor (MOs). Coordinates are in millimeters relative to lambda and bregma (open red circles). Area boundaries drawn from Allen Mouse Brain Connectivity Atlas [54]. (**B**) Schematic of hemodynamic intrinsic-signal optical imaging (IOI) for retinotopic mapping of V1 (VISp). Hemodynamic responses were measured to upward-drifting square wave gratings (2 Hz, 0.1 cycles/°) masked with a circular boundary (10°diameter) at four locations in the right monocular field of view. (**C**) Picture of a large-diameter (5 mm) cranial window allowing access to the visual and somatosensory cortices. Yellow outline: estimated edge of V1 via retinotopic responses; colored numbers: centroid of IOI response corresponding to the stimulus positions in panel B; solid lines of the same colors: areas yielding responses within 50% of peak response. (**D**) Example physiology probe placement and blue LED overlap for optogenetic stimulation in V1. (**E**) Picture of a smaller-diameter (3 mm) window implant allowing access to the motor cortex. (**F**) Histological section showing location of electrode tracks in M1/PM. AP bregma+0. Blue: DAPI. Red: DiI deposited on electrodes before insertion.

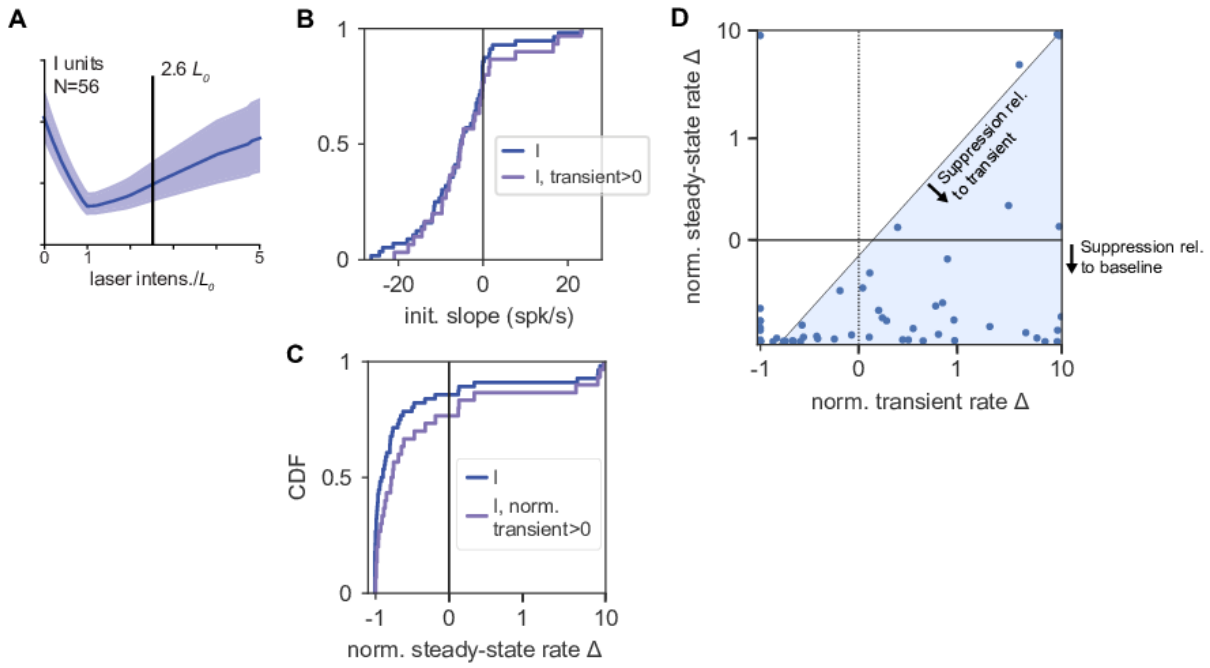

FIG. S2 **Relationship between inhibitory initial transient and paradoxical suppression.** (**A**) Schematic of laser intensity used here and in Fig. 3. This intensity ($2.6 \cdot L_0$) is above the point of maximal paradoxical steady-state suppression ($1.0 \cdot L_0$), but shows the initial transient (Fig. 3AB) clearly. (**B**) Paradoxical suppression (initial slopes < 0) is seen even when selecting only the inhibitory units that show a positive initial transient ("transient > 0"), those units with a an increase in spike rate over an interval [1,12] ms after light onset. ("I": all recorded I units, N=56). (**C**) Paradoxical suppression is preserved for transient-only units: same as (B), plotting the distribution of steady-state responses. Normalized responses < 0, to the left of the vertical line at 0, are units that are suppressed (i.e. that show paradoxical suppression) in steady state, after the end of the transient, for this laser intensity. (**D**) The majority of recorded inhibitory units show steady-state paradoxical suppression relative to both baseline rate and to transient rate. y-axis: steady-state rate change (normalized to baseline rate; i.e. $\frac{rate-baseline}{baseline}$, a value < 0, labeled at right, signifies paradoxical suppression). x-axis: transient rate change (normalized to baseline rate; I units with an initial positive transient have values > 0). The majority of the units are in the lower right quadrant (transient > 0, steady-state < 0). A few units show a positive transient and elevated steady-state rate (upper right quadrant); these are likely I units showing non-paradoxical steady-state rate increases. Blue shaded region shows units whose normalized steady-state rate is lower than their normalized transient rate. The fact that most I units fall into this region shows that steady-state suppression does not arise from mixing E and I waveforms, but instead that we accurately record I units — which often show initial transients followed by suppression as predicted by ISN models.

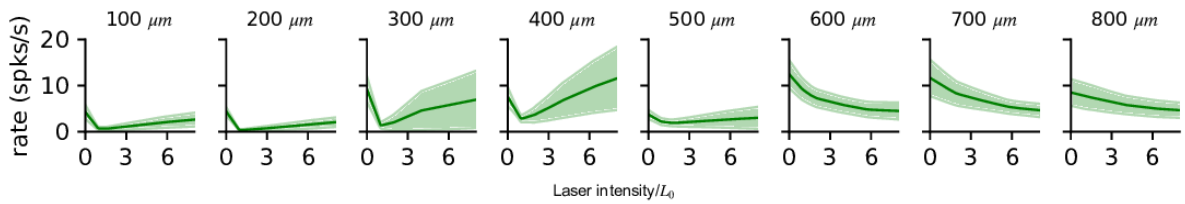

FIG. S3 **Unit response to light as a function of depth.** Panels show the average unit firing rate as a function of laser intensity at sites of different depths. The recording probes have 8 sites per shank (four shanks, 32 total sites) with $100\mu m$ spacing between sites. Each plot shows avearge of all inhibitory units recorded from V1 (data as in Fig. 2) without blockers present. The shallowest four sites (left) show strong responses to laser stimulation, such that at maximum laser power firing rates increase as excitatory rates are suppressed and inhibitory cells respond strongly to stimulation. The deeper sites show suppression, but do not show a transition to increased firing rates, even at maximum laser intensity. This is consistent with the fact that the laser intensity experienced by neurons decays with depth due to scattering and absorption. Because of the decrease in laser effectiveness at deeper sites, in the analysis described in the main text we used only units found in the four more superficial recording sites, i.e. approximately 400μm from the brain surface. In our recordings, we aimed to advance the electrodes so that the top sites were shallow enough to record no neuronal signals, or just record the most shallow units observed during advancing. For this reason, site depth is an approximate measure of neuronal depth, but this procedure likely resulted in the V1 data analyzed in the paper containing a majority of units from layer 2/3 or 4. Note that inhibitory cells from 300-400μm depth show the largest increases at high laser power, but units from 100-400μm depth all show initial paradoxical suppression, as expected in an ISN.

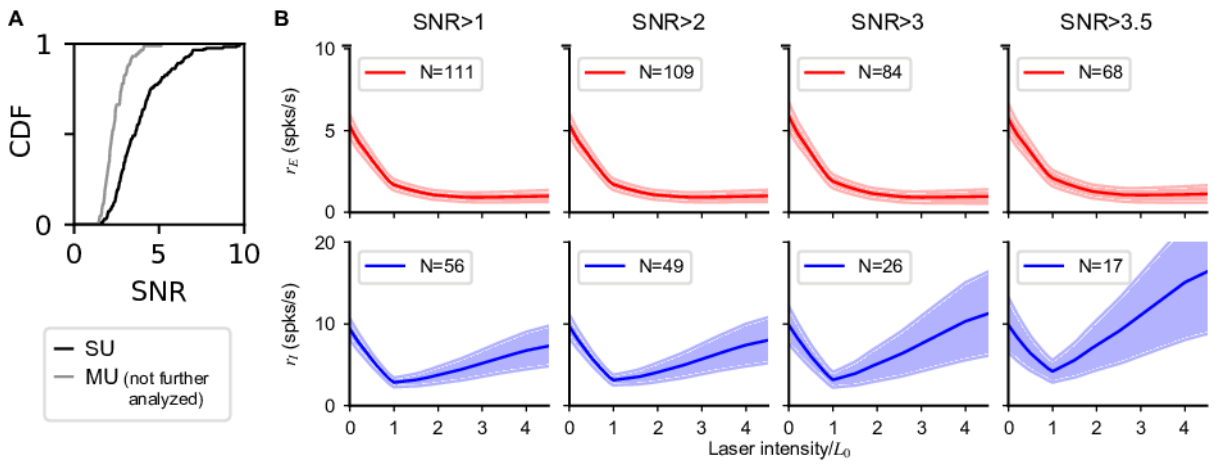

FIG. S4 **Experimental results are robust to quality of unit isolation.** (**A**) Signal-to-noise ratio [55] of single (black) and multi (gray) units. Single units are classified manually (blind to initial slope) based on waveform principal component separation from noise and other units, change in waveform separation over time, and inter-spike-interval distribution. As expected, classified single units have a larger SNR (single-unit E median SNR: 3.97, I median 2.83) than multi-units. Units classified as multi-units were excluded from all further analysis and are shown here only for comparison purposes. (**B**) Population response computed in a subset of SU with SNR higher than the threshold indicated in title. As the constraint becomes more stringent (left to right), the number units used decreases, and the paradoxical effect is preserved (for all panels, initial slope median is significantly negative: all $p < 10^{-5}$ except bottom right: $p < 0.003$; one-sided Mann-Whitney U test). Additional evidence that the paradoxical effect is not dependent on our unit isolation comes from the fact that the PV-Chronos recordings were conducted with identical experimental procedures, and show no average paradoxical effect (see Results and Fig. 6).

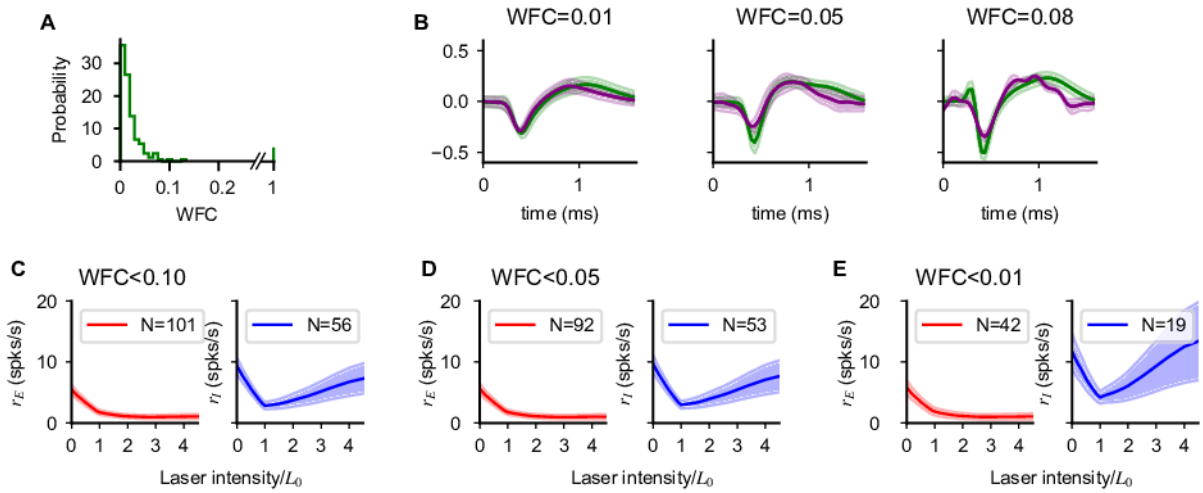

FIG. S5 **Waveform drift has no qualitative effect on population response.** (**A**) Distribution of relative waveform change index (WFC). WFC is defined as, for every unit, the quantity

$$WFC = \frac{1}{norm} \sum_i \frac{(w_0(t_i) - w_1(t_i))^2}{\sigma_0^2(t_i) + \sigma_1^2(t_i)}, \quad norm = \sum_i \frac{(w_0(t_i) + w_1(t_i))^2}{\sigma_0^2(t_i) + \sigma_1^2(t_i)}, \tag{14}$$

where the waveform and its variance are given by $w_j(t)$ and $\sigma_j^2(t)$, and $j = 0, 1$ indexing the blocker, no-blocker cases. t_i ranges over time samples of the recorded waveforms (45 samples/waveform, 1.5 ms, 30kHz acquisition rate). Eq. (14) measures the relative change in waveform of a unit before and after the addition of blockers. Note that a small number of units did not generate any spike after the addition of blockers (WFC=1, N=9); these could either be excitatory units that become silent, or units that drifted substantially during the recording. (**B**) Example single unit before (green) and after (purple) the addition of synaptic blockers. (**C**-**E**) Population response of excitatory and inhibitory cells computed using single unit selected with a threshold on the maximum waveform change. As the constraint becomes more stringent (left to right), the number units used decreases but the structure of the response is not modified and the paradoxical effect is preserved. Our waveform-sorting procedure uses a contour, fixed over time, in the 2d axis of the first two principal components (PCs) to classify units. Therefore, larger WFC values signify shifts in waveform (e.g. WFC=0.08 example, B, right panel), despite the fact the limits of maximum waveform shift are constrained to a range of possible PC values specified by the contour. Including units that show relatively larger waveform shifts (compare D and E) does not modify our effects.

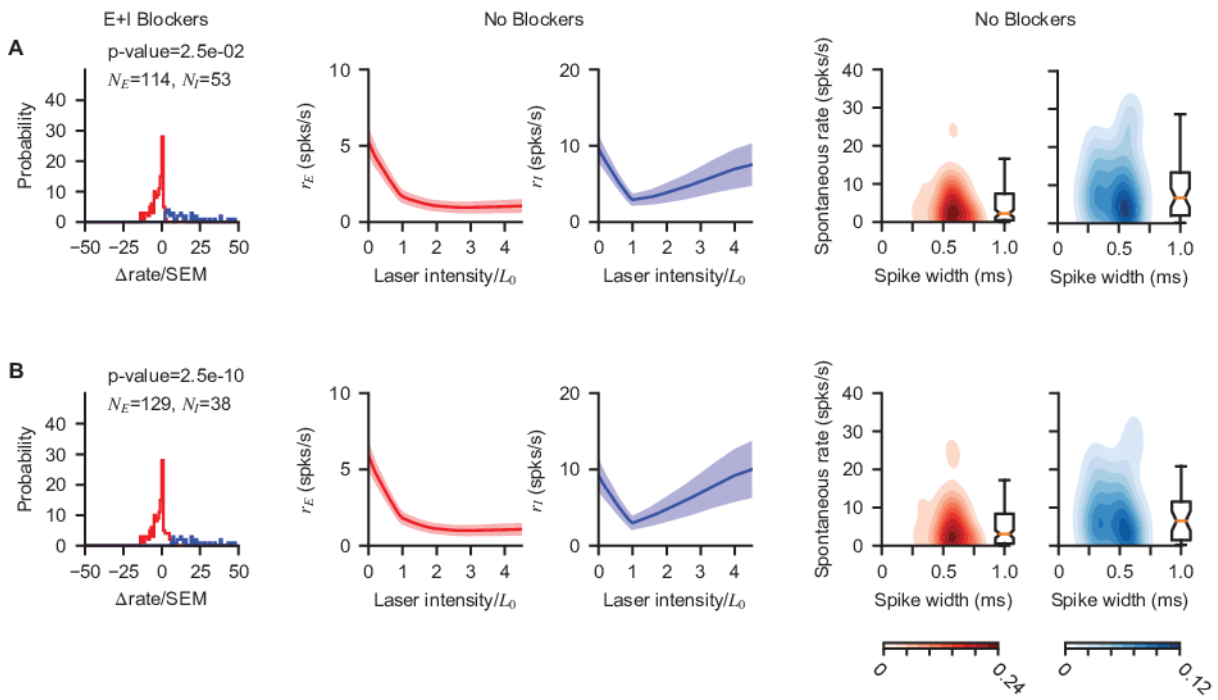

FIG. S6 **Paradoxical V1 inhibitory response does not depend on statistical E/I classification threshold.** The pharmacological classification method (based on firing rate increase to optogenetic stimulation when E+I blockers had been applied; Fig. 2) is robust to choice of classification threshold applied to that firing rate change. (In addition, Supp Fig. S4 shows paradoxical effect is preserved as waveform isolation via SNR is varied; Supp. Fig. S5 shows potential waveform drift over time does not affect results.) **A**: Left column, distribution of response slopes shown with a permissive classification threshold ($p < 0.05$; to make the test two-tailed we used $p < 2.5 \cdot 10^{-2}$; Welch's t-test allowing unequal variances). Middle two columns: resulting average response to stimulation. Right two columns: distribution of spontaneous rates and waveform widths. **B**: Same, but for a more restrictive classification threshold ($p < 10^{-9}$). Decreasing the threshold reduces the number of units classified as inhibitory, (13 units re-classified as excitatory once the more stringent threshold is applied), but does not modify the conclusions of the main text; in particular paradoxical inhibitory modulation is preserved.

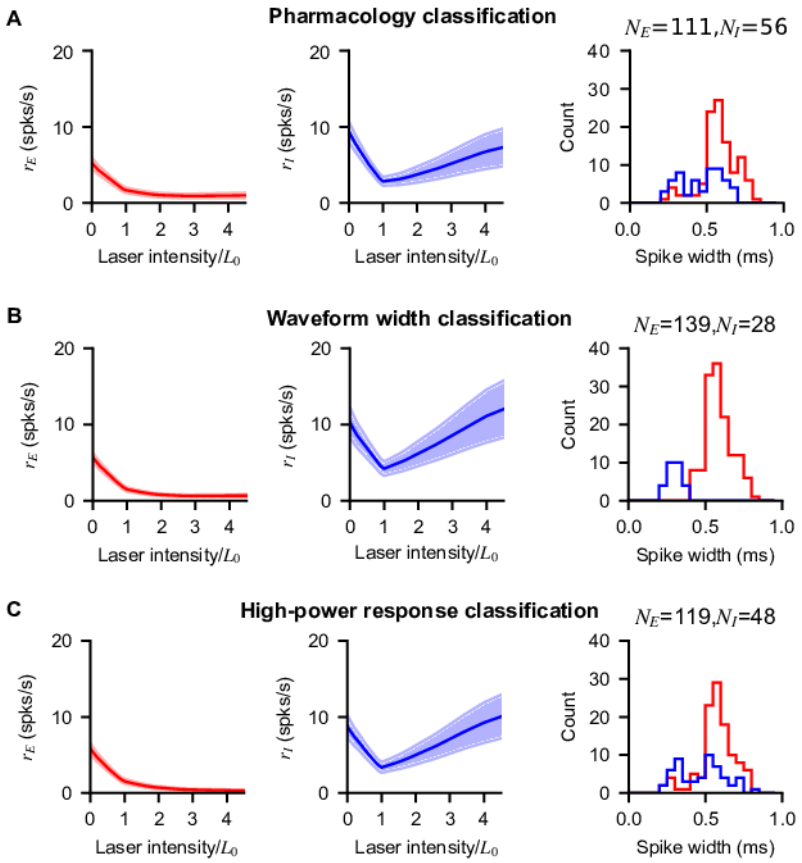

FIG. S7 **Paradoxical effects in V1 are preserved when inhibitory units are classified via pharmacology, via waveform width, and via response at high laser power.** Population response in V1 of excitatory and inhibitory cells as defined by: response with blockers (**A**), waveform width (**B**), and firing rate increase at high laser intensity (**C**). Fig. 2 of the main text uses the classification in shown in (A), but a similar population response results using the two alternative methods of classification. (**B**): units are classified as inhibitory if their waveform width is less than 0.45ms, a value determined by the bimodal distribution observed in the population (left panel). (**C**): units are classified as inhibitory if their firing at maximum laser intensity ($5L/L_0$ in the figure) is one SEM larger than their firing at L_0 (the minimum inhibitory network response).

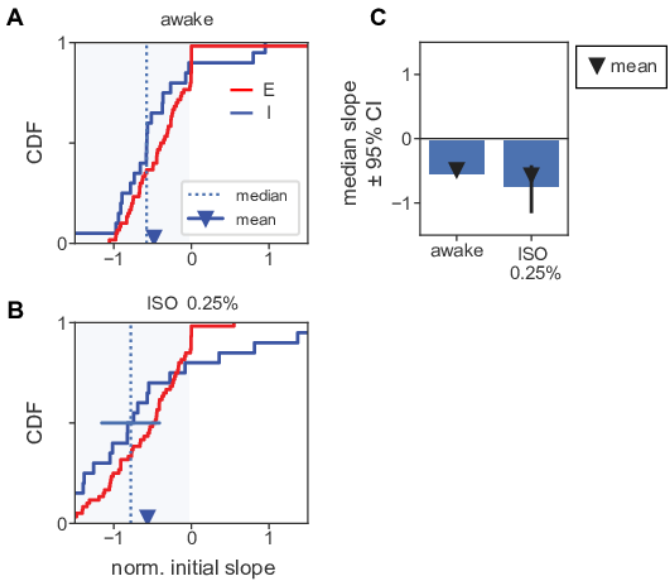

FIG. S8 **Normalized initial slopes for units recorded in anesthesia experiments. (A)** Normalized slopes for E and I units in awake recordings. Conventions as in Fig. 5d. **(B)** Normalized slopes for E and I units in recordings with 0.25% isoflurane. More I units have a positive initial slope than in (A), but the mean and median remain negative. This increase in heterogeneity may be due to the network moving closer to an ISN transition point as firing rates are decreased by anesthesia. **(C)** Summary of mean and median slopes for I units. Both mean and median are significantly negative in both cases. (Mean, awake: $p < 0.01$, iso: $p < 0.05$, via t-test; median, awake: $p < 10^{-5}$, iso: $p < 10^{-3}$, Mann-Whitney U.) Solid black bars: 95% confidence interval for median via bootstrap.

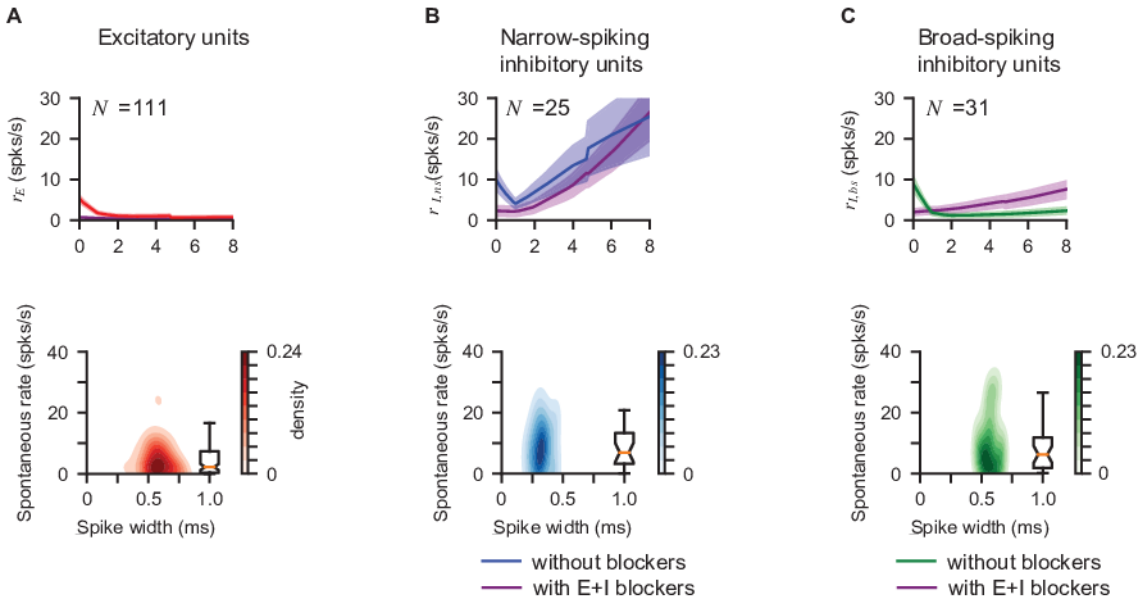

FIG. S9 **Responses of broad- and narrow-spiking V1 (VGAT-ChR2) inhibitory units.** Average rate (first row) and distribution of spontaneous rates vs spike width (second row) of excitatory units (**A**, $N = 112$), narrow-spiking inhibitory units (**B**, waveform width$<$ 0.45ms, $N = 24$), and broad-spiking inhibitory units (**C**, waveform width$>$ 0.45ms, $N = 30$). Corresponding average response with E+I blockers is plotted to show elevated responses with light (first row, purple). Cells are classified as E or I via pharmacological classification (as in Fig. 2). The two inhibitory populations both show initial paradoxical suppression. Narrow-spiking units show a more pronounced increase in firing rate in the non-ISN (high laser power) regime. This difference seems unlikely to result from experimental differences like variation in opsin expression levels, and more likely may be generated by recurrent inhibition. (I.e. narrow spiking putative basket/PV cells provide strong inhibition to wider-spiking somatostatin or other inhibitory cells, and at high laser powers PV cells suppress the wider-spiking units.) In Fig. 6, for both the viral PV-Chronos and transgenic PV-ReachR experiments, we classify cells as inhibitory based on elevated response at high laser power. Consistent with the V1 VGAT-ChR2 data shown above, the cells in PV experiments classfied as I by their high-power response also have narrow waveforms on average.

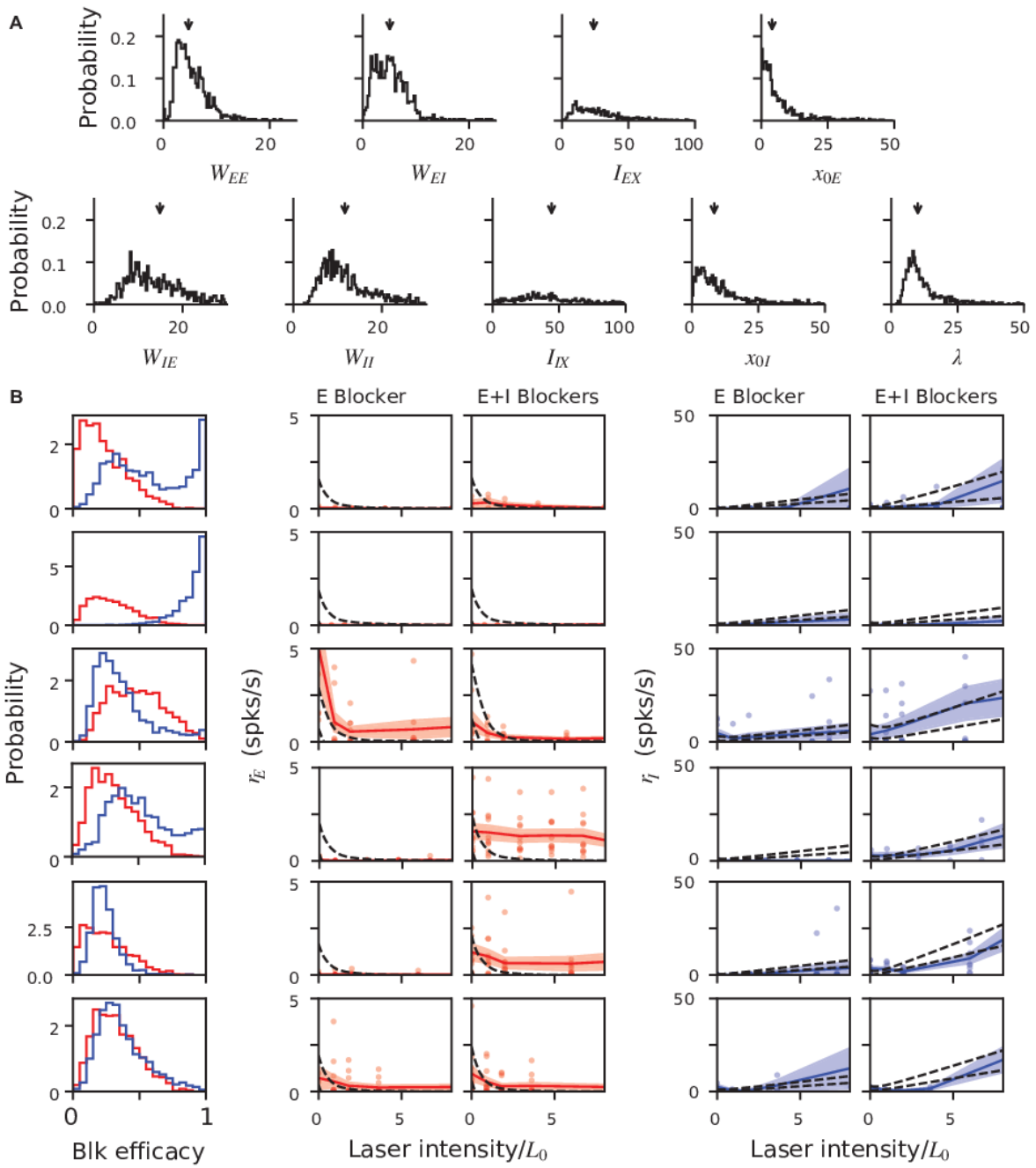

FIG. S10 **V1 model parameter stability shown via data bootstrap.** To examine the stability of the model fit shown in Fig. 4, here we show the distribution of the parameter estimates obtained via bootstrapping (randomly resampling) the data (Methods). (**A**) Distribution (solid lines) of the network parameters (defined by Eq. (6)), and corresponding medians (arrows) over all the bootstrap runs. (**B**) Distribution of the parameters for excitatory (red) and inhibitory (blue) blocker efficacy (left column) with corresponding measured and predicted responses (middle column, red: E cells; right column, blue: I cells). Different rows correspond to different recording sessions. Shaded areas: ±1 SEM across units. The efficacy of blockers varies somewhat from day to day, justifying our use of different efficacy parameters for different recording days in the model.

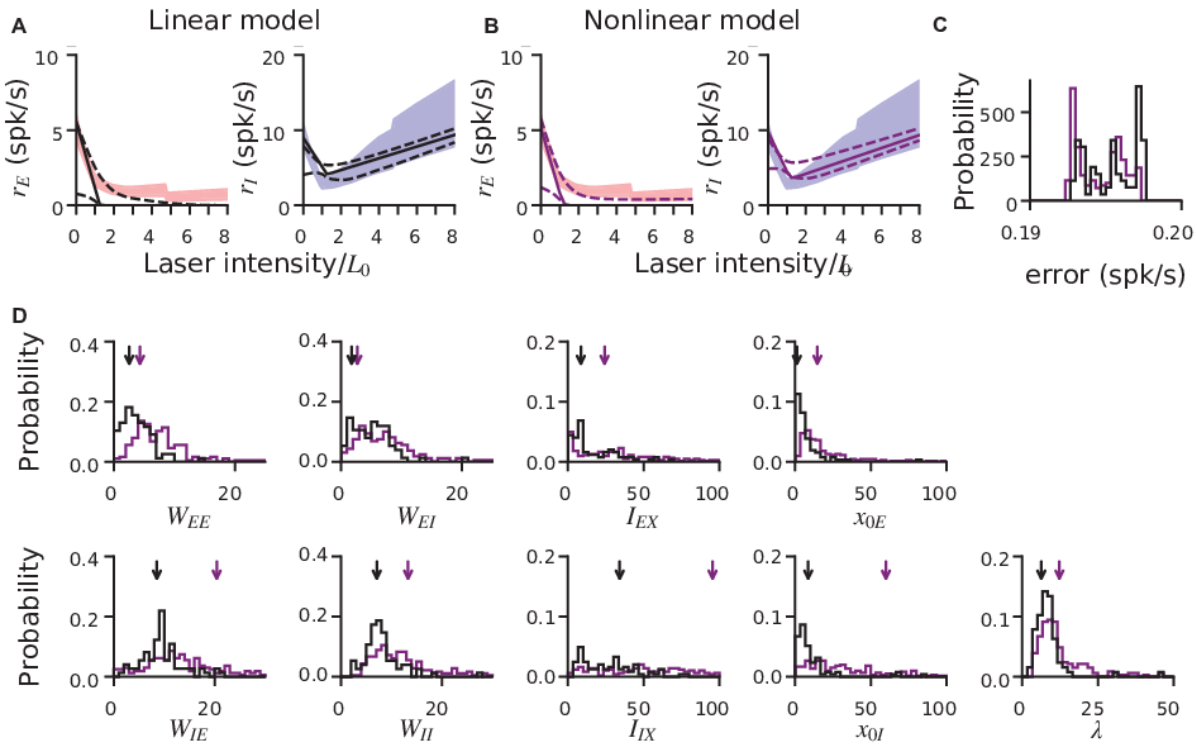

FIG. S11 **Comparison of linear and nonlinear rate models, and global optimization method.** (**A-B**) Similar characterizations of the V1 data are provided by linear and non-linear rate models (i.e. models with either piecewise-linear or non-linear single-cell transfer functions; Methods), even though the non-linear model has one extra parameter (which controls the transfer function nonlinearity; Methods). Shaded regions: means ±1 SEM for E (red) and I (blue) units; data is recorded from V1 of awake animals, same data set as Figs. 2-4. Solid black lines: optimal (lowest-error) fit, across many optimization runs initialized with different parameter values. The lowest-error fit is the global minimum of the optimization cost (error) function. To include only plausible solutions, we retained optimzation results only where error is within 2.5% of optimal. Dashed lines: ±1 s.d. across optimization runs. (**C**) Error values for the two models (black: linear; purple: nonlinear). Error (units: spk/s) is the absolute difference between model and data mean, averaged over all laser intensities. (**D**) Parameters corresponding to solution shown in **A,B**. While the optimal solutions (arrows) show some differences between the two models, the distribution of parameters across runs overlaps substantially, showing the inferred network parameters are robust to the specific model used. Note that the distributions shown here are taken over optimization initialization values, in contrast to Fig. 4, which shows variability of the optimal solution over bootstrap runs.

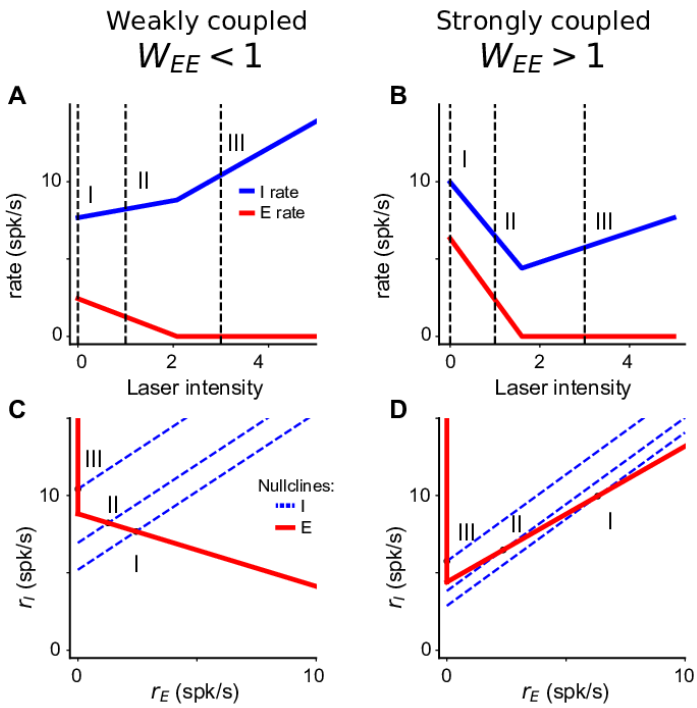

FIG. S12 **Nullclines characterizing the network response at different levels of inhibitory drive.** (**A-B**) Firing rate of E (red) and I (blue) populations computed from Eq. (1) as a function of laser intensity when the excitatory sub-network is stable (A, $W_{EE} < 1$; weak coupling) or unstable (B, $W_{EE} > 1$, strong coupling) . As described in main text, I cells increase firing rate at high laser intensities in both networks. Differences are seen at lower laser intensities, where the inhibition stabilized network ($W_{EE} > 1$) responds paradoxically. (**C-D**) Differences in I responses are seen in the corresponding nullclines (defined as the stationary solutions of the two lines of Eq. (1)) For $W_{EE} < 1$ (C), at low laser intensity (I), the excitatory (red) and the inhibitory (blue) nullclines intersect at a nonzero value of r_E (which represents the stationary point of the network). Increasing the laser intensity shifts the inhibitory nullcline to position (II) and produces a new intersection point with larger r_I and lower r_E. At large laser intensities (III), the two nullclines intersect at r_E equal to zero and the corresponding r_I value increases with the laser intensity. For $W_{EE} > 1$ (D), when the laser intensity increases from (I) to (II), both the stationary value of r_I and r_E decrease. This happens because the excitatory nullcline is an increasing function of r_E, a property which appears when that the excitatory population is unstable [10]. At large laser intensity (III), the excitatory population is silent and stable, both the excitatory and inhibitory nullclines behave as in the $W_{EE} < 1$ case, and increasing the laser intensity increases the inhibitory rate.

REFERENCES

[1] V. Braitenberg and A. Schüz. *Anatomy of the Cortex: Statistics and Geometry*. Studies of Brain Function. Springer Berlin Heidelberg, 2013.

[2] Sandrine Lefort, Christian Tomm, J. C. Floyd Sarria, and Carl C.H. Petersen. The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex. *Neuron*, 2009.

[3] T. Binzegger. A Quantitative Map of the Circuit of Cat Primary Visual Cortex. *Journal of Neuroscience*, 2004.

[4] Alex Thomson and Christophe Lamy. Functional maps of neocortical local circuitry. *Frontiers in Neuroscience*, 1:2, 2007.

[5] C van Vreeswijk and H Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science (New York, N.Y.)*, 274(5293):1724–6, 1996.

[6] C van Vreeswijk and H Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–71, 1998.

[7] D J Amit and N Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 7(3):237–252, 1997.

[8] Ran Rubin, L F Abbott, and Haim Sompolinsky. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *Proceedings of the National Academy of Sciences*, 114(44):E9366–E9375, 2017.

[9] Brendan K. Murphy and Kenneth D. Miller. Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns. *Neuron*, 61(4):635–648, 2009.

[10] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L Mcnaughton. Paradoxical Effects of External Modulation of Inhibitory Interneurons. *Journal of Neuroscience*, 17(11):4382–4388, 1997.

[11] Alfonso Renart, Jaime de la Rocha, Peter Bartho, Liad Hollender, Néstor Parga, Alex Reyes, and Kenneth D Harris. The asynchronous state in cortical circuits. *Science (New York, N.Y.)*, 327(5965):587–590, 01 2010.

[12] N Brunel. Dynamics of sparsely connected networls of excitatory and inhibitory neurons. *Computational Neuroscience*, 8:183–208, 2000.

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

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

[15] Hirofumi Ozeki, Ian M. Finn, Evan S. Schaffer, Kenneth D. Miller, and David Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578 – 592, 2009.

[16] Bassam V. Atallah, William Bruns, Matteo Carandini, and Massimo Scanziani. Parvalbumin-Expressing Interneurons Linearly Transform Cortical Responses to Visual Stimuli. *Neuron*, 73(1):159–170, 2012.

[17] Diego Adrian Gutnisky, Jianing Yu, Samuel Andrew Hires, Minh-Son To, Michael Ross Bale, Karel Svoboda, and David Golomb. Mechanisms underlying a thalamocortical transformation during active tactile sensation. *PLoS computational biology*, 13(6):e1005576; e1005576–e1005576, 06 2017.

[18] Hiroyuki K. Kato, Samuel K. Asinof, and Jeffry S. Isaacson. Network-Level Control of Frequency Tuning in Auditory Cortex, 2017.

[19] Alexandra K. Moore, Aldis P. Weible, Timothy S. Balmer, Laurence O. Trussell, and Michael Wehr. Rapid Rebalancing of Excitation and Inhibition by Cortical Circuitry. *Neuron*, 97(6):1341–1355.e6, mar 2018.

[20] Ashok Litwin-Kumar, Robert Rosenbaum, and Brent Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *Journal of Neurophysiology*, 115(3):1399–1409, 2016. PMID: 26740531.

[21] Benjamin R. Arenkiel, Joao Peca, Ian G. Davison, Catia Feliciano, Karl Deisseroth, George J. Augustine, Michael D. Ehlers, and Guoping Feng. In vivo light-induced activation of neural circuitry in transgenic mice expressing channelrhodopsin-2. *Neuron*, 54(2):205 – 218, 2007.

[22] Julie A. Harris, Stefan Mihalas, Karla E. Hirokawa, Jennifer D. Whitesell, Joseph Knox, Amy Bernard, Philip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, David Feng, Nathalie Gaudreault, Nile Graddis, Peter A. Groblewski, Alex Henry, Anh Ho, Robert Howard, Leonard Kuan, Jerome Lecoq, Jennifer Luviano, Stephen McConoghy, Marty T. Mortrud, Maitham Naeemi, Lydia Ng, Seung W. Oh, Benjamin Ouellette, Staci A. Sorensen, Wayne Wakeman, Quanxin Wang, Ali Williford, John W. Phillips, Allan Jones, Christof Koch, and Hongkui Zeng. The organization of intracortical connections by layer and cell class in the mouse brain. *bioRxiv*, 2018.

[23] Javier DeFelipe, Pedro L. López-Cruz, Ruth Benavides-Piccione, Concha Bielza, Pedro Larrañaga, Stewart Anderson, Andreas Burkhalter, Bruno Cauli, Alfonso Fairén, Dirk Feldmeyer, Gord Fishell, David Fitzpatrick, Tamás F. Freund, Guillermo González-Burgos, Shaul Hestrin, Sean Hill, Patrick R. Hof, Josh Huang, Edward G. Jones, Yasuo Kawaguchi, Zoltán Kisvárday, Yoshiyuki Kubota, David A. Lewis, Oscar Marín, Henry Markram, Chris J. McBain, Hanno S. Meyer, Hannah Monyer, Sacha B. Nelson, Kathleen Rockland, Jean Rossier, John L. R. Rubenstein, Bernardo Rudy, Massimo Scanziani, Gordon M. Shepherd, Chet C. Sherwood, Jochen F. Staiger, Gábor Tamás, Alex Thomson, Yun Wang, Rafael Yuste, and Giorgio A. Ascoli. New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nature Reviews Neuroscience*, 14(3):202–216, mar 2013.

[24] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits. *Neuron*, 91(2):260–292, jul 2016.

[25] Troy W. Margrie, Michael Brecht, and Bert Sakmann. In vivo, low-resistance, whole-cell recordings from neurons in the anaesthetized and awake mammalian brain. *Pflügers Archiv*, 444(4):491–498, Jul 2002.

[26] Bruno A. Olshausen and David J. Field. How close are we to understanding v1? *Neural Computation*, 17(8):1665–1699, 2005.

[27] D. A. McCormick, Barry W Connors, James W Lighthall, and David A Prince. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *Journal*

of Neurophysiology, 54(4):782–806, oct 1985.

[28] Garrett T. Neske, Saundra L. Patrick, and Barry W. Connors. Contributions of diverse excitatory and inhibitory neurons to recurrent network activity in cerebral cortex. *Journal of Neuroscience*, 35(3):1089–1105, 2015.

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

[30] L. L. Glickfeld, M. H. Histed, and J. H. R. Maunsell. Mouse Primary Visual Cortex Is Used to Detect Both Orientation and Contrast Changes. *Journal of Neuroscience*, 33(50):19416–19422, 2013.

[31] Dennis R Sparta, Nanna Hovelso, Alex O Mason, Pranish A Kantak, Randall L Ung, Heather K Decot, and Garret D Stuber. Activation of prefrontal cortical parvalbumin interneurons facilitates extinction of reward-seeking behavior. *J Neurosci*, 34(10):3699–3705, Mar 2014.

[32] Nathan C Klapoetke, Yasunobu Murata, Sung Soo Kim, Stefan R Pulver, Amanda Birdsey-Benson, Yong Ku Cho, Tania K Morimoto, Amy S Chuong, Eric J Carpenter, Zhijian Tian, Jun Wang, Yinlong Xie, Zhixiang Yan, Yong Zhang, Brian Y Chow, Barbara Surek, Michael Melkonian, Vivek Jayaraman, Martha Constantine-Paton, Gane Ka-Shu Wong, and Edward S Boyden. Independent optical excitation of distinct neural populations. *Nature Methods*, 11(3):338–346, mar 2014.

[33] Simon Hippenmeyer, Eline Vrieseling, Markus Sigrist, Thomas Portmann, Celia Laengle, David R Ladle, and Silvia Arber. A developmental switch in the response of drg neurons to ets transcription factor signaling. *PLoS biology*, 3(5):e159; e159–e159, 05 2005.

[34] Hiroki Taniguchi, Jiangteng Lu, and Z Josh Huang. The spatial and temporal origin of chandelier cells in mouse neocortex. *Science*, 339(6115):70–74, Jan 2013.

[35] Akiya Watakabe, Masanari Ohtsuka, Masaharu Kinoshita, Masafumi Takaji, Kaoru Isa, Hiroaki Mizukami, Keiya Ozawa, Tadashi Isa, and Tetsuo Yamamori. Comparative analyses of adeno-associated viral vector serotypes 1, 2, 5, 8 and 9 in marmoset, mouse and macaque cerebral cortex. *Neuroscience Research*, 93:144–157, 2015.

[36] John Y Lin, Per Magne Knutsen, Arnaud Muller, David Kleinfeld, and Roger Y Tsien. Reachr: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat Neurosci*, 16(10):1499–1508, Oct 2013.

[37] Hillel Adesnik. Synaptic Mechanisms of Feature Coding in the Visual Cortex of Awake Mice. *Neuron*, 95(5):1147–1159.e4, aug 2017.

[38] Hillel Adesnik, William Bruns, Hiroki Taniguchi, Z Josh Huang, and Massimo Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–231, 2013.

[39] Nuo Li, Susu Chen, Zengcai V. Guo, Han Chen, Yan Huo, Hidehiko K. Inagaki, Courtney Davis, David Hansel, Caiying Guo, and Karel Svoboda. Spatiotemporal limits of optogenetic manipulations in cortical circuits. *bioRxiv*, 2019.

[40] Davi D. Bock, Wei Chung Allen Lee, Aaron M. Kerlin, Mark L. Andermann, Greg Hood, Arthur W. Wetzel, Sergey Yurgenson, Edward R. Soucy, Hyon Suk Kim, and R. Clay Reid.

Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 2011.

[41] Kazuhiro Sohya, Katsuro Kameyama, Yuchio Yanagawa, Kunihiko Obata, and Tadaharu Tsumoto. Gabaergic neurons are less selective to stimulus orientation than excitatory neurons in layer ii/iii of visual cortex, as revealed by in vivo functional ca2+ imaging in transgenic mice. *Journal of Neuroscience*, 27(8):2145–2149, 2007.

[42] Aaron M. Kerlin, Mark L. Andermann, Vladimir K. Berezovskii, and R. Clay Reid. Broadly Tuned Response Properties of Diverse Inhibitory Neuron Subtypes in Mouse Visual Cortex. *Neuron*, 2010.

[43] Elodie Fino and Rafael Yuste. Dense inhibitory connectivity in neocortex. *Neuron*, 69(6):1188–1203, 03 2011.

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

[45] Gianluigi Mongillo, David Hansel, and Carl van Vreeswijk. Bistability and spatiotemporal irregularity in neuronal networks with nonlinear synaptic transmission. *Phys. Rev. Lett.*, 108:158101, Apr 2012.

[46] Ying Ma, Mohammed A. Shaik, Sharon H. Kim, Mariel G. Kozberg, David N. Thibodeaux, Hanzhi T. Zhao, Hang Yu, and Elizabeth M.C. Hillman. Wide-field optical mapping of neural activity and brain haemodynamics: Considerations and novel approaches, oct 2016.

[47] Cynthia H. Chen-Bee, Teodora Agoncillo, Ying Xiong, and Ron D. Frostig. The triphasic intrinsic signal: Implications for functional imaging. *Journal of Neuroscience*, 27(17):4572–4586, 2007.

[48] J. Alexander Heimel, Robin J. Hartman, Josephine M. Hermans, and Christiaan N. Levelt. Screening mouse vision with intrinsic signal optical imaging. *European Journal of Neuroscience*, 25(3):795–804, 2019/05/01 2007.

[49] Yinghong Xiao, David C Martin, Xinyan Cui, and Mahesh Shenai. Surface modification of neural probes with conducting polymer poly(hydroxymethylated-3,4- ethylenedioxythiophene) and its biocompatibility. *Appl Biochem Biotechnol*, 128(2):117–130, Feb 2006.

[50] Mark H. Histed and John H. R. Maunsell. Cortical neural populations can guide behavior by integrating inputs linearly, independent of synchrony. *Proceedings of the National Academy of Sciences*, 111(1):E178–E187, 2014.

[51] Ryan C. Kelly, Matthew A. Smith, Jason M. Samonds, Adam Kohn, A. B. Bonds, J. Anthony Movshon, and Tai Sing Lee. Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. *Journal of Neuroscience*, 27(2):261–264, 2007.

[52] Mark H. Histed. Feedforward inhibition allows input summation to vary in recurrent cortical networks. *eNeuro*, 2018.

[53] H R Wilson and J D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 01 1972.

[54] Seung Wook Oh, Julie A. Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M. Henry, Marty T. Mortrud, Benjamin Ouellette, Thuc Nghi Nguyen, Staci A. Sorensen, Clifford R. Slaughterbeck, Wayne Wakeman, Yang Li, David Feng, Anh Ho, Eric Nicholas, Karla E. Hirokawa, Phillip Bohn, Kevin M.

Joines, Hanchuan Peng, Michael J. Hawrylycz, John W. Phillips, John G. Hohmann, Paul Wohnoutka, Charles R. Gerfen, Christof Koch, Amy Bernard, Chinh Dang, Allan R. Jones, and Hongkui Zeng. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, apr 2014.

[55] Ryan C. Kelly, Matthew A. Smith, Jason M. Samonds, Adam Kohn, A. B. Bonds, J. Anthony Movshon, and Tai Sing Lee. Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. *Journal of Neuroscience*, 27(2):261–264, 2007.