Untangling stability and gain modulation in cortical circuits with multiple interneuron classes

Hannah Bos[1,2], Anne-Marie Oswald[2,3,4], and Brent Doiron[1,2,4,5]

[1]Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, USA
[2]Center for the Neural Basis of Cognition, Pittsburgh, PA, USA
[3]Department of Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA
[4]Department of Neurobiology, University of Chicago, Chicago, IL, USA
[5]Department of Statistics, University of Chicago, Chicago, IL, USA

Abstract

Synaptic inhibition is the mechanistic backbone of a suite of cortical functions, not the least of which is maintaining overall network stability as well as modulating neuronal gain. Past cortical models have assumed simplified recurrent networks in which all inhibitory neurons are lumped into a single effective pool. In such models the mechanics of inhibitory stabilization and gain control are tightly linked in opposition to one another – meaning high gain coincides with low stability and vice versa. This tethering of stability and response gain restricts the possible operative regimes of the network. However, it is now well known that cortical inhibition is very diverse, with molecularly distinguished cell classes having distinct positions within the cortical circuit. In this study, we analyze populations of spiking neuron models and associated mean-field theories capturing circuits with pyramidal neurons as well as parvalbumin (PV) and somatostatin (SOM) expressing interneurons. Our study outlines arguments for a division of labor within the full cortical circuit where PV interneurons are ideally positioned to stabilize network activity, whereas SOM interneurons serve to modulate pyramidal cell gain. This segregation of inhibitory function supports stable cortical dynamics over a large range of modulation states. Our study offers a blueprint for how to relate the circuit structure of cortical networks with diverse cell types to the underlying population dynamics and stimulus response.

Introduction

A prominent feature of cortical neurons is that their responses to stimuli are quite malleable, depending upon a host of factors. For instance, the global structure of a sensory scene activates surround receptive fields, normalizing the response of cortical neurons to their preferred stimuli (Adesnik et al., 2012; Carandini and Heeger, 2012; Reynolds and Heeger, 2009; Vinje and Gallant, 2000). Alternatively, top-down projections can also mediate how cortical responses change with directed attention (Cohen and Maunsell, 2009; Ruff et al., 2018), subject arousal (McGinley et al., 2015), or task engagement (Downer et al., 2015). On the whole, these neuronal response shifts are reliably signaled by changes in neuronal firing rates (Carandini and Heeger, 2012; Harris and Thiele, 2011; Niell and Stryker, 2010; Ruff et al., 2018), the spectral content of local field potentials (Harris and Thiele, 2011; Niell and Stryker, 2010) as well as single neuron membrane potentials (Poulet and Petersen, 2008), and finally the magnitude of spike train correlations across a population (Doiron et al., 2016). These observations have prompted studies that focus on how inhibitory circuits are key mediators of cortical state modulations. Indeed, inhibition has been implicated in the suppression of neuronal activity (Adesnik, 2017; Adesnik et al., 2012; Haider et al., 2013; Kato et al., 2017), gain control of pyramidal neuron firing rates (Ferguson and Cardin, 2020; Katzner et al., 2011; Phillips and Hasenstaub, 2016; Silver, 2010) and correlated neuronal fluctuations (Okun and Lampl, 2008), rhythmic population

activity (Atallah and Scanziani, 2009; Womelsdorf et al., 2014), and spike timing of pyramidal neurons (Berman and Maler, 1998; Wehr and Zador, 2003). However, inhibition must also prevent runaway cortical activity that would otherwise lead to pathological activity (Haider et al., 2013; Ozeki et al., 2009; Veit et al., 2017), enforcing constraints on how inhibition can modulate pyramidal neuron activity. In sum, using inhibitory interactions to expose the physiological and circuit basis for how cortical activity changes depending upon processing or cognitive needs has been a longstanding and popular avenue of study (Isaacson and Scanziani, 2011).

While inhibition has been long measured (Eccles et al., 1954; Hartline et al., 1956; Lloyd, 1946), the past fifteen years have witnessed a newfound appreciation of its diversity. The invention and widespread use of cell-specific labeling and optogenetic control (Fenno et al., 2011), combined with the detailed genetic and physiological characterization of cortical interneurons (Jiang et al., 2015; Markram et al., 2004) has painted a complex picture of a circuit that was previously considered to be simpler (Douglas et al., 1989). The standard cortical circuit now includes (at a minimum) somatostatin (SOM) and parvalbumin (PV) expressing interneuron classes, with distinct synaptic interactions between these classes as well as with pyramidal neurons (Jiang et al., 2015; Kepecs and Fishell, 2014; Pfeffer et al., 2013; Tremblay et al., 2016). This new circuit reality presents some clear challenges (Cardin, 2018; Ferguson and Cardin, 2020; Urban-Ciecko and Barth, 2016; Wood et al., 2017; Yavorska and Wehr, 2016), foremost being to uncover how state changes that were previously associated with inhibition in a broad sense, should be distributed over these diverse interneuron classes.

An attractive hypothesis is that SOM and PV interneurons are within-group functionally homogeneous, yet each class performs functions that are distinct from those of the other classes (Hattori et al., 2017; Kepecs and Fishell, 2014; Wang et al., 2004). In line with this idea, early optogenetic studies of PV and SOM neurons in the mouse visual cortex showed differential multiplicative and subtractive modulations of excitatory neuron responses (Atallah et al., 2012; Wilson et al., 2012). However, such tidy arithmetic of inhibitory modulation is likely an over-abstraction, and a detailed analysis of data from the auditory cortex shows a mixed modulatory influence of both SOM and PV interneurons (Seybold et al., 2015). Another popular functional distinction between interneuron classes is their roles in disinhibitory cortical circuits (Fu et al., 2014; Large et al., 2018; Lee et al., 2013; Pi et al., 2013; Wang and Yang, 2018), specifically when one interneuron class inhibits another class and thereby releasing pyramidal neurons from a source of inhibition. Yet inhibition can be complicated to dissect in the full cortical circuit. For instance, the SOM $\rightarrow$ E inhibitory pathway competes with the SOM $\rightarrow$ PV $\rightarrow$ E disinhibitory pathway for pyramidal neuron (E) influence. This competition could underlie differences in how SOM neurons suppress pyramidal neuron activity in layer 2/3 of visual and auditory cortex (Adesnik, 2017; Adesnik et al., 2012; Kato et al., 2017) yet increase activity in layer 4 of somatosensory cortex (Xu et al., 2013). These difficulties in interpretation likely arise from an incomplete analysis: selected feedforward sub motifs within the full recurrent circuit are considered, yet with a tacit ignorance of the other pathways that will nevertheless still contribute to the cortical response. A proper untangling of the complexity of a fully recurrent cortical circuit requires a modeling framework where a rigorous analysis can be performed.

Cortical models with both excitatory and inhibitory pathways have a long history of study (Griffith, 1963; Wilson and Cowan, 1972). Models with just a single inhibitory interneuron class have successfully explained a wide range of cortical behavior; from contrast dependent nonlinearities in cortical response (Ozeki et al., 2009; Rubin et al., 2015), to the genesis of irregular and variable spike discharge (Brunel and Hakim, 1999; van Vreeswijk and Sompolinsky, 1996), and finally to the mechanisms underlying high-frequency cortical network rhythms (Bos et al., 2016; Wang, 2010). On the surface, it may seem surprising that cortical models that ignore inhibitory diversity can nevertheless account for such a range of cortical behavior. However, these models are often designed to capture only a single aspect of cortical responses, and rarely do they account for how a set of neuronal correlates (firing rates, response gain, neuronal correlations, etc.) shift with the cortical state. In truth, perhaps the most compelling reason that theorists lag behind the reality of a diverse interneuron cortical circuit is that there are serious complications in model analysis when multiple interneuron classes are introduced. Indeed, the set of model parameters and the landscape of possible solutions

both increase dramatically with the addition of new neuron classes. In total, there is a real need for a cogent analysis of the behavior of multi-interneuron cortical circuits models.

We present a circuit theory for previously developed multi-interneuron cortical circuit models (del Molino et al., 2017; Kuchibhotla et al., 2017; Litwin-Kumar et al., 2016; Mahrach et al., 2020; Veit et al., 2017) with the goal of giving a mechanistic understanding of how diverse inhibitory interneurons participate in the circuit level modulation of cortical responses. In particular, we consider modulatory inputs to SOM neurons which aim to shift the operative state of the full E – PV – SOM neuron circuit. We expose a circuit-based relationship between how SOM neuron modulations co-determine network stability and response gain. More to the point, the underlying E – PV – SOM circuit supports a division of labor (Wang et al., 2004), whereby PV neurons are well-positioned to provide network stability, allowing SOM neurons the freedom to modulate response gain. Our theoretical framework offers an attractive platform to probe how interneuron circuit structure determines gain and stability which may generalize well beyond the sensory cortices where these interneuron circuits are currently best characterized.

Results

The inhibitory and disinhibitory pathways of the E – PV – SOM circuit

There is strong *in vivo* evidence that SOM interneurons play a critical role in the modulation of cortical response (Urban-Ciecko and Barth, 2016; Yavorska and Wehr, 2016). However, the complex wiring between excitatory and inhibitory neurons (Jiang et al., 2015; Pfeffer et al., 2013; Tremblay et al., 2016) presents a challenge when trying to expose the specific mechanisms by which SOM neurons modulate cortical response. Two distinct inhibitory circuit pathways are often considered when disentangling the impact of SOM inhibition on pyramidal neuron (E) response. We introduce these pathways with a pair of studies that are emblematic of these circuit motifs.

In the first line of study, SOM neuron activity in layer 2/3 of the mouse visual cortex was recruited by expanding the spatial scale of an orientated drifting grating visual stimulus (Adesnik, 2017; Adesnik et al., 2012). This resulted in decreased activity of putative layer 2/3 E neurons whose spatial receptive field was in the center of a visual image. The simplest interpretation is that the increased SOM activity inhibited E neurons via direct SOM $\rightarrow$ E connectivity. Similar suppression of E neuron response from direct SOM inhibition has been implicated in other studies (Wang and Yang, 2018), and a disinhibition of SOM $\rightarrow$ E projections is often mediated through vasoactive intestinal-peptide (VIP) neurons projections to SOM neurons (Fu et al., 2014; Pi et al., 2013). In the second study, layer 4 SOM neurons in mouse somatosensory cortex were optogenetically silenced (Xu et al., 2013). This resulted in increased activity of PV neurons, and a subsequent decreased activity of E neurons. The authors intuited a suppression of the disinhibitory pathway SOM $\rightarrow$ PV $\rightarrow$ E as the source of reduced E neuron activity. Taken together, the pair of studies seem in opposition to one another, with SOM neuron activity providing either a source or a relief of E neuron suppression. This response dichotomy prompted us to consider what physiological and circuit properties of the E – PV – SOM circuit are critical determinants of whether an increase in SOM neuron activity results in an increase or a decrease in E neuron response.

An answer to this question requires consideration of the full recurrent connectivity within the E – PV – SOM neuron circuit, as opposed to analysis restricted to just the SOM $\rightarrow$ E and SOM $\rightarrow$ PV $\rightarrow$ E sub motifs within the circuit. Fortunately, there is a detailed physiological characterization of the specific connectivity patterns of inhibitory subtypes within the cortical circuit (Jiang et al., 2015; Pfeffer et al., 2013; Tremblay et al., 2016). Briefly, PV neurons couple strongly to other PV neurons as well as to E neurons, while SOM neurons connect strongly to PV and E neurons, but not to other SOM neurons. Finally, both PV and SOM neurons receive inputs from E neurons. We incorporate these circuit details in a network of leaky integrate-and-fire model neurons where the E, PV, and SOM subclasses are represented (Fig. 1A; see Methods 1.1). The simplicity of the spiking dynamics makes our model amenable to a mean-field reduction which captures the bulk spiking activity of each subpopulation of neurons (where r_{E}, r_{P}, and r_{S} denote the E, PV, and SOM populations respectively;

see Methods 1.2), as has been done by similar studies of the E – PV – SOM cortical circuit (del Molino et al., 2017; Kuchibhotla et al., 2017; Litwin-Kumar et al., 2016; Mahrach et al., 2020).

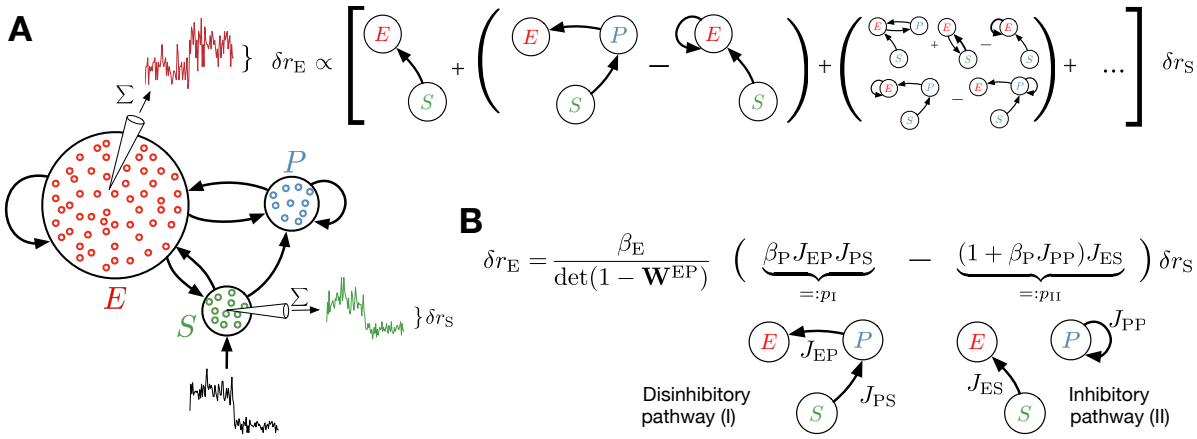

Figure 1: **Tradeoff between two inhibitory motifs in the E – PV – SOM cortical circuit. A,** Sketch of the network model with neuron class-specific connections motivated by (Pfeffer et al., 2013). A modulatory input (black line) is applied to the SOM neurons. This input results in a change in the SOM neuron population firing rate (δr_{S}; green line), which in turn causes a change of excitatory neuron activity (δr_{E}; red line). Assuming a linear relation between δr_{S} and δr_{E} allows δr_{E} to be determined by summing over all possible pathways through the network by which SOM activity can influence E activity (see Eq. (23)). **B,** The relation between δr_{S} and δr_{E} after summing over all paths (see Eq. (26)). Sketches visualizes the tradeoff between the inhibitory and disinhibitory pathways. $J_{\kappa\gamma}$ summarizes the number and strength of connections from population γ to population κ (κ ,$\gamma \in \{\mathrm{E}, \mathrm{P}, \mathrm{S}\}$, see Eq. (7)); β_κ denotes the cellular gain of population κ (see Eq. (16)), and $\mathbf{W}^{\mathrm{EP}}$ the effective connectivity matrix between the E and PV populations (see Eq. (17)).

Using our model we ask how a modulation of the SOM neuron activity, resulting in $r_{\mathrm{S}} \to r_{\mathrm{S}} + \delta r_{\mathrm{S}}$, is transferred to a modulation of E neuron activity yielding $r_{\mathrm{E}} \to r_{\mathrm{E}} + \delta r_{\mathrm{E}}$ (Fig. 1A, green and red inset timeseries). If δr_{S} is sufficiently small we can linearize around a given dynamical state of the model so that $\delta r_{\mathrm{E}} = L_{\mathrm{ES}} \delta r_{\mathrm{S}}$, where L_{ES} is the transfer coefficient between SOM and E neuron modulations. In principle L_{ES} depends on the synaptic matrix J that defines the coupling between neuron classes, as well as the cellular gain, β, of all neurons classes (see Methods 1.3). Since the neuron transfer is nonlinear (see Eq. (10)), the β values, and by extension L_{ES}, depend upon the operating point of the network. While L_{ES} is straightforward to compute (del Molino et al., 2017; Litwin-Kumar et al., 2016) it can be cumbersome, and extracting clear insight about how J and β influence modulation can be difficult.

It is instructive to express L_{ES} as a sum over all possible synaptic pathways by which SOM neuron activity can influence E neuron activity (Fig. 1A, right; see Eq. (23)). Fortunately, this sum can be simplified so that just two network motifs determine the sign of L_{ES} (Fig. 1B; see Methods 1.3). These motifs reflect both the disinhibitory component of the network (the SOM $\to$ PV $\to$ E connections, labeled pathway p_{I} in Fig. 1B) and the inhibitory component (the PV $\to$ PV and SOM $\to$ E connections, labeled pathway p_{II} in Fig. 1B). Interestingly, these motifs corresponds to the pathways evoked in our motivating pair of experimental studies (p_{I} in (Xu et al., 2013) and p_{II} in (Adesnik et al., 2012)). The net pathway is a tradeoff between these two motifs, where p_{I} vies for disinhibition ($L_{\mathrm{ES}} > 0$) while p_{II} vies for inhibition ($L_{\mathrm{ES}} < 0$).

Whether the full motif is inhibitory or disinhibitory depends (somewhat expectantly) on the three connection strengths J_{EP}, J_{PS}, and J_{ES} (J_{EP} is the synaptic strength of the PV $\to$ E neuron pathway; the other connections have the same nomenclature). However, what is unexpected is that it also depends on the recurrent strength between PV neurons, J_{PP} (the $\beta_{\mathrm{P}} J_{\mathrm{PP}}$ term in pathway p_{II}). Further, since β_{P} depends on the operating point of the network then the tradeoff between the two

pathways can be controlled by PV neuron modulation. In particular, since β_{P} increases with r_{P} then L_{ES} can transition from effectively inhibitory for low PV activity (small β_{P}) to effectively disinhibitory for higher PV activity (large β_{P}). However, this is only possible if the combined SOM $\rightarrow$ PV and PV $\rightarrow$ E connections are stronger than the combined PV $\rightarrow$ PV and SOM $\rightarrow$ E connections. We remark that other connections and the activity of the excitatory and SOM neurons only contribute to the amplitude of the effective pathway (as reflected in the prefactor in Fig. 1B). Thus, changing the activity of E and SOM neurons does not explicitly change the sign of L_{ES}.

Pfeffer et al. (Pfeffer et al., 2013) report that in L2/3 and L5 of the mouse visual cortex the SOM $\rightarrow$ E connection is stronger than the SOM $\rightarrow$ PV connection, while the connections from PV $\rightarrow$ E neurons and PV $\rightarrow$ PV neurons are comparably strong. Given these facts, our analysis suggests that the inhibitory pathway p_{II} outweighs the disinhibitory pathway p_{I}, so that $L_{\mathrm{ES}} < 0$. This is consistent with the SOM suppression of E neuron activity by large drifting gratings reported by Adesnik et al. (Adesnik et al., 2012) in L2/3 of mouse visual cortex. Alternatively, Xu et al. (Xu et al., 2013) find stronger SOM $\rightarrow$ PV than SOM $\rightarrow$ E connections in L4 mouse somatosensory cortex. However, since the relative strength of the PV $\rightarrow$ E and PV $\rightarrow$ PV connection strengths are unknown, our analysis cannot predict whether the effective pathway is inhibitory or disinhibitory.

In sum, while the full E – PV – SOM recurrent circuit invokes a multitude of polysynaptic pathways, a tradeoff between the two central disynaptic pathways often appealed to in literature, p_{I} and p_{II}, does indeed determine the modulatory influence of SOM neurons upon E neurons. Having now identified the central role of these two pathways, in the following sections we investigate how they control the stimulus – response gain of E neurons.

Comparison of gain modulation by the two inhibitory pathways

Gain modulation refers to changes in the sensitivity of neuron activity to changes in a driving input (Ferguson and Cardin, 2020; Silver, 2010; Williford and Maunsell, 2006). Typically, the driving input is a feature of a sensory scene and individual neuron responses show tuning to specific feature value. In a later section we will consider gain modulation of tuned responses, but to begin we simply consider a homogeneous input (I_{stim}) that targets both E and PV neurons (Tremblay et al., 2016), and we compute the E neuron *network gain* as the $dr_{\mathrm{E}}/dI_{\mathrm{stim}}$ (Eq. (27)). Here, network gain measures the sensitivity of r_{E} owing to the activity of the full recurrent circuit in response to a change in I_{stim}. This is opposed to the cellular gain β_{E} which measures the sensitivity of r_{E} to a change in the full input current to E neurons due to both the external stimulus and internal interactions. Our circuit model allows us to compare and contrast the effectiveness of network gain modulation via the disinhibitory pathway from SOM $\rightarrow$ PV $\rightarrow$ E neurons (p_{I}; Fig. 2Ai) to modulation via the direct SOM $\rightarrow$ E pathway (p_{II}; Fig. 2Bi). In what follows we explore the E – PV – SOM circuit with both simulations of populations of integrate-and-fire model neurons (Methods 1.1) as well as an associated firing rate model (Methods 1.2).

We first address modulation via the disinhibitory pathway p_{I} with SOM neurons being depolarized by a modulatory input ($I_{\mathrm{mod}} > 0$). Examples such a modulation include suppressed VIP inhibition onto SOM neurons (Pi et al., 2013), activation of pyramidal cells located outside the circuit yet preferentially projecting to SOM neurons (Adesnik et al., 2012), and direct cholinergic modulation of SOM neurons (Kuchibhotla et al., 2017; Urban-Ciecko and Barth, 2016). Modulations that increase SOM neurons activity (Fig. 2Aii, green) suppress the activity of PV neurons (Fig. 2Aii, blue), thereby removing inhibition from E neurons and increasing their firing rates (Fig. 2Aii, red). For small modulations ($I_{\mathrm{mod}} < 15$ Hz), the slight increase in E neuron firing rates causes an increase in E neuron network gain by increasing cellular gain β_E (Fig. 2Aiii, black). For larger modulations ($I_{\mathrm{mod}} > 20$ Hz), PV neuron firing rates increase due to increased excitatory feedback from E $\rightarrow$ PV neurons which overcomes the feedforward inhibition of the SOM $\rightarrow$ PV neuron pathway. The switch from suppressed PV activity to enhanced PV activity marks the transition of the E – PV sub-circuit from a non-inhibition stabilized network (non-ISN) to an inhibition stabilized network (ISN) (Litwin-Kumar et al., 2016; Ozeki et al., 2009; Tsodyks et al., 1997). In the ISN state, the recurrently coupled E neurons are not stable by themselves but are stabilized by recurrent inhibition from PV neurons. Our firing rate model predicts that increased recurrent inhibition also suppresses network response to

input and can decrease E neuron network gain for sufficiently strong modulations (Fig. 2Aiii, black curve). The maximum in response gain occurs after the network has transitioned to the ISN regime (see Methods 1.4).

At high values of modulatory input ($I_{\text{mod}} > 30$ Hz) there is a severe disagreement between the firing rates and network gain obtained with the integrate-and-fire neuron simulations compared to those from analysis of the firing rate model (Fig. 2Aiii, compare points to the curve). The firing rate description assumes near Poisson spiking dynamics and negligible correlations between neuron spike trains (Methods 1.2). Thus, the discrepancy between theory and simulations hints at a shift in the spiking network dynamics away from the asynchronous irregular regime (Brunel, 2000; Renart et al., 2010). Indeed, while the population spike times of the E neurons in the integrate-and-fire simulations show irregular spiking for low modulation (Fig. 2Aiv, top), this gives way to more synchronous activity when reaching maximal gain (Fig. 2Aiv, middle) and eventually pathologic synchronous population bursts for large modulation (Fig. 2Aiv, bottom). Past modeling work has shown how fast recurrent inhibition within a cortical circuit can maintain an asynchronous spiking dynamic (Huang et al., 2019; Renart et al., 2010; Tetzlaff et al., 2012). Thus, it is to be expected that suppressing the normally strong recurrent PV inhibition via pathway p_{I} will lead to a breakdown of the asynchronous state that is supported in the E – PV subcircuit. In total, modulation of E neuron network gain through the disinhibitory pathway p_{I} is restricted to only sufficiently weak modulations.

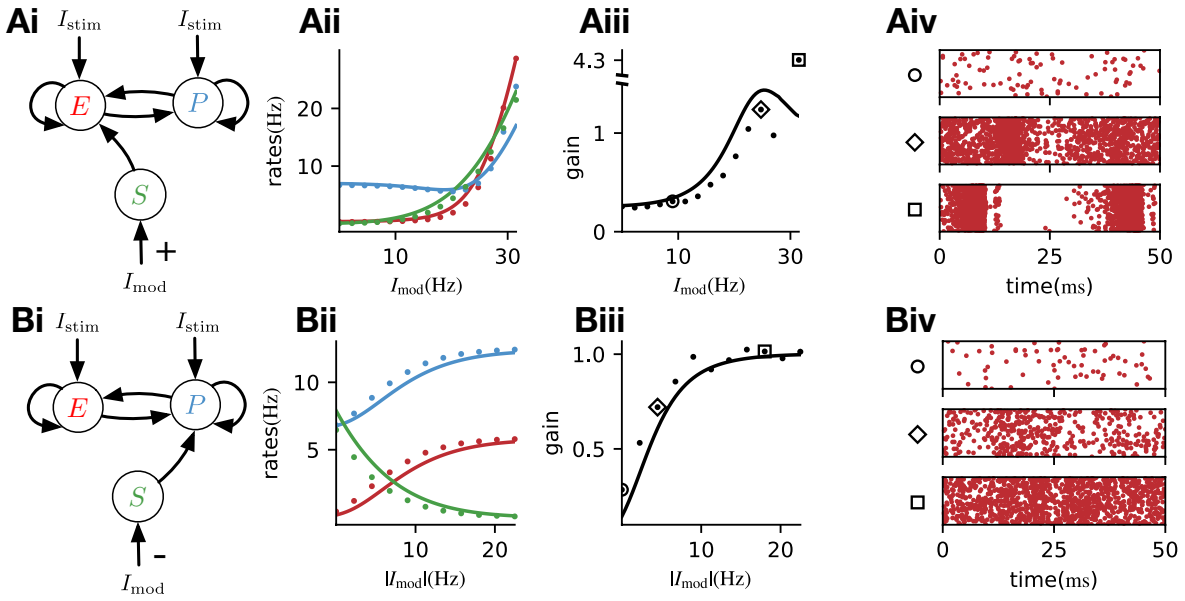

Figure 2: **Comparison of two inhibitory pathways.** Sketches of the model for disinhibition via SOM → PV → E (pathway p_{I}, **Ai**) and for inhibition via SOM → E (pathway p_{II}, **Bi**). The stimulus targets E and PV neurons and modulatory input targets SOM neurons. **A**, Disinhibition via SOM→PV→E: ii) Stationary population firing rates of excitatory neurons (red), PV neurons (blue) and SOM neurons (green) dependent on modulation of the SOM neurons. Dots correspond to simulation results of populations of leaky-integrate-and-fire neurons (Methods 1.1). Solid lines denote analytical predictions obtained from mean-field theory (Eq. (13)). iii) Gain of excitatory neurons (black, Eq. (27)) depending on the modulation of SOM neurons. Simulation (dots) and theoretical results (solid lines) in A and B are normalized separately to their maxima in B. iv) Raster plots showing all spike times of the excitatory neurons in a 50 ms segment for three levels of modulation (see markers in iii). **B** Inhibition via SOM → E: ii-iv) as in A. *Parameters*: $p_{\text{PS}} = 0.1$ (I), $p_{\text{PS}} = 0$ (II), $p_{\text{ES}} = 0$ (I), $p_{\text{ES}} = 0.1$ (II), $p_{\text{SE}} = 0$ (I, II), $p_{\text{S}}^{\text{ext,inh}}$ varies between 0.0225 and 0.00675 in (I) and between 0.01125 and 0.0225 in (II) (see also Eq. (8), $p_{\text{E}}^{\text{ext,inh}} = 0.01$ (I).

In the second pathway, p_{II}, SOM neurons project directly to E neurons, so we consider modulations that inhibit SOM neurons ($I_{\text{mod}} < 0$) and result in net disinhibition of E neurons. Consequently, E neurons increase their firing rates and then drive PV neurons to increased rates (Fig. 2Bii). Through-

out the modulation E neuron network gain increases monotonically from low to high values (Fig. 2Biii). Further, network dynamics remain in the asynchronous, irregular spiking regime (Fig. 2Biv). Indeed, this asynchrony permits the firing rate model theory to match the integrate-and-fire simulations across low and high gain values (Fig. 2Biii, compare points to the curve). Thus, in contrast to the modulation through inhibitory pathway p_{I}, pathway p_{II} supports a network gain increase that is robust throughout the modulation range.

In summary, these initial examples show that inhibitory modulations via the indirect SOM $\rightarrow$ PV $\rightarrow$ E and direct SOM $\rightarrow$ E pathways and are quite distinct. There is a clear need to establish a more nuanced view on how diverse inhibition controls network gain and stability; this is the central goal of our study. Towards this end, in the next section, we formalize the relations between gain and stability in simple cortical networks with only one inhibitory class. This will provide us a platform to build a broader theory in E – PV – SOM circuits, which will be the focus of the later sections.

Gain and stability modulation are opposed in E – PV circuits

In order to provide insight in how stability and network gain are co-affected by a modulation we first consider a reduced E – PV circuit (Fig. 3B) where stimuli target both E and PV neurons $\mathbf{I}_{\mathrm{stim}} = (I^{\mathrm{E}}_{\mathrm{stim}}, I^{\mathrm{P}}_{\mathrm{stim}})^{\mathrm{T}}$. For this circuit the network gain is (Methods 1.3):

$$g_E \equiv \frac{dr_{\mathrm{E}}}{d\mathbf{I}_{\mathrm{stim}}} = \beta_{\mathrm{E}} \frac{(1+\beta_{\mathrm{P}} J_{\mathrm{PP}}) I^{\mathrm{E}}_{\mathrm{stim}} - \beta_{\mathrm{P}} J_{\mathrm{EP}} I^{\mathrm{P}}_{\mathrm{stim}}}{\det(\mathbf{1}-\mathbf{W})}. \tag{1}$$

The first term on the right-hand side (in the numerator) describes how E neuron response is amplified by the network, while the second one describes response cancellation by PV neurons. Since the cellular gains β_{E} and β_{P} depend upon the operating point from which the circuit dynamics is linearized, the tradeoff between amplification and cancellation can be controlled through an external modulation that shifts this point.

To compare network gain across different network states we consider a grid of possible firing rates $(r_{\mathrm{E}}, r_{\mathrm{P}})$. A given network state is found by numerically determining the external input required to position the network at that rate (see Eq. (9)). For each network state we linearize the network dynamics (i.e. determine the βs) and compute the network gain via Eq. (1) (see heatmap in Fig. 3A). It is immediately apparent that network gain is largest for high r_{E} and low r_{P}, and smallest vice versa. Gain modulation is most effective when it connects two network states which are orthogonal to a line of constant gain (gain isolines in Fig. 3A). Thus, for most network states the highest gain increase occurs for modulations that increase E neuron rates while simultaneously decreasing PV neuron rates.

Unstable firing rate dynamics are typified by runaway activity, when recurrent excitation is not stabilized by recurrent inhibition (Griffith, 1963; Ozeki et al., 2009; van Vreeswijk and Sompolinsky, 1996; Wilson and Cowan, 1972). In networks of spiking neurons, this instability can manifest as network-wide synchronized, oscillatory dynamics (as observed in Fig. 2Aiv). To quantify these frequency-dependent state transitions we use a stability measure defined as the minimal distance (across oscillatory frequency) to the transition between stable and unstable oscillatory dynamics (see Methods 1.5). As was done for network gain we consider how stability depends on network activity $(r_{\mathrm{E}}, r_{\mathrm{P}})$ about which the network is linearized (color in Fig. 3C). Network dynamics are most stable for large PV and low E neuron rates (the white region in Fig. 3C). For higher E rates the recruited PV activity gives rise to larger amplitude network oscillations, and as a result the distance to instability decreases (the orange region in Fig. 3C). Finally, for sufficiently large E and small PV rates, the network transitions into fully unstable firing rate dynamics (black lines in Fig. 3A,C).

This analysis of gain and stability reveals an inverse relationship between these two network features (compare Fig. 3A and C). This is evident by plotting a near constant gain for $(r_{\mathrm{E}}, r_{\mathrm{P}})$ ranging over a stability isoline (Fig. 3D). Further, over a large region of $(r_{\mathrm{E}}, r_{\mathrm{P}})$ we observe that higher gain is accompanied by lower stability (Fig. 3A,C). This connection arises because the recurrent E $\rightarrow$ PV $\rightarrow$ E loop enforces both a dynamic cancellation required to stabilize network activity (Ozeki et al., 2009; Renart et al., 2010; Tetzlaff et al., 2012) as well as a cancellation that attenuates E neuron response to $\mathbf{I}_{\mathrm{stim}}$ (Stern et al., 2018; Sutherland et al., 2009). Thus, any modulation that increases network gain in the E – PV circuit will necessarily result in a less stable network.

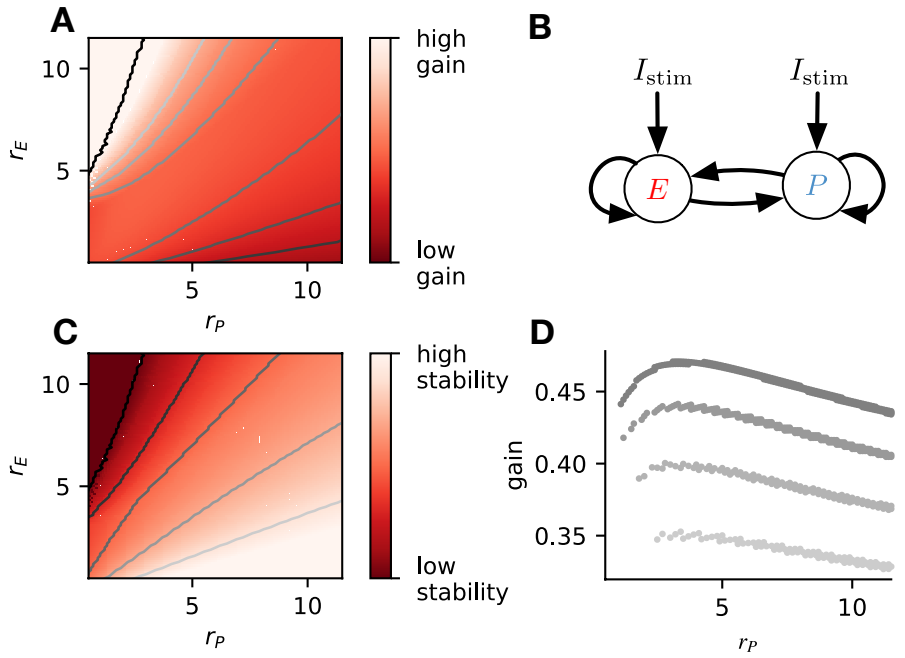

Figure 3: **Gain and stability in a network of E and PV neurons.** **A**, Every dot in the heatmap is a fixed point of the population rate dynamics of coupled E and PV neurons (Eq. (9)). Different fixed points correspond to different external inputs to the populations. The color denotes the network gain (Eq. (1)) of the system at the given fixed point. Lines of constant gain are shown in gray (from dark to light gray normalized gain 0.1 to 0.7 in steps of 0.1). The black line marks where the deterministic rate dynamics become unstable. Parameter combinations for which no fixed point could be found are set to maximum gain and minimum stability. **B**, Sketch of the model containing E and PV neurons and a stimulus that targets both populations. **C**, Same as in A but the color denotes how close the system operates to instability (Eq. (35)). The deterministic rate dynamics becomes unstable at distance (d_{min}) zero. Stability isolines are plotted in gray (from dark to light gray: $d_{\text{min}} = 0.3, 0.5, 0.7, 0.9$). **D**, Gain along stability isolines of $d_{\text{min}} = 0.6, 0.7, 0.8, 0.9$. *Parameters*: $p_{\text{ES}} = p_{\text{PS}} = p_{\text{SE}} = 0$, $p_{\text{E}}^{\text{ext}}$ and $p_{\text{P}}^{\text{ext}}$ vary on each point of the grid in A and C.

These results prompt the question: can a cortical circuit be modulated through inhibition to a higher gain regime without compromising network stability? In the next section, we investigate whether direct modulation of SOM neurons can shift the E – PV –SOM circuit from a low to a high gain state while stability is maintained.

Gain and stability are differentially mediated by SOM and PV neurons

Our initial exploration of the full E – PV – SOM circuit showed that while disinhibition through pathway p_{I} (SOM $\rightarrow$ PV $\rightarrow$ E) can modulate network gain it also destabilizes circuit activity. In contrast, disinhibition through pathway p_{II} (SOM $\rightarrow$ E) ensured a stable asynchronous state throughout modulation (Fig. 2). However, our analysis was restricted to a network with a single unmodulated state, and it remains to show that this result is general over a wide range of network states.

To simplify our analysis we neglect the E $\rightarrow$ SOM connections in the full E – PV – SOM circuit (Fig. 4, left column). Consequently, SOM neuron modulation can only affect the stability and gain of E neurons by changing the dynamical state of the E – PV subcircuit. Positive or negative input modulations to SOM neurons increase or decrease their steady-state firing rate, which in turn affects the steady-state rates of the E and PV neurons (Fig. 4Ai,ii). A specific modulation can be visualized as a vector $(\Delta r_{\text{E}}, \Delta r_{\text{P}})$ in the previously introduced $(r_{\text{E}}, r_{\text{P}})$ firing rate grid (Fig. 4Aiii). The direction of the vector indicates where the E – PV network state would move to if SOM neurons are weakly modulated. We remark that the modulation $(\Delta r_{\text{E}}, \Delta r_{\text{P}})$ not only depends on the feedforward SOM projections to E and PV neurons, but also on the dynamical regime (i.e linearization) of the

unmodulated state (r_E, r_P).

To build intuition we first consider only the SOM $\rightarrow$ PV connection and set the SOM $\rightarrow$ E connection to zero, thereby isolating pathway p_I (Fig. 2Ai). A depolarizing modulation to SOM neurons is applied ($I_{mod} > 0$), ultimately causing $\Delta r_E > 0$ through PV neuron disinhibition. If the unmodulated network state has low r_E and high r_P then the modulation vector field shows a transition from non-ISN to ISN dynamics, indicated by $\Delta r_P < 0$ for low r_E yet shifting to $\Delta r_P > 0$ for larger r_E (Fig. 4Bi,ii). In this regime, the modulation vectors cross the gain isolines, and network gain robustly increases over a wide range of initial (r_E, r_P) network states (Fig. 4Bii). However, the vectors also cross the stability isolines, showing that the modulation compromises network stability (Fig. 4Bi). If we extend our analysis by including weak SOM $\rightarrow$ E connectivity (Fig. 4Ci,ii), the SOM $\rightarrow$ PV connection continues to dominate and maintains a disinhibitory effect on E neurons. The vector field changes so that the modulation now strongly increases gain but also shifts the circuit more directly into the unstable region. Note also that the response reversal of the PV neurons at the transition from non-ISN to ISN has vanished. In total, disinhibition through pathway p_I has the general property that increased gain comes at the cost of decreased stability.

We next consider the modulation of E neuron activity via removal of direct SOM mediated inhibition through pathway p_{II}. To begin we set the SOM $\rightarrow$ PV connection to zero (Fig. 4C) and consider an external hyperpolarizing modulation of SOM neurons ($I_{mod} < 0$). This results in $\Delta r_E > 0$, which in turn gives $\Delta r_P > 0$ through the E $\rightarrow$ PV connection (Fig. 4Di,ii). The vector field crosses gain isolines at low r_E values, but aligns with a them at high r_E (Fig. 4Dii). That means that a modulation yields gain increases for low network activity but loses its effectiveness at high activity. However, the vector field also shows how the modulation aligns with the stability isolines at high r_E, indicating that stability is not comprised by modulation. Finally, including the SOM $\rightarrow$ PV connection shows that the arrow field aligns with a lower stability isoline (Fig. 4Ei). The high gain state is, therefore, more stable, but at the price of a smaller overall gain increase (Fig. 4Eii).

In sum, our analysis shows that if PV neurons stabilize E neuron activity then increasing the gain of E neurons by inhibiting PV neurons drives the network dynamics towards instability. By contrast, shifting the network to a high gain state by removing SOM inhibition from the E neurons yields rate increases of both the E and PV neurons and a stable high gain state. The network stability is improved if SOM neurons project to both E and PV neurons. In total, our analysis supports the simple idea that interneurons that stabilize network activity should not be involved in the modulation of network gain.

Recurrent feedback from E to SOM neurons amplifies gain modulation

Neglecting the E $\rightarrow$ SOM connection in the E – PV –SOM circuit makes SOM activity simply an intermediate step in a feedforward modulation of the E – PV subcircuit. In this section, we consider how the recurrent E $\leftrightarrow$ SOM interactions determine how an external modulation to SOM neurons affects E neuron gain.

First, in the absence of E $\rightarrow$ SOM coupling, we consider the network in three conditions: non-stimulated (NSt), stimulated (St), or modulated and stimulated (M+St). In the NSt condition E and PV neuron activity is low (Fig. 5Ai,ii; left bars), while SOM neurons are moderately activity (Fig. 5Aiii; left bar). In the St condition a depolarizing stimulus ($I_{stim} > 0$) is given to the E and PV neurons and their activity naturally increases (Fig. 5Ai,ii; middle bars). However, since E neurons do not project to SOM neurons then SOM activity in the S condition remains the same as in the NS condition (Fig. 5Aiii; middle bar). Finally, in the M+St condition a hyperpolarizing modulation ($I_{mod} < 0$) is applied to the SOM neurons in conjunction with $I_{stim} > 0$ to both E and PV neurons. In this case, the E and PV neurons rise their activity further compared to the St condition (Fig. 5Ai,ii; right bars) because of disinhibition via SOM neuron suppression (Fig. 5Aiii; right bar).

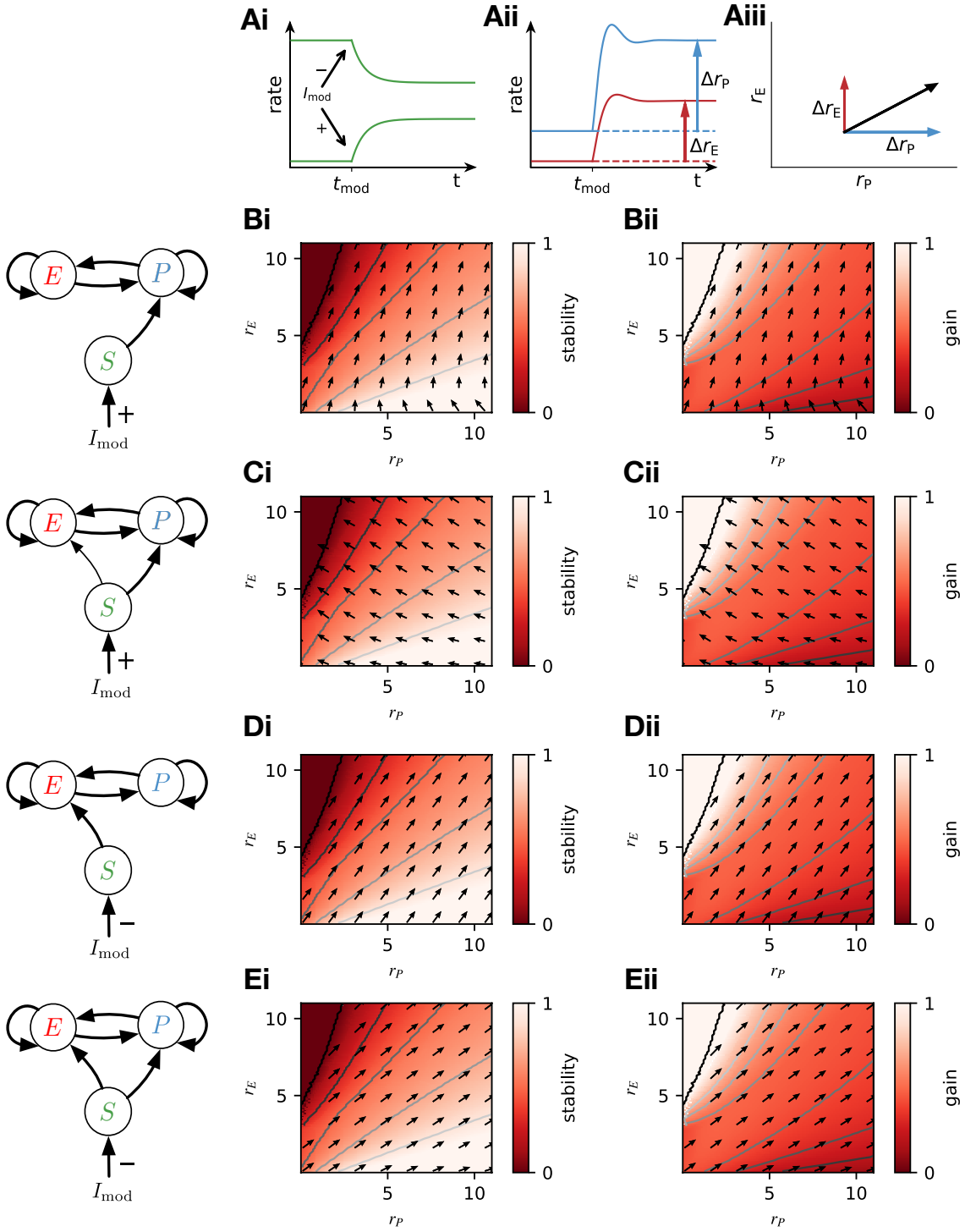

Figure 4: **How modulation of SOM neurons affect network gain and stability. A**, Sketches illustrating how the vector fields in panels B-E are obtained. Modulation of SOM neurons yields a steady-state rate change of SOM (i), which yields altered steady-state rates of E and PV neurons (ii). The arrows indicate in which direction a fixed point of the rate dynamics is changed by the modulation (iii). **B**-**E** Heatmaps as in Fig. 3 for network stability (i) and gain (ii). Arrow fields as defined in panel A. SOM neurons do not receive recurrent feedback from excitatory neurons ($p_{SE} = 0$). B: SOM neurons receive positive modulation and connect only to PV neurons ($p_{ES} = 0$, $p_{PS} = 0.1$). C: SOM neurons receive positive modulation and connect to both PV and E neurons ($p_{ES} = 0.07$, $p_{PS} = 0.1$). D: SOM neurons receive negative modulation and connect only to E neurons ($p_{ES} = 0.1$, $p_{PS} = 0$). E: SOM neurons receive negative modulation and connect to both PV and E neurons ($p_{ES} = 0.1$, $p_{PS} = 0.07$) .

The above analysis ignored E → SOM coupling; we now compare it to the case when E → SOM projections are intact. There is little difference in the NSt condition since E neuron activity is low and hence the influence of E neurons upon SOM neurons is small (Fig. 5Bi,ii,iii; left bars). The same is true for the M+St condition since the modulation continues to suppress SOM neuron activity despite the E projections to SOM (Fig. 5Bi,ii,iii; right bars). However, in the St condition the E → SOM coupling increases SOM activity (Fig. 5Biii; right bars), which in turn lowers the E response (and by extension the PV response) compared to the case when E → SOM coupling is absent (Fig. 5Bi,ii; right bars). The above analysis provides an important observation: when we compare the St and M+St cases, the relative increase in E neuron rates with SOM modulation is greater when E → SOM connections are present (Fig. 5 Ai vs Fig. 5Bi).

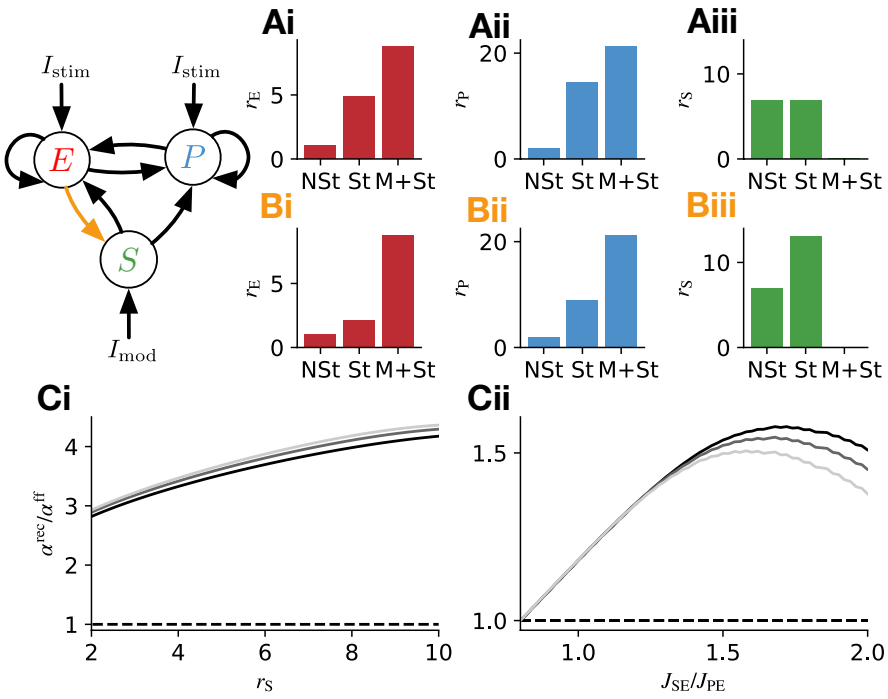

Figure 5: **Feedback from E to SOM cells enhances gain modulation. A**, Network without feedback from E to SOM neurons ($J_{\text{SE}} = 0$). Firing rates of E (i), PV (ii), and SOM (iii) neurons without stimulus (NSt), with stimulus (St) and with stimulus and (negative) modulation (M+St). **B**, Same as panel A but with feedback from E to SOM neurons. **C**, Ratio between the network gain modulation for networks with and without recurrent feedback from E to SOM neurons, $\alpha^{\text{rec}}/\alpha^{\text{ff}}$. The ratio increases with the initial firing rate of the SOM neurons (i) and the relative coupling strength between E to SOM neurons and E to PV neurons (ii). The curves correspond to three different initial conditions of the rates $(r_{\text{E}}, r_{\text{P}})$ (black: $(2\,\text{Hz}, 3\,\text{Hz})$, gray: $(2\,\text{Hz}, 2\,\text{Hz})$, light gray: $(2\,\text{Hz}, 1.5\,\text{Hz})$).

To show that this result is robust we employ our linearized rate description to give a general theory for how E – SOM recurrence increases E neuron gain control via SOM neuron modulation. To this end we consider the ratio of network gains in the modulated (m) and unmodulated (u) circuits: $\alpha = g^{\text{m}}/g^{\text{u}}$ (recall that $g = dr_{\text{E}}/dI_{\text{stim}}$). We do this for feedforward (ff) and recurrent (rec) networks which either lack or preserve E → SOM coupling, respectively. To simplify our analysis we assume that the modulation ($I_{\text{mod}} < 0$) completely silences SOM activity. With this simplification the gain modulation in the recurrent network, α^{rec}, is related to the gain modulation in the feedforward network, α^{ff}, as follows (see Eq. (45)):

$$\alpha^{\text{rec}} = \alpha^{\text{ff}}\left(1 - \beta_{\text{S}} J_{\text{SE}} \frac{dr_{\text{E}}}{dr_{\text{S}}}\right). \tag{2}$$

In circuits where SOM neurons effectively inhibit E neurons (i.e $p_{\text{II}} > p_{\text{I}}$, see Fig. 1) we always have that $dr_{\text{E}}/dr_{\text{S}} < 0$. Consequently, Eq. (2) shows that the gain modulation in the recurrent system

is then always larger than the modulation in the feedforward system. Indeed, $\alpha^{\text{rec}}/\alpha^{\text{ff}} > 1$ over a large range of both unmodulated SOM neuron activity (Fig. 5Ci) and the strength of the E $\rightarrow$ SOM connection (Fig. 5Cii). In sum, the theory which produced Eq. (2) generalizes the results of our specific network example (Fig. 5A and B).

Modulation of SOM neurons has additive and multiplicative effects on tuning curves

In the previous sections, we measured network gain as the increase of E neuron activity in response to a small increase in stimulus intensity. We now extend our gain modulation analysis to E – PV – SOM circuits with distributed responses, whereby individual neurons are tuned to a particular value of a stimulus (i.e the preferred orientation of a bar in a visual scene or the frequency of an acoustic tone). In what follows the stimulus θ is parametrized with an angle ranging from 0° to 180°.

We begin by giving all E and PV neurons feedforward input $\mu^{\text{ff}}(\theta)$ which is tuned to $\theta = 90°$ with a Gaussian profile (see Eq. (47) and Fig. 6Ai). This homogeneous input to all E and PV neurons ensures that neurons have similar tuning curves, and thus it suffices to consider the population average tuning curve, $\mathbf{r}(\theta) = [r_{\text{E}}(\theta), r_{\text{P}}(\theta), r_{\text{S}}(\theta)]$. The slope of the input tuning $d\mu^{\text{ff}}/d\theta$ (Eq. (47)) is translated to response gain $d\mathbf{r}/d\theta$ via (Eq. (49) and see Methods 1.8):

$$\frac{dr_\kappa}{d\theta} = \underbrace{\sum_{\gamma\in\{\text{E,P}\}} P_{\kappa\gamma}\beta_\gamma}_{:=g_\kappa} \frac{d\mu^{\text{ff}}}{d\theta}, \quad \text{where} \quad \mathbf{P} = (\mathbf{1} - \mathbf{W})^{-1}, \quad \kappa \in \{\text{E, P, S}\}. \tag{3}$$

Here $\mathbf{W}$ is the effective connectivity matrix for the E – PV – SOM circuit ($W_{\kappa\gamma} = \beta_\kappa J_{\kappa\gamma}$).

We have written Eq. (3) in a compact form where the mapping from input gain to response gain for population κ is given by $dr_\kappa/d\theta = g_\kappa d\mu^{\text{ff}}/d\theta$, where g_κ is the gain coefficient. The coefficient g_κ has a complicated dependence on the cellular gain β; explicitly through the products $P_{\kappa\gamma}\beta_\gamma$ as well as through the matrix $\mathbf{P}$'s dependence on the effective connectivity $\mathbf{W}$. If the neuronal transfer is truly linear, so that β is independent of the operating point, then g cannot be changed by an applied modulation. Consequently, any circuit modulation evokes only an additive (or subtractive) shift of response tuning. By similar logic, a nonlinear neuronal transfer would cause the cellular gain β to change under any modulation that shifts the network operating point. In this case, g_{E} will now also depend on the modulation (unless the circuit $\mathbf{W}$ and cellular transfer function are very finely tuned), and thus the modulation cannot result in a purely additive tuning curve shift and rather will show some multiplicative (or divisive) component.

To explore how SOM neuron modulation mediates multiplicative gain control of tuning curves we first set the unstimulated state of the model to have a low firing rate (~ 0.1 Hz) for both E and PV neurons. Applying stimuli with varying θ shows that all populations are tuned (Fig. 6Bi,ii,iii). This is expected since both E and PV neurons receive tuned feedforward input and SOM neurons receive input from E neurons which all have the same preferred angle. When the network is modulated through a suppression of SOM neuron activity ($I_{\text{mod}} < 0$) SOM tuning is both subtractively and divisively modulated (Fig. 6Biii), while the tuning of E and PV neurons undergo both additive and multiplicative changes (Fig. 6Bi,ii). The additive or subtractive shift reflects the overall rate increase of E and PV neurons and decrease of SOM neurons, respectively. Similarly, the multiplicative or divisive shift arises from an increase in the transfer coefficient g for E and PV neurons and a decrease in g_S SOM neuron suppression. Indeed, $\Delta g_\kappa = g_\kappa^{\text{m}} - g_\kappa^{\text{u}}$ is positive for all θ for the E and P populations (Fig. 6Ci,ii) and negative for all θ for the S population (Fig. 6Ciii), showing the respective multiplicative and divisive nature of the modulation on the circuit tuning.

It is currently under debate whether tuning in mouse V1 is imposed by tuned input or a combination of weakly tuned input and recurrent connections. Theoretical studies showed that balanced networks extract the weakly tuned component by subtracting the large untuned input component that is supplied to every neuron through recurrent inhibition (Hansel and van Vreeswijk, 2012; Sadeh and Rotter, 2015) and therefore produce sparse and selective responses (Pehlevan and Sompolinsky, 2014). We test whether SOM mediated gain amplification extends to a network where E and PV

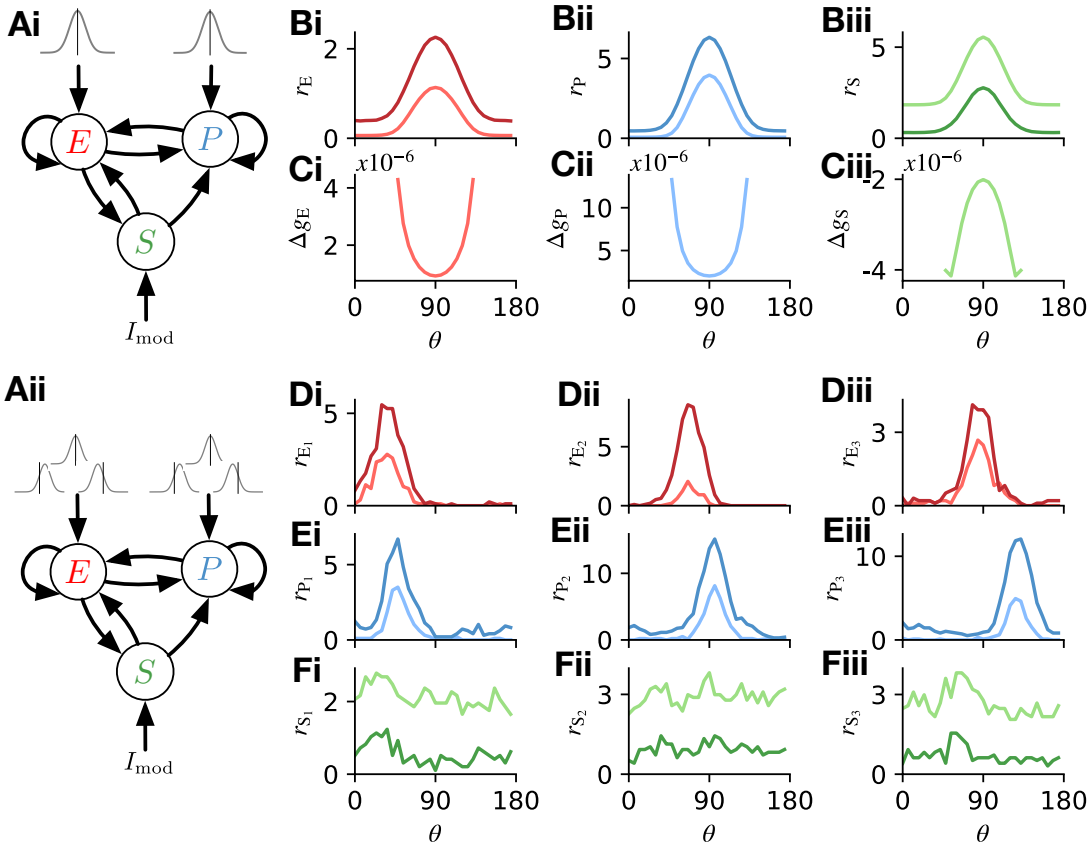

Figure 6: **SOM activity modulates tuning curves of E and PV neurons. A**, Sketch of all neurons receiving input tuned to the same (i) and different (ii) preferred θ. **B**, All E and PV neurons receive tuned feedforward input with a preferred θ of 90°. Modulatory input suppresses the activity of SOM neurons. Population tuning curves of E (i), PV (iii), and SOM neurons (iii) without (light color) and with modulation (dark color). **C**, Difference of gain coefficients $\Delta g_\kappa = g_\kappa^{\mathrm{m}} - g_\kappa^{\mathrm{u}}$ for the modulated (m) and unmodulated (m) networks. Δg_i is only shown for angles θ at which the input $\mu^{\mathrm{ff}}(\theta)$ is larger than 1% of its peak value. **D-F**, Same as in panel B, but individual neurons receive tuned feedforward input with different preferred θ. Panels show tuning curves of three example neurons (i-iii) within the E (D), PV (E), and SOM (F) populations. *Parameters:* $p_{\mathrm{E}}^{\mathrm{ext}} = 0.0185$, $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0175$, $p_{\mathrm{S}}^{\mathrm{ext}} = 0.0445$ $p_{\mathrm{S}}^{\mathrm{ext,inh}} = 0.0125$ without modulation, $p_{\mathrm{S}}^{\mathrm{ext,inh}} = 0.0175$ with modulation.

neurons have heterogeneous tuning curves. We supply E and PV neurons differently tuned input so that their θ preference is now distributed (Fig. 6D-F show the tuning curves of three representative E and PV neurons). Suppression of SOM activity yields both additive and multiplicative shifts of the tuning curves of E and PV neurons (Fig. 6D-F, solid vs faded curves). Further, since SOM neurons receive convergent input from E neurons with different preferred frequencies, then they are not tuned themselves (Fig. 6F). Thus, SOM neurons can regulate the network gain of E neurons to stimuli of different directions without having direct access to the stimulus information through tuned E input. The gain of individual E neurons is amplified once inhibition by SOM neurons that target all E neurons and suppress responses of highly active neurons is reduced.

E – PV – SOM circuitry promotes a division of labor between PV and SOM neurons.

One central conclusion from our study is that modulations of SOM neurons in the E – PV – SOM cortical circuit can segregate the mechanics of gain control and network stability. We established this by comparing and contrasting the modulatory influence of the direct SOM → E pathway (p_{II})

to that of the indirect SOM → PV → E pathway (p_{I}). However, perhaps the most salient circuit distinction between SOM and PV interneurons is the strong PV → PV connectivity in comparison to the complete lack of SOM → SOM coupling, at least as reported in the mouse sensory neocortex (Pfeffer et al., 2013; Tremblay et al., 2016; Urban-Ciecko and Barth, 2016). In this concluding section, we investigate how gain control and stability are affected by self-inhibition (or lack thereof) within the E – PV – SOM cortical circuit.

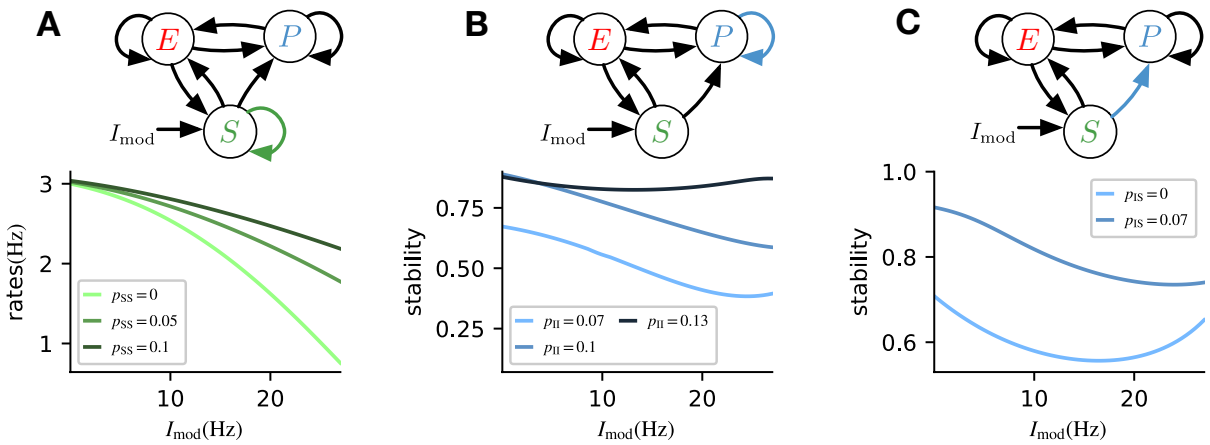

Figure 7: **How inhibitory to inhibitory coupling affects rate and stability.** **A**, Rate of E neurons depending on the modulation of SOM neurons for different values of SOM → SOM coupling. **B**, Stability of the circuit depending on the modulation of SOM neurons for different values of PV → PV coupling. The external input is chosen such that the rates at zero modulation are the same as in panel A for all values of p_{II}. **C,** Stability of the circuit depending on the modulation of SOM neurons for different values of SOM → PV coupling. The external input is chosen such that the rates at zero modulation are the same as in panel A for all values of p_{IS}. *Parameters:* $p_{\mathrm{E}}^{\mathrm{ext}} = 0.0527$, $p_{\mathrm{S}}^{\mathrm{ext,inh}} = 0.0225$, A: $p_{\mathrm{IS}} = 0.1$, $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0403$, $p_{\mathrm{S}}^{\mathrm{ext}} = 0.0408$, B: $p_{\mathrm{IS}} = 0.1$, $p_{\mathrm{S}}^{\mathrm{ext}} = 0.041$, $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0288$ ($p_{\mathrm{II}} = 0.07$), $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0403$ ($p_{\mathrm{II}} = 0.1$), $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0529$ ($p_{\mathrm{II}} = 0.13$), C: $p_{\mathrm{S}}^{\mathrm{ext}} = 0.041$, $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0374$ ($p_{\mathrm{IS}} = 0.0$), $p_{\mathrm{P}}^{\mathrm{ext}} = 0.0394$ ($p_{\mathrm{IS}} = 0.07$).

We consider the full circuit with connectivity parameters where a depolarizing SOM modulation ($I_{\mathrm{mod}} > 0$) leads to a reduction in E neuron firing rate r_{E} (i.e $p_{\mathrm{II}} > p_{\mathrm{I}}$; Fig 7A, light green curve). As a first exercise, we modify the circuit to include SOM → SOM neuron coupling ($p_{\mathrm{SS}} > 0$). A clear consequence is that modulatory recruitment of SOM activity is less effective at suppressing E neuron activity with stronger SOM self-inhibition (Fig 7A, medium and dark green curves). The intuition for this is straightforward. Recurrent inhibition is a well known form of divisive gain control (Sadeh et al., 2014; Stern et al., 2018; Sutherland et al., 2009). In our circuit, SOM neurons act as an intermediate stage in a modulatory path to E neurons. Thus, SOM → SOM coupling will ultimately reduce the efficacy (or gain) of I_{mod} to enact a change in E neuron activity. In other words, SOM self-inhibition would act to counter the modulatory influence of the I_{mod} → SOM → E pathway. This provides a functional benefit for the observed lack of SOM → SOM coupling in cortical networks. While SOM neurons lack self-inhibition (as a group), PV neurons connect strongly to other PV neurons (Pfeffer et al., 2013). In our circuit PV neurons stabilize network dynamics, counteracting the large E → E recurrence which if left unchecked would cause an explosion of network activity. In order to best quench runaway excitation, PV neurons must dynamically track E neuron activity so as to provide inhibition which effectively cancels recurrent excitation (Ozeki et al., 2009; Tsodyks et al., 1997; van Vreeswijk and Sompolinsky, 1996). This is best accomplished when PV neurons receive the same synaptic inputs as E neurons. Thus, the strong PV → E pathway necessitates a comparable PV → PV pathway. Indeed, over a large range of modulatory states network stability is increased with stronger PV → PV activity (Fig 7B). Along similar lines, the strong SOM → E pathway needed for effective network modulation should be matched by a strong SOM → PV pathway so that PV neurons are subject to the same SOM inhibition that E neurons receive. This allows PV neurons to better

stabilize overall network activity (Fig 7C).

In total, a division of labor within the PV – SOM sub-circuit, whereby SOM neurons mediate modulations of network response and PV neurons are responsible for network stability, is supported by the known synaptic interactions between the PV and SOM neurons.

Discussion

Cortical inhibition is quite diverse, with molecularly distinguished cell classes having distinct placement within the cortical circuit (Jiang et al., 2015; Markram et al., 2004; Pfeffer et al., 2013; Tremblay et al., 2016). Cell specific optogenetic perturbations are a critical probe used to relate circuit wiring to cortical function. In many cases, a preliminary analysis of these new optogenetic datasets involves building circuit intuition only from the dominant direct synaptic pathways while neglecting indirect or disynaptic pathways. This is understandable given the newly realized complexity of the circuit; however, this is precisely the situation where a more formal modeling approach can be very fruitful. Toward this end, recent modelling efforts both at the large (Billeh et al., 2020) and smaller (del Molino et al., 2017; Kuchibhotla et al., 2017; Litwin-Kumar et al., 2016; Mahrach et al., 2020; Veit et al., 2017) scales have incorporated key aspects of interneuron diversity. These studies typically explore which aspects of cellular or circuit diversity are required to replicate a specific experimental finding.

In our study, we provide a general theoretical framework that dissects the full E – PV – SOM into interacting sub-circuits. We then identify how specific inhibitory sub-circuits support both network stability and E neuron gain control; two ubiquitous functions often associated with inhibition (Ferguson and Cardin, 2020; Haider et al., 2013; Isaacson and Scanziani, 2011; Ozeki et al., 2009; Veit et al., 2017). In this way, our approach gives an expanded view of the mechanics of cortical function when compared to more classical results that focus only on how circuit structure supports a single feature of cortical dynamics. The theoretical framework we develop can be adopted to investigate other structure - function relationships in complicated multi-class cortical circuits.

Division of labor between PV and SOM interneurons

Compelling theories for both network stability (Griffith, 1963; Ozeki et al., 2009; van Vreeswijk and Sompolinsky, 1996) and gain control (Stern et al., 2018; Sutherland et al., 2009) have been developed using simple cortical models having only one inhibitory neuron class. Thus, stability and gain control do not necessarily require cortical circuits with diverse inhibition. What our study points out is that for a cortical circuit to perform gain modulation robustly through disinhibition yet remaining in a regime where the activity is stable, segregating the circuits responsible for gain control and stability has significant advantages.

If we accept this division of inhibitory labor hypothesis, then a natural question arises: Why would PV neurons be assigned to stabilization while SOM neurons assigned to E neuron modulation? In fact, there is evidence for the reverse labor assignment, namely that optogenetic perturbation of PV neurons can shift E neuron response gain (Atallah et al., 2012; Seybold et al., 2015; Wilson et al., 2012), and SOM neurons can suppress E neuron firing which in principle would also quench runaway E neuron activity (Adesnik, 2017; Adesnik et al., 2012). However, the interpretation of optogenetic perturbations can be fraught with subtle but important difficulties (Ferguson and Cardin, 2020; Phillips and Hasenstaub, 2016); in this case, the activation of PV neurons without concomitant E neuron activation is somewhat artificial. Further, while it is true that SOM neurons can mediate suppression, this is distinct from conferring stability since stable networks can allow E neurons to have high (but not runaway) activity. Our approach was not to infer a specific division of labor assignment from a synthesis of varied experimental results, yet base our conclusions from the known circuit structure of the E – PV – SOM circuit.

Two key circuit features support our division of labor breakdown. Firstly, E neurons and PV neurons experience very similar synaptic environments. Both receive excitatory drive from upstream areas (Tremblay et al., 2016), and both receive strong recurrent excitation, as well as PV- and SOM-mediated inhibition (Pfeffer et al., 2013). This symmetry in the synaptic input to E and PV neurons

allow PV neurons to dynamically track E neuron activity. Consequently, any spurious increase in excitatory drive to E neurons, that could cause a cascade of E population activity due to recurrent E → E connections, is quickly countered by an associated increase in PV inhibition. Secondly, SOM neurons famously do not connect to other SOM neurons (Jiang et al., 2015; Pfeffer et al., 2013; Urban-Ciecko et al., 2015). Since SOM neurons do provide strong inhibition to E neurons this lack of input symmetry makes them less fit to stabilize E neuron activity than PV neurons. However, it is precisely the lack of SOM neuron self inhibition which allows a high gain for any top-down modulatory signal to induce a change in E neuron response. A large component of the analysis in our manuscript is devoted to establishing this circuit based view of a division of inhibitory labor in E – PV – SOM cortical circuits.

There are circuit and cellular distinctions between PV and SOM neurons that were not considered in our study, but could nonetheless still contribute to a division of labor between network stability and modulation. Pyramidal neurons have widespread dendritic arborizations, while by comparison PV neurons have restricted dendritic trees (Markram et al., 2004). Thus, the dendritic filtering of synaptic inputs that target distal E neurons dendrites would be quite distinct from that of the same inputs onto PV neurons. Fortunately, PV neurons target both the cell bodies and proximal dendrites of both PV and E neurons (Di Cristo et al., 2004; Markram et al., 2004; Tremblay et al., 2016), so that the symmetry of PV inhibition onto PV and E neurons as viewed by action potential initiation is maintained. In stark contrast, SOM neurons inhibit the distal dendrites of E neurons (Markram et al., 2004). Dendritic inhibition has been shown to gate burst responses in pyramdial neurons greatly reducing cellular gain (Larkum et al., 2004; Mehaffey et al., 2005), and recent theoretical work shows how such gating allows for a richer, multiplexed spike train code (Naud and Sprekeler, 2018). Further, dendritic inhibition is localized near the synaptic site for E → E coupling, and recent modelling (Yang et al., 2016) and experimental (Adler et al., 2019) work shows how such dendritic inhibition can control E synapse plasticity. This implies that SOM neurons may be an important modulator not only of cortical response but also of learning.

The E – PV – SOM cortical circuit is best characterized in superficial layers of sensory neocortex (Pfeffer et al., 2013; Tremblay et al., 2016; Urban-Ciecko and Barth, 2016), and the wiring of our model network relied heavily on that literature. However, cell densities and connectivity patterns of interneuron populations change across the brain (Kim et al., 2017) and across cortical layers (Jiang et al., 2015; Tremblay et al., 2016). Our circuit based division of labor thus predicts that any differences in inhibitory connectivity compared to the one we studied will be reflected in changes of the roles that interneurons play in distinct cortical functions.

Feedback from E to SOM neurons expands range of gain modulation

Our initial analysis of the E – PV – SOM circuit neglects E → SOM cell coupling (Figs 2 and 4). While this permits a simplified analysis, a consequence is that SOM activity serves only as a feedforward input to the recurrent E – PV circuit. This restriction constrains the range over which the gain of E neurons can be varied through disinhibition of E and PV neurons. Including the E → SOM connection recruits strong SOM-mediated inhibitory feedback which suppresses the E neuron response to stimuli. Subsequent suppression of SOM neuron activity (through VIP inhibition, for instance) releases E neurons from SOM inhibition so that stimuli now drove high activity (Fig. 5). In effect, removing SOM inhibition when it controls E neuron responses through a recurrent E – SOM circuit has a far greater effect than simply removing feedforward SOM → E inhibition. This result is in line with previous findings that recurrent inhibition is more effective at gain modulation than feedforward inhibition (Sadeh et al., 2014; Stern et al., 2018; Sutherland et al., 2009).

To study SOM-mediated gain control, we simply include or remove the E → SOM coupling in the E – PV – SOM circuit. In truth, the E → SOM synaptic coupling is malleable and dependent upon E neuron activity. Short term synaptic dynamics in cortical circuits often show net depression (Zucker and Regehr, 2002), however, the E → SOM connection famously facilitates with increasing pre-synaptic activity (Beierlein et al., 2003; Reyes et al., 1998; Thomson, 1997; Tremblay et al., 2016; Urban-Ciecko and Barth, 2016; Yavorska and Wehr, 2016). Indeed, prolonged activation of E neurons recruits SOM activity through this facilitation (Beierlein et al., 2003). Thus, this enhanced

gain control would require a strong and long lasting drive to E neurons to facilitate the E $\rightarrow$ SOM synapses. Further work incorporating short term plasticity models (Tsodyks et al., 1998) into the E – PV – SOM model circuit will be required to fully explore how evoked E activity shapes the modulation of E neurons via SOM disinhibition.

Impact of SOM neuron modulation on tuning curves

Neuronal gain control has a long history of investigation (Ferguson and Cardin, 2020; Salinas and Thier, 2000; Williford and Maunsell, 2006), with mechanisms that are both bottom-up (Schwartz and Simoncelli, 2001) and top-down (Reynolds and Heeger, 2009; Ruff et al., 2018) mediated. A vast majority of early studies focused on single neuron mechanisms; examples include the role of spike frequency adaptation (Ermentrout, 1998), interactions between fluctuating synaptic conductances and spike generation mechanics (Chance et al., 2002), and dendritic-dependent burst responses (Larkum et al., 2004; Mehaffey et al., 2005). These studies often dichotomized gain modulations into a simple arithmetic where they are classified as either additive (subtractive) or multiplicative (divisive) (Silver, 2010; Williford and Maunsell, 2006). More recently, this arithmetic has been used to dissect the modulations imposed by SOM and PV neuron activity onto E neuron tuning (Atallah et al., 2012; Lee et al., 2014; Wilson et al., 2012). Initially, the studies framed a debate about how subtractive and divisive gain control should be assigned to PV and SOM neuron activation. However, a pair of studies in the auditory cortex gave a sobering account whereby activation and inactivation of PV and SOM neurons had both additive and multiplicative effects on tuning curves (Phillips and Hasenstaub, 2016; Seybold et al., 2015), challenging the tidy assignment of modulation arithmetic into interneuron class.

Past modelling efforts have specifically considered how tuned or untuned SOM and PV projections combine with nonlinear E neuron spike responses to produce subtractive or divisive gain changes (Litwin-Kumar et al., 2016; Seybold et al., 2015). However, the insights in these studies were primarily restricted to feedforward SOM and PV projections to E neurons, and ignored E neuron recurrence within the circuit. In our study we explicitly consider the role of recurrent wiring through how the effective connectivity matrix $\mathbf{W}$ will change with network state. We derived that changes in $\mathbf{W}$ scale E neuron gain in a multiplicative (or divisive) fashion (see Eq. (3) and (Sadeh et al., 2014)). By contrast, additive (or subtractive) gain changes can occur through feedforward inhibition and a roughly linear E neuron transfer (β_{E} does not change). The combination of these observations prompts a testable prediction. The large heterogeneity of subtractive and divisive gain control reported in various studies (Atallah et al., 2012; Lee et al., 2014; Natan et al., 2017; Seybold et al., 2015; Wilson et al., 2012) may not reflect differences in SOM vs PV projections to a specific E neuron, yet rather how embedded that recorded E neuron and the activated interneurons are in the full recurrent cortical circuit.

Acknowledgments

We thank Xinruo Yang, Fereshteh Lagzi and Gregory Handy for useful comments on the manuscript. Funding was provided by the National Institutes of Health Grants 1U19NS107613 (BD), CRCNS R01DC015139 (AMO, BD), and R01EB026953 (BD), the Vannevar Bush Faculty Fellowship ONR-N00014-18-1-2002 (BD, AMO), and an award from the Simons Foundation Collaboration on the Global Brain 542967 (BD).

Code

Packaged Python code to replicate simulation and theory results is freely available at `https://github.com/hannahbos/disinhibitory_pathways`

1 Methods

1.1 Population model

E, PV and SOM neurons are modeled as leaky-integrate-and-fire (LIF) neurons connected with exponentially decaying synapses

$$\tau_{\mathrm{m}} \frac{dV_i}{dt} = -(V_i - E_{\mathrm{L}}) + RI_i \tag{4}$$

$$\tau_{\mathrm{s}} \frac{dI_i}{dt} = -I_i + \tau_{\mathrm{s}} \sum_j w_{ij} \sum_n \delta(t - t_j^n). \tag{5}$$

Here V_i describes the membrane potential of the i-th neuron and I_i its synaptic current. R denotes the resistance of the membrane, E_{L} the resting potential, τ_{m} the membrane time constant, τ_{s} the synaptic time constant, w_{ij} weight of the synapse from neurons j to neuron i and t_j^n the time of the n-th spike of neuron j. When the membrane potential of a neurons exceeds its threshold (V_{th}), it is reset to its resting potential (V_{r}), where it is clipped for the refractory period (τ_{ref}). All weights are drawn from a Gaussian distribution with mean w and standard deviation $0.1w$. All parameters are given in Table 1. The dynamics are simulated with NEST (Kunkel et al., 2017).

Neurons are subdivided into N_{E} excitatory (E), N_{P} Parvalbumin-positive (PV) neurons, and N_{S} Somatostatin-positive (SOM) neurons. We assume that 20% of all neurons are inhibitory ($\eta = 1/4$) and that the ratio of PV and SST densities is given by $\rho = N_{\mathrm{S}}/N_{\mathrm{P}} = 0.83$ (Pfeffer et al., 2013), the population sizes are given by

$$N_{\mathrm{E}} = N, \quad N_{\mathrm{P}} = \frac{\eta}{1+\rho} N = \eta_{\mathrm{P}} N, \quad N_{\mathrm{S}} = \rho \eta_{\mathrm{P}} N. \tag{6}$$

Pairs of neurons are connected randomly with $p_{\kappa\gamma}$ being the probability of a connection from a neuron in population γ to a neuron in population κ ($\kappa, \gamma \in \{\mathrm{E}, , \mathrm{S}\}$). Thus the connectivity between populations is described by

$$\mathbf{J} = \tilde{w} N \begin{pmatrix} p_{\mathrm{EE}} & -g p_{\mathrm{EP}} \eta_{\mathrm{P}} & -g p_{\mathrm{ES}} \rho \eta_{\mathrm{P}} \\ p_{\mathrm{PE}} & -g p_{\mathrm{PP}} \eta_{\mathrm{P}} & -g p_{\mathrm{PS}} \rho \eta_{\mathrm{P}} \\ p_{\mathrm{SE}} & -g p_{\mathrm{SP}} \eta_{\mathrm{P}} & -g p_{\mathrm{SS}} \rho \eta_{\mathrm{P}} \end{pmatrix} = \begin{pmatrix} J_{\mathrm{EE}} & -J_{\mathrm{EP}} & -J_{\mathrm{ES}} \\ J_{\mathrm{PE}} & -J_{\mathrm{PP}} & -J_{\mathrm{PS}} \\ J_{\mathrm{SE}} & -J_{\mathrm{SP}} & -J_{\mathrm{SS}} \end{pmatrix}, \tag{7}$$

where the inhibitory connection strength is amplified by $g = 4$ and the effective weight is given by $\tilde{w} = \tau_{\mathrm{s}} R w$. Each neuron in population κ receives additional input from external Poisson sources with connection strength $\tilde{w} N p_\kappa^{\mathrm{ext}}$ and firing rate $r^{\mathrm{ext}} = 8\,\mathrm{Hz}$. In matrix representation the external connection strength is given by $\mathbf{J}^{\mathrm{ext}} = \tilde{w} N \mathrm{diag}(p_{\mathrm{E}}^{\mathrm{ext}}, p_{\mathrm{P}}^{\mathrm{ext}}, p_{\mathrm{S}}^{\mathrm{ext}})$. In some cases, populations receive input from inhibitory external Poisson sources with connection strength $-g\tilde{w} N p_\kappa^{\mathrm{ext,inh}}$ which could represent, for example, VIP neurons. External input due to stimulus or modulation is given the same connection strength. Modulatory input is given by

$$\mathbf{J}_{\mathrm{mod}} = -g \tilde{w} N \mathrm{diag}(0, 0, p_{\mathrm{S}}^{\mathrm{ext,inh}}), \tag{8}$$

such that $\mathbf{I}_{\mathrm{mod}} = p_{\mathrm{S}}^{\mathrm{ext,inh}} r^{\mathrm{ext}}$.

1.2 Theoretical description of the population dynamics

The population rate dynamics of E, PV and SOM neurons ($\mathbf{r} = (r_{\mathrm{E}}, r_{\mathrm{P}}, r_{\mathrm{S}})$) are described by a firing rate model (Wilson and Cowan, 1972)

$$\tau \frac{d\mathbf{r}}{dt} = -\mathbf{r} + f(\mathbf{r}, \mathbf{q}), \quad \text{with} \quad f(\mathbf{r}, \mathbf{q}) = \Phi(\mu(\mathbf{r}, \mathbf{q}), \sigma(\mathbf{r}, \mathbf{q})), \tag{9}$$

where $\mathbf{q}$ denotes the firing rate of external input due to a stimulus or modulation. The static transfer function Φ is given by the inverse of the mean first-passage time of the membrane potential of the

neurons. It follows from diffusion approximation of the membrane potentials of the neurons within one population, assuming that all neurons are uncorrelated (Fourcaud and Brunel, 2002)

$$\Phi(\mu,\sigma) = \Big(\tau_{\mathrm{m}}\sqrt{\pi}\int_{(\tilde{V}_r-\mu)/\sigma}^{(\tilde{V}_{th}-\mu)/\sigma} e^{s^2}(1+\mathrm{erf}(s))ds\Big)^{-1}, \tag{10}$$

with

$$\tilde{V}_r = V_r + \sigma\frac{\alpha}{2}\sqrt{\frac{\tau_{\mathrm{s}}}{\tau_{\mathrm{m}}}}, \quad \tilde{V}_{th} = V_{th} + \sigma\frac{\alpha}{2}\sqrt{\frac{\tau_{\mathrm{s}}}{\tau_{\mathrm{m}}}}, \quad \alpha = \frac{1}{\sqrt{2}}\Big|\zeta\Big(\frac{1}{2}\Big)\Big|, \tag{11}$$

with the Riemann zeta function ζ. The mean and variance of the input current to the neurons are given by

$$\mu(\mathbf{r},\mathbf{q}) = \mathbf{J}\mathbf{r} + \mathbf{J}^{\mathrm{ext}}\mathbf{q} \quad \text{and} \quad \sigma^2(\mathbf{r},\mathbf{q}) = \frac{\tilde{w}}{\sqrt{N}}\mathbf{J}^{\sigma}\mathbf{r} + \frac{\tilde{w}}{\sqrt{N}}\mathbf{J}^{\mathrm{ext}}\mathbf{q}, \tag{12}$$

with $J^{\sigma}_{i\mathrm{E}} = J_{i\mathrm{E}}$, $J^{\sigma}_{i\mathrm{P}} = gJ_{i\mathrm{P}}$ and $J^{\sigma}_{i\mathrm{S}} = gJ_{i\mathrm{S}}$.

In the steady-state the population averaged rates are given by the self-consistent equation

$$\mathbf{r}(\mathbf{q}) = f(\mathbf{r},\mathbf{q}). \tag{13}$$

Noisy rate dynamics induced by the finite-size of the network can be described be the following dynamical variable (Grytskyy, 2013):

$$\mathbf{y} = \mathbf{r} + \mathbf{x}, \quad \tau\frac{d\mathbf{r}}{dt} = -\mathbf{r} + f(\mathbf{y},\mathbf{q}), \tag{14}$$

where $\mathbf{x}$ is a vector of white noise terms x_κ with variance r_κ/N_κ.

1.3 Network gain

Changes of the steady-state rates induced by small changes in the external rate can be described by linearization around the fixed point (del Molino et al., 2017; Litwin-Kumar et al., 2016)

$$\frac{d\mathbf{r}}{d\mathbf{q}} = \frac{d\Phi}{d\boldsymbol{\mu}}\frac{d\mu(\mathbf{r},\mathbf{q})}{d\mathbf{q}} = \mathbf{B}\Big(\mathbf{J}\frac{d\mathbf{r}}{d\mathbf{q}} + \mathbf{J}^{\mathrm{ext}}\Big), \tag{15}$$

with

$$\mathbf{B} = \mathrm{diag}(\beta_{\mathrm{E}},\beta_{\mathrm{P}},\beta_{\mathrm{S}}) \quad \text{and} \quad \beta_i = d\Phi(\mu_{\mathrm{i}},\sigma_{\mathrm{i}})/d\mu_{\mathrm{i}} \tag{16}$$

yielding

$$\frac{d\mathbf{r}}{d\mathbf{q}} = \Big(\mathbf{1}-\mathbf{W}\Big)^{-1}\mathbf{W}^{\mathrm{ext}}, \quad \text{with} \quad \mathbf{W} = \mathbf{B}\mathbf{J} \quad \text{and} \quad \mathbf{W}^{\mathrm{ext}} = \mathbf{B}\mathbf{J}^{\mathrm{ext}}. \tag{17}$$

Here $\mathbf{1}$ denotes the identity matrix. Note that we omitted changes in the fixed point due to the rate dependence of the variance of the input since it is small compared to changes in the mean input. Thus the change of population rates $\delta\mathbf{r}$ induced by a change in the external rate $\delta\mathbf{q}$ is given by

$$\delta\mathbf{r} = \frac{d\mathbf{r}}{d\mathbf{q}}\delta\mathbf{q}. \tag{18}$$

and the matrix $\frac{d\mathbf{r}}{d\mathbf{q}}$ has been termed a response matrix (del Molino et al., 2017). If all eigenvalues of $\mathbf{W}$ are smaller than 1 the response matrix can be written as:

$$\frac{d\mathbf{r}}{d\mathbf{q}} = \sum_{i=0}^{\infty}\mathbf{W}^i\mathbf{W}^{\mathrm{ext}}. \tag{19}$$

and the response of the excitatory neurons to modulation of the SOM neurons gives

$$\frac{dr_{\mathrm{E}}}{dq_{\mathrm{S}}} = \beta_{\mathrm{S}} J_{\mathrm{S}}^{\mathrm{ext}} \sum_{i=0}^{\infty} W_{13}^{i} \tag{20}$$

$$= \beta_{\mathrm{S}} J_{\mathrm{S}}^{\mathrm{ext}} \left(\beta_{\mathrm{E}} J_{\mathrm{ES}} - \beta_{\mathrm{E}}^2 J_{\mathrm{EE}} J_{\mathrm{ES}} + \beta_{\mathrm{E}} \beta_{\mathrm{P}} J_{\mathrm{EP}} J_{\mathrm{PS}} \right. \tag{21}$$

$$+ (\beta_{\mathrm{P}} J_{\mathrm{EP}} J_{\mathrm{PE}} + \beta_{\mathrm{S}} J_{\mathrm{ES}} J_{\mathrm{SE}} - \beta_{\mathrm{E}} J_{\mathrm{EE}}^2) \beta_{\mathrm{E}}^2 J_{\mathrm{ES}} \tag{22}$$

$$\left. + (\beta_{\mathrm{E}} J_{\mathrm{EE}} J_{\mathrm{EP}} - \beta_{\mathrm{P}} J_{\mathrm{EP}} J_{\mathrm{PP}}) \beta_{\mathrm{E}} \beta_{\mathrm{P}} J_{\mathrm{PS}} + \ldots \right). \tag{23}$$

Here W_{13}^{i} denotes the element in the first row and third column of the matrix. Our expression shows that the response matrix describes the summed effect of all possible pathways through the network whereby an externally applied signal could influence population E rates.

Assuming that modulation only targets SOM neurons $\delta\mathbf{q} = (0, 0, I_{\mathrm{mod}})$, the rate change of excitatory neurons induced by modulation is given by

$$\delta r_{\mathrm{E}} = \frac{dr_{\mathrm{E}}}{dI_{\mathrm{mod}}} \beta_{\mathrm{S}} \tilde{I}_{\mathrm{mod}} = \frac{\beta_{\mathrm{E}} J_{\mathrm{EP}} \beta_{\mathrm{P}} J_{\mathrm{PS}} - (1 + \beta_{\mathrm{P}} J_{\mathrm{PP}}) \beta_{\mathrm{E}} J_{\mathrm{ES}}}{\det(\mathbf{1} - \mathbf{W})} \beta_{\mathrm{S}} \tilde{I}_{\mathrm{mod}} \tag{24}$$

with $\tilde{I}_{\mathrm{mod}} = J_{\mathrm{S}}^{\mathrm{ext}} I_{\mathrm{mod}}$ and the rate change of SOM neurons by

$$\delta r_{\mathrm{S}} = \frac{\det(\mathbf{1} - \mathbf{W}^{\mathrm{EP}})}{\det(\mathbf{1} - \mathbf{W})} \beta_{\mathrm{S}} \tilde{I}_{\mathrm{mod}} \quad \text{with} \quad \mathbf{W}^{\mathrm{EP}} = \begin{pmatrix} W_{\mathrm{EE}} & W_{\mathrm{EP}} \\ W_{\mathrm{PE}} & W_{\mathrm{PP}} \end{pmatrix}. \tag{25}$$

Thus the rate change of the excitatory neurons can be expressed as a function of the rate change of SOM neurons as

$$\delta r_{\mathrm{E}} = \beta_{\mathrm{E}} \frac{J_{\mathrm{EP}} \beta_{\mathrm{P}} J_{\mathrm{PS}} - (1 + \beta_{\mathrm{P}} J_{\mathrm{PP}}) J_{\mathrm{ES}}}{\det(\mathbf{1} - \mathbf{W}^{\mathrm{EP}})} \delta r_{\mathrm{S}}. \tag{26}$$

Network gain is defined as the rate change of neurons in response to a stimulus, assuming that stimuli target E and PV neurons $\delta\mathbf{q} = \mathbf{I}_{\mathrm{stim}} = (I_{\mathrm{stim}}^{\mathrm{E}}, I_{\mathrm{stim}}^{\mathrm{P}}, 0)$, E neuron network gain is given by

$$g_E = \frac{dr_{\mathrm{E}}}{d\mathbf{I}_{\mathrm{stim}}} = \beta_{\mathrm{E}} \frac{(1 + \beta_{\mathrm{P}} J_{\mathrm{PP}}) \tilde{I}_{\mathrm{stim}}^{\mathrm{E}} - \beta_{\mathrm{P}} J_{\mathrm{EP}} \tilde{I}_{\mathrm{stim}}^{\mathrm{P}}}{\det(\mathbf{1} - \mathbf{W})} \quad \text{with} \quad \tilde{I}_{\mathrm{stim}}^{\mathrm{E,P}} = J_{\mathrm{E,P}}^{\mathrm{ext}} I_{\mathrm{stim}}^{\mathrm{E,P}}. \tag{27}$$

1.4 Paradoxical responses and gain maximum

The response of the PV neurons to SOM modulation is given by

$$\frac{dr_{\mathrm{P}}}{dI_{\mathrm{mod}}} = -\beta_{\mathrm{P}} \frac{\beta_{\mathrm{E}} J_{\mathrm{ES}} J_{\mathrm{PE}} + (1 - \beta_{\mathrm{E}} J_{\mathrm{EE}}) J_{\mathrm{PS}}}{\det(\mathbf{1} - \mathbf{W})} = -\beta_{\mathrm{P}} \frac{J_{\mathrm{PS}} + \beta_{\mathrm{E}} (J_{\mathrm{ES}} J_{\mathrm{PE}} - J_{\mathrm{EE}} J_{\mathrm{PS}})}{\det(\mathbf{1} - \mathbf{W})} \tag{28}$$

When SOM neurons only project to PV neurons ($J_{\mathrm{ES}} = 0$), the rate of PV neurons decreases if the E – PV circuit is in the non-ISN regime ($\beta_{\mathrm{E}} J_{\mathrm{EE}} < 1$) and increases otherwise. The latter case has been termed paradoxical response (Tsodyks et al., 1997). If SOM neurons also project to E neurons, PV neurons get additional negative drive from the lack of E feedback yielding decreased PV rates even in the ISN regime. Hence we only expect paradoxical responses if the product of connection strength $J_{\mathrm{EE}} J_{\mathrm{PS}}$ is large. Thus the observation of paradoxical responses of PV neurons in response to suppression via SOM neurons cannot disclose whether the E neurons operate in the ISN or non-ISN regime if SOM neurons also suppress the activity of E neurons.

Modulation controls the firing rates and cellular gains (βs) of E and PV population, and network gain (g) explicitly depends on the latter. Differentiating Eq. (27) with respect to β_{E} we have:

$$\frac{dg_E}{d\beta_{\mathrm{E}}} = \frac{1 + \beta_{\mathrm{P}} J_{\mathrm{PP}}}{\det(\mathbf{1} - \mathbf{W})^2} \left(\tilde{I}_{\mathrm{stim}}^{\mathrm{E}} + \beta_{\mathrm{P}} (J_{\mathrm{PP}} \tilde{I}_{\mathrm{stim}}^{\mathrm{E}} - J_{\mathrm{EP}} \tilde{I}_{\mathrm{stim}}^{\mathrm{P}}) \right). \tag{29}$$

We see that $\frac{dg_E}{d\beta_{\mathrm{E}}}$ increases if PV $\rightarrow$ PV coupling is strong and PV neurons are not driven strongly by the stimulus.

Differentiating Eq. (27) with respect to β_{P} gives:

$$\frac{dg_E}{d\beta_{\mathrm{P}}} = \frac{\beta_{\mathrm{E}} J_{\mathrm{EP}}}{\det(\mathbf{1}-\mathbf{W})^2}\Big(- \tilde{I}^{\mathrm{E}}_{\mathrm{stim}} + \beta_{\mathrm{E}}(J_{\mathrm{EE}}\tilde{I}^{\mathrm{P}}_{\mathrm{stim}} - J_{\mathrm{PE}}\tilde{I}^{\mathrm{E}}_{\mathrm{stim}})\Big) + J_{\mathrm{SE}}A, \tag{30}$$

with

$$A = \frac{\beta_{\mathrm{E}}^2 J_{\mathrm{EP}}\beta_{\mathrm{S}}}{\det(\mathbf{1}-\mathbf{W})^2}\Big(J_{\mathrm{PS}}\tilde{I}^{\mathrm{E}}_{\mathrm{stim}} - J_{\mathrm{ES}}\tilde{I}^{\mathrm{P}}_{\mathrm{stim}}\Big). \tag{31}$$

Here we see that when β_{P} increases then g_E decreases if E $\rightarrow$ E coupling is not too strong, PV neurons are not driven strongly by the stimulus, and SOM neurons project in a feedforward manner to both E and PV neurons.

When increasing g_E via the disinhibitory pathway SOM $\rightarrow$ PV $\rightarrow$ E neurons in the ISN regime, the rate of E and PV neurons increases. Assuming that the circuit operates below the saturating regime of the transfer function, the cellular gains β_{E} and β_{P} therefore also increase. The increased E rate then increases g_E, but the increased PV rate decreases g_E, which explains the maximum at g_E. (Fig. 2Biii).

1.5 Quantifying network stability

Let us consider the linearized deterministic rate dynamics stemming from Eq. (9):

$$\tau\frac{d\mathbf{r}}{dt} = -\mathbf{r} + \mathbf{W}(\mathbf{r}+\mathbf{q}). \tag{32}$$

Stability can be inferred from the eigenvalues of the Jacobian $\mathbf{W}-\mathbf{1}$ (where $\mathbf{1}$ is the identity matrix). The steady-state rate dynamics become unstable if one of the eigenvalues of $\mathbf{W}$ has a real part that is larger than one. Going beyond this binary view of stability (stable vs unstable) we use a measure of distance to instability which can inform us whether the system becomes more or less stable with state changes. A frequency-dependent proximity to instability can be inferred from the Nyquist plot of the transfer function of the rate dynamics (Doyle et al., 2009). For simplicity, consider the dynamics of the least stable mode associated to the eigenvalue λ of $\mathbf{W}$:

$$\tau\dot{r} = -r + \lambda(r+q), \tag{33}$$

where r and q denote the rate $\mathbf{r}$ and input $\mathbf{q}$ projected onto the least stable mode, respectively. We remark that we consider q to be time-dependent with Laplace transform $Q(\omega)$. Solving the equation above in the Laplace domain gives the following relation between input and output

$$R(\omega) = \frac{\lambda(\omega)}{1-\lambda(\omega)}Q(\omega) \quad \text{with} \quad \lambda(\omega) = \frac{\lambda}{1+i\omega\tau}. \tag{34}$$

The term $1/(1-\lambda(\omega))$ is referred to as sensitivity function in control theory (Doyle et al., 2009) and its inverse measures the distance of the Nyquist curve to the critical point, which is here given by $\Re(\lambda) = 1$. The curve generated by $\lambda(\omega)$ provides a distance measure of the rate dynamics to instabilities for $\omega > 0$ (Fig. 8). Thus, it is natural to measure the proximity of the rate dynamics to instability via:

$$d_{\min} = \min_{\omega\tau} |1-\lambda(\omega)|. \tag{35}$$

Since $\lambda(\omega)$ converges to zero for large frequencies $d_{\min}$ is upper bounded by one.

The same distance measure quantifies how easily the dynamics are destabilized by internally generated noise, which is amplified by the circuit as

$$Y(\omega) = \frac{1}{1-\lambda(\omega)}X(\omega). \tag{36}$$

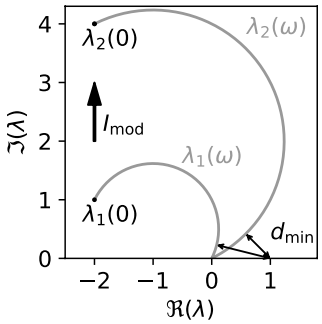

Figure 8: **Measure of distance to instability.** Sketch of Nyquist plots corresponding to the system at rest ($\lambda_1(\omega)$) and the modulated system ($\lambda_2(\omega)$). Only the mode closest to instability is shown. Modulation increases the imaginary part of the eigenvalue but leaves the real part unaltered. d_{min} is defined as the minimal distance of $\lambda(\omega)$ to the critical value $\Re(\lambda) = 1$.

Here, we simplified the dynamic transfer function as a first-order low pass filter. A more rigorous mapping of the frequency-dependent input-output relation between the rate and the spiking model can be achieved by applying linear response theory when deriving the transfer function from the Fokker-Planck equation (Brunel and Hakim, 1999; Lindner and Schimansky-Geier, 2001). For LIF-neurons in the balanced regime, this results in small modifications of the Nyquist plots from first-order low pass filters (Bos et al., 2016).

1.6 Stability of the E – PV sub-circuit

Considering the circuit composed of E and PV neurons, the eigenvalues of the Jacobian $(\mathbf{W}^{\mathrm{EP}} - 1)$ are given by

$$\lambda_{1,2} = \frac{\beta_{\mathrm{E}} J_{\mathrm{EE}} - \beta_{\mathrm{P}} J_{\mathrm{PP}} - 2}{2} \pm \sqrt{\frac{(\beta_{\mathrm{E}} J_{\mathrm{EE}} - \beta_{\mathrm{P}} J_{\mathrm{PP}} - 2)^2}{4} + \beta_{\mathrm{E}} J_{\mathrm{EP}} \beta_{\mathrm{P}} J_{\mathrm{PE}}}. \tag{37}$$

Thus the real part of the largest eigenvalue fulfills

$$\Re(\lambda_{\mathrm{max}}) \geq \frac{\beta_{\mathrm{E}} J_{\mathrm{EE}} - \beta_{\mathrm{P}} J_{\mathrm{PP}} - 2}{2}. \tag{38}$$

Hence if there is no coupling between the PV neurons $J_{\mathrm{PP}} = 0$, the activity of excitatory neurons cannot be dynamically stabilized and the susceptibility at which the dynamics become unstable is given by

$$\beta_{\mathrm{E}}^{\mathrm{critical}} = 2/J_{\mathrm{EE}}. \tag{39}$$

1.7 Gain amplification

The E neuron network gain in the unmodulated system with feedforward inhibition by SOM neurons is given by:

$$g_u^{\mathrm{ff}} \equiv \frac{dr_{\mathrm{E}}}{dI_{\mathrm{stim}}} = \frac{\beta_{\mathrm{E}}(1 + \beta_{\mathrm{P}} J_{\mathrm{PP}}) I_{\mathrm{E}} - \beta_{\mathrm{E}} \beta_{\mathrm{P}} J_{\mathrm{EP}} I_{\mathrm{P}}}{\det(\mathbf{1} - \mathbf{W})} := \frac{f(\beta_{\mathrm{E}}, \beta_{\mathrm{P}})}{h(\beta_{\mathrm{E}}, \beta_{\mathrm{P}})}. \tag{40}$$

Modulating the circuit by removing inhibition from SOM neurons shifts E and PV neurons to a new fixed point with cellular gains β_{E}^* and β_{P}^*, respectively. Similarly, the network gain in the modulated state is:

$$g_{\mathrm{m}}^{\mathrm{ff}} = \frac{f(\beta_{\mathrm{E}}^*, \beta_{\mathrm{P}}^*)}{h(\beta_{\mathrm{E}}^*, \beta_{\mathrm{P}}^*)}. \tag{41}$$

Modulation then amplifies gain by the factor $\alpha^{\mathrm{ff}} = g_{\mathrm{m}}^{\mathrm{ff}} / g_u^{\mathrm{ff}}$.

Starting from the same fixed point as in the feedforward system, the gain in the unmodulated circuit with recurrently connected SOM neurons is then:

$$g_{\mathrm{u}}^{\mathrm{rec}} = \frac{f(\beta_{\mathrm{E}}, \beta_{\mathrm{P}})}{h(\beta_{\mathrm{E}}, \beta_{\mathrm{P}}) - u(\beta_{\mathrm{E}}, \beta_{\mathrm{P}}, \beta_{\mathrm{S}})}, \tag{42}$$

with

$$u(\beta_{\mathrm{E}}, \beta_{\mathrm{P}}, \beta_{\mathrm{S}}) = (\beta_{\mathrm{E}}\beta_{\mathrm{P}} J_{\mathrm{EP}} J_{\mathrm{PS}} - \beta_{\mathrm{E}}(1 + \beta_{\mathrm{P}} J_{\mathrm{PP}}) J_{\mathrm{ES}})\beta_{\mathrm{S}} J_{\mathrm{SE}} = h(\beta_{\mathrm{E}}, \beta_{\mathrm{P}})\beta_{\mathrm{S}} J_{\mathrm{SE}} \frac{dr_{\mathrm{E}}}{dr_{\mathrm{S}}} \tag{43}$$

and $dr_{\mathrm{E}}/dr_{\mathrm{S}}$ for the corresponding feedforward circuit. Assuming that modulation silences SOM neurons, the circuit is shifted to the same fixed point as in the feedforward case and therefore

$$g_{\mathrm{m}}^{\mathrm{rec}} = \frac{f(\beta_{\mathrm{E}}^*, \beta_{\mathrm{P}}^*)}{h(\beta_{\mathrm{E}}^*, \beta_{\mathrm{P}}^*)} = g_{\mathrm{m}}^{\mathrm{ff}}. \tag{44}$$

The gain amplification in the circuit with recurrently connected SOM neurons is hence given by

$$\alpha^{\mathrm{rec}} = \frac{g_{\mathrm{m}}^{\mathrm{rec}}}{g_{u}^{\mathrm{rec}}} = \alpha^{\mathrm{ff}}\Big(1 - \beta_{\mathrm{S}} J_{\mathrm{SE}} \frac{dr_{\mathrm{E}}}{dr_{\mathrm{S}}}\Big). \tag{45}$$

Since $dr_{\mathrm{E}}/dr_{\mathrm{S}}$ is always negative for the circuit that supports gain modulation by removing SOM activity, the gain amplification in the recurrent system is always larger than in the feedforward system.

1.8 Gain modulation in tuned populations

We first assume that the E, PV and SOM population each has a stable tuning curve with respect to some stimuli $\theta \in (0°, 180°)$; we denote the population tuning curve as $\mathbf{r}(\theta) = (r_{\mathrm{E}}(\theta), r_{\mathrm{P}}(\theta), r_{\mathrm{S}}(\theta))$. In equilibrium (i.e neglecting transients) the change of the tuning curves with respect to the angle is given by

$$\frac{d\mathbf{r}}{d\theta} = \frac{d}{d\theta} f(\boldsymbol{\mu}^{\mathrm{rec}}(\theta), \boldsymbol{\mu}^{\mathrm{ff}}(\theta)), \tag{46}$$

with the static transfer function f (Eq. (9) and Eq. (10)) and the mean input from recurrent and feedforward connections $\boldsymbol{\mu}^{\mathrm{rec}}$ and $\boldsymbol{\mu}^{\mathrm{ff}}$ (Eq. (12)). We assume that the feedforward input is tuned with a Gaussian profile and that it only targets E and PV neurons:

$$\boldsymbol{\mu}^{\mathrm{ff}}(\theta) = \tilde{w} N p_\theta e^{-(\theta - \theta^p)^2/\sigma_\theta^2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \tag{47}$$

with the preferred angle $\theta^p = 90°$. Analyzing the sensitivity of firing rates to changes in preferred orientation (analogously to changes in the input intensity as analyzed in Methods 1.3) gives the stimulus gain mapping from feedforward input to firing rate output:

$$\frac{d\mathbf{r}}{d\theta} = (\mathbf{1} - \mathbf{W})^{-1}\tilde{\mathbf{I}}(\theta), \quad \tilde{I}_\kappa = \beta_\kappa \frac{d\mu_\kappa^{\mathrm{ff}}}{d\theta}, \tag{48}$$

where $\kappa \in \{\mathrm{E}, \mathrm{P}, \mathrm{S}\}$. Since E and PV receive the same amount of feedforward input we can drop the index of μ_κ^{ff} and write the slope of the population tuning curves as

$$\frac{dr_\kappa}{d\theta} = \underbrace{\sum_{\gamma \in \{\mathrm{E},\mathrm{P}\}} P_{\kappa\gamma}\beta_\gamma}_{:= g_\kappa} \frac{d\mu^{\mathrm{ff}}}{d\theta}, \quad \text{where} \quad \mathbf{P} = (\mathbf{1} - \mathbf{W})^{-1}, \quad \alpha \in \{\mathrm{E}, \mathrm{P}, \mathrm{S}\}, \tag{49}$$

showing that the slope of the input tuning curve of population κ is effectively multiplied by the factor g_κ.

We can extend the analysis to networks where individual E and PV neurons receive differently tuned input (i.e distributed tuning). In that case the response sensitivity $d\mathbf{r}/d\theta$ and the mean inputs $\boldsymbol{\mu}^{\text{rec}}$, $\boldsymbol{\mu}^{\text{ff}}$ are of dimension N (the number of neurons). The input to the i^{th} neuron of population κ is given by:

$$\mu^{\text{ff}}_{i\kappa}(\theta) = \tilde{w}Np_\theta e^{-(\theta-\theta^p_{i\kappa})^2/\sigma^2_\theta}, \tag{50}$$

with the preferred angle $\theta^p_{i\kappa} = 180i/N_\kappa$. Thus the input vector for each θ is sparse and the response of the balanced network is also sparse and selective (Pehlevan and Sompolinsky, 2014). The amplification of sparse input vectors can be understood as the projection of the input onto the eigenvectors of the $N \times N$ dimensional Jacobian $(\mathbf{1} - \mathbf{W}^{N\times N})$ weighted by the inverse of the associated eigenvalues (Sadeh et al., 2014). Since some eigenvalues of the random connectivity matrix $\mathbf{W}^{N\times N}$ are close to one, their modes get amplified. When the activity of SOM neurons is removed the product of the population eigenvalues ($\det(\mathbf{1} - \mathbf{W})$) becomes smaller which subsequently increases gain. Since the eigenvectors of population eigenvalues affect all neurons in one population equally, the tuning curve slopes of individual neurons is increased (Fig. 6D-F).

In particular, if SOM neurons only project feedforward to the E and PV neurons, the population response scaling ($\det(\mathbf{1} - \mathbf{W})^{-1} = \det(\mathbf{1} - \mathbf{W}^{\text{EP}})^{-1}$) is determined by the changes of cellular gains of the E and PV neurons alone. In contrast, if SOM neurons are part of the recurrent circuit ($J_{\text{SE}} \neq 0$), removing their activity directly affects the population response scaling ($\det(\mathbf{1} - \mathbf{W})^{-1} = (\det(\mathbf{1} - \mathbf{W}^{\text{EP}}) - \beta_{\text{E}}(p_{\text{I}} - p_{\text{II}})\beta_{\text{S}}J_{\text{SE}})^{-1}$) resulting in stronger multiplicative scaling.

Parameter	Value	Description
τ_{m}	10 ms	membrane time constant
R	40 MΩ	membrane resistance
E_{L}	-65 mV	resting potential
V_{th}	-50 mV	threshold potential
V_{r}	-65 mV	reset potential
τ_{s}	0.5 ms	synaptic time constant
τ_{ref}	2 ms	absolute refractory period
N	4136	number of excitatory neurons
p_{EE}	0.03	connection probability between excitatory neurons
p_{PE}	0.05	connection probability between E and PV neurons
p_{SE}	0.05	connection probability between E and SOM neurons
p_{EP}	0.1	connection probability between PV and E neurons
p_{PP}	0.1	connection probability between PV neurons
p_{SP}	0.0	connection probability between PV and SOM neurons
p_{ES}	0.1	connection probability between SOM and E neurons
p_{PS}	0.07	connection probability between SOM and PV neurons
p_{SS}	0.0	connection probability between SOM neurons
$p_{\mathrm{E}}^{\mathrm{ext}}$	0.055	connection probability of external input to E neurons
$p_{\mathrm{P}}^{\mathrm{ext}}$	0.05	connection probability of external input to PV neurons
$p_{\mathrm{S}}^{\mathrm{ext}}$	0.05	connection probability of inhibitory external input to SOM neurons
$p_{\mathrm{E}}^{\mathrm{ext,inh}}$	0	connection probability of inhibitory external input to E neurons
$p_{\mathrm{P}}^{\mathrm{ext,inh}}$	0	connection probability of inhibitory external input to PV neurons
$p_{\mathrm{S}}^{\mathrm{ext,inh}}$	0.025	connection probability of external inhibitory input to SOM neurons
$w = w_{i\mathrm{E}}$	610.56 pA	synaptic strength of excitatory connection
$w_{i\mathrm{P}}$, $w_{i\mathrm{S}}$	$-gw$	synaptic strength of inhibitory connection
p_θ	0.025	connection probability of tuned input to E neurons
σ_θ	20°	standard deviation of input tuning

Table 1: **Default model parameter**. Deviations are specified in the text or caption of figures.

References

Adesnik, H. (2017). Synaptic Mechanisms of Feature Coding in the Visual Cortex of Awake Mice. Neuron *95*, 1147–1159.e4.

Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z.J., and Scanziani, M. (2012). A neural circuit for spatial summation in visual cortex. Nature *490*, 226–231.

Adler, A., Zhao, R., Shin, M.E., Yasuda, R., and Gan, W.B. (2019). Somatostatin-expressing interneurons enable and maintain learning-dependent sequential activation of pyramidal neurons. Neuron *102*, 202–216.

Atallah, B.V., Bruns, W., Carandini, M., and Scanziani, M. (2012). Parvalbumin-expressing interneurons linearly transform cortical responses to visual stimuli. Neuron *73*, 159.

Atallah, B.V. and Scanziani, M. (2009). Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. Neuron *62*, 566–577.

Beierlein, M., Gibson, J.R., and Connors, B.W. (2003). Two dynamically distinct inhibitory networks in layer 4 of the neocortex. Journal of neurophysiology *90*, 2987–3000.

Berman, N.J. and Maler, L. (1998). Inhibition evoked from primary afferents in the electrosensory lateral line lobe of the weakly electric fish (apteronotus leptorhynchus). Journal of Neurophysiology *80*, 3173–3196.

Billeh, Y.N., et al. (2020). Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. Neuron .

Bos, H., Diesmann, M., and Helias, M. (2016). Identifying Anatomical Origins of Coexisting Oscillations in the Cortical Microcircuit. PLoS computational biology *12*, e1005132–34.

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. Journal of computational neuroscience *8*, 183–208.

Brunel, N. and Hakim, V. (1999). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. Neural computation *11*, 1621–1671.

Carandini, M. and Heeger, D.J. (2012). Normalization as a canonical neural computation. Nature Reviews Neuroscience *13*, 51.

Cardin, J.A. (2018). Inhibitory interneurons regulate temporal precision and correlations in cortical circuits. Trends in neurosciences *41*, 689–700.

Chance, F.S., Abbott, L.F., and Reyes, A.D. (2002). Gain modulation from background synaptic input. Neuron *35*, 773–782.

Cohen, M.R. and Maunsell, J.H. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nature neuroscience *12*, 1594.

del Molino, L.C.G., Yang, G.R., Mejias, J.F., and Wang, X.J. (2017). Paradoxical response reversal of top-down modulation in cortical circuits with three interneuron types. Elife *6*, e29742.

Di Cristo, G., et al. (2004). Subcellular domain-restricted gabaergic innervation in primary visual cortex in the absence of sensory and thalamic inputs. Nature neuroscience *7*, 1184–1186.

Doiron, B., Litwin-Kumar, A., Rosenbaum, R., Ocker, G.K., and Josić, K. (2016). The mechanics of state-dependent neural correlations. Nature neuroscience *19*, 383–393.

Douglas, R.J., Martin, K.A., and Whitteridge, D. (1989). A canonical microcircuit for neocortex. Neural computation *1*, 480–488.

Downer, J.D., Niwa, M., and Sutter, M.L. (2015). Task engagement selectively modulates neural correlations in primary auditory cortex. Journal of Neuroscience *35*, 7565–7574.

Doyle, J.C., Francis, B.A., and Tannenbaum, A.R., *Feedback Control Theory* (Dover Publications, 2009).

Eccles, J.C., Fatt, P., and Koketsu, K. (1954). Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones. The Journal of physiology *126*, 524–562.

Ermentrout, B. (1998). Linearization of F-I curves by adaptation. Neural computation *10*, 1721–1729.

Fenno, L., Yizhar, O., and Deisseroth, K. (2011). The development and application of optogenetics. Annual review of neuroscience *34*, 389–412.

Ferguson, K.A. and Cardin, J.A. (2020). Mechanisms underlying gain modulation in the cortex. Nature Reviews Neuroscience 1–13.

Fourcaud, N. and Brunel, N. (2002). Dynamics of the firing probability of noisy integrate-and-fire neurons. Neural computation *14*, 2057–2110.

Fu, Y., et al. (2014). A Cortical Circuit for Gain Control by Behavioral State. Cell *156*, 1139–1152.

Griffith, J. (1963). On the stability of brain-like structures. Biophysical journal *3*, 299–308.

Grytskyy, D. (2013). A unified view on weakly correlated recurrent networks 1–19.

Haider, B., Häusser, M., and Carandini, M. (2013). Inhibition dominates sensory responses in the awake cortex. Nature *493*, 97.

Hansel, D. and van Vreeswijk, C. (2012). The mechanism of orientation selectivity in primary visual cortex without a functional map. The Journal of neuroscience *32*, 4049–4064.

Harris, K. and Thiele, A. (2011). Cortical state and attention. Nature reviews neuroscience *12*, 509–523.

Hartline, H.K., Wagner, H.G., and Ratliff, F. (1956). Inhibition in the eye of limulus. The Journal of general physiology *39*, 651–673.

Hattori, R., Kuchibhotla, K.V., Froemke, R.C., and Komiyama, T. (2017). Functions and dysfunctions of neocortical inhibitory neuron subtypes. Nature neuroscience *20*, 1199.

Huang, C., et al. (2019). Circuit models of low-dimensional shared variability in cortical networks. Neuron *101*, 337–348.

Isaacson, J.S. and Scanziani, M. (2011). How inhibition shapes cortical activity. Neuron *72*, 231–243.

Jiang, X., et al. (2015). Principles of connectivity among morphologically defined cell types in adult neocortex. Science *350*, aac9462.

Kato, H.K., Asinof, S.K., and Isaacson, J.S. (2017). Network-level control of frequency tuning in auditory cortex. Neuron *95*, 412–423.

Katzner, S., Busse, L., and Carandini, M. (2011). Gabaa inhibition controls response gain in visual cortex. Journal of Neuroscience *31*, 5931–5941.

Kepecs, A. and Fishell, G. (2014). Interneuron cell types are fit to function. Nature *505*, 318.

Kim, Y., et al. (2017). Brain-wide Maps Reveal Stereotyped Cell-Type-Based Cortical Architecture and Subcortical Sexual Dimorphism. Cell *171*, 456–469.e22.

Kuchibhotla, K.V., et al. (2017). Parallel processing by cortical inhibition enables context-dependent behavior. Nature neuroscience *20*, 62–71.

Kunkel, S., et al. (2017). Nest 2.12.0.

Large, A.M., et al. (2018). Differential inhibition of pyramidal cells and inhibitory interneurons along the rostrocaudal axis of anterior piriform cortex. Proceedings of the National Academy of Sciences *115*, E8067–E8076.

Larkum, M.E., Senn, W., and Lüscher, H.R. (2004). Top-down dendritic input increases the gain of layer 5 pyramidal neurons. Cerebral cortex *14*, 1059–1070.

Lee, S., Kruglikov, I., Huang, Z.J., Fishell, G., and Rudy, B. (2013). A disinhibitory circuit mediates motor integration in the somatosensory cortex. Nature Publishing Group *16*, 1662–1670.

Lee, S.H., Kwan, A.C., and Dan, Y. (2014). Interneuron subtypes and orientation tuning. Nature *508*, E1–E2.

Lindner, B. and Schimansky-Geier, L. (2001). Transmission of noise coded versus additive signals through a neuronal ensemble. Physical Review Letters *86*, 2934–2937.

Litwin-Kumar, A., Rosenbaum, R., and Doiron, B. (2016). Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. Journal of Neurophysiology *115*, 1399–1409.

Lloyd, D.P. (1946). Facilitation and inhibition of spinal motoneurons. Journal of Neurophysiology *9*, 421–438.

Mahrach, A., Chen, G., Li, N., van Vreeswijk, C., and Hansel, D. (2020). Mechanisms underlying the response of mouse cortical networks to optogenetic manipulation. Elife *9*, e49967.

Markram, H., et al. (2004). Interneurons of the neocortical inhibitory system. Nature reviews neuroscience *5*, 793.

McGinley, M.J., et al. (2015). Waking state: rapid variations modulate neural and behavioral responses. Neuron *87*, 1143–1161.

Mehaffey, W.H., Doiron, B., Maler, L., and Turner, R.W. (2005). Deterministic multiplicative gain control with active dendrites. Journal of Neuroscience *25*, 9968–9977.

Natan, R.G., Rao, W., and Geffen, M.N. (2017). Cortical Interneurons Differentially Shape Frequency Tuning following Adaptation. Cell reports *21*, 878–890.

Naud, R. and Sprekeler, H. (2018). Sparse bursts optimize information transmission in a multiplexed neural code. Proceedings of the National Academy of Sciences *115*, E6329–E6338.

Niell, C.M. and Stryker, M.P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. Neuron *65*, 472–479.

Okun, M. and Lampl, I. (2008). Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. Nature neuroscience *11*, 535.

Ozeki, H., Finn, I.M., Schaffer, E.S., Miller, K.D., and Ferster, D. (2009). Inhibitory Stabilization of the Cortical Network Underlies Visual Surround Suppression. Neuron *62*, 578–592.

Pehlevan, C. and Sompolinsky, H. (2014). Selectivity and sparseness in randomly connected balanced networks. PloS one *9*, e89992.

Pfeffer, C.K., Xue, M., He, M., Huang, Z.J., and Scanziani, M. (2013). Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. Nature neuroscience *16*, 1068–1076.

Phillips, E.A. and Hasenstaub, A.R. (2016). Asymmetric effects of activating and inactivating cortical interneurons. eLife *5*, e18383.

Pi, H.J., et al. (2013). Cortical interneurons that specialize in disinhibitory control. Nature *503*, 521.

Poulet, J.F. and Petersen, C.C. (2008). Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. Nature *454*, 881.

Renart, A., et al. (2010). The asynchronous state in cortical circuits. Science *327*, 587–590.

Reyes, A., et al. (1998). Target-cell-specific facilitation and depression in neocortical circuits. Nature neuroscience *1*, 279–285.

Reynolds, J.H. and Heeger, D.J. (2009). The normalization model of attention. Neuron *61*, 168–185.

Rubin, D.B., Van Hooser, S.D., and Miller, K.D. (2015). The Stabilized Supralinear Network: A Unifying Circuit Motif Underlying Multi-Input Integration in Sensory Cortex. Neuron *85*, 402–417.

Ruff, D.A., Ni, A.M., and Cohen, M.R. (2018). Cognition as a window into neuronal population space. Annual review of neuroscience *41*, 77–97.

Sadeh, S., Cardanobile, S., and Rotter, S. (2014). Mean-field analysis of orientation selectivity in inhibition-dominated networks of spiking neurons. SpringerPlus *3*, 148.

Sadeh, S. and Rotter, S. (2015). Orientation selectivity in inhibition-dominated networks of spiking neurons: effect of single neuron properties and network dynamics. PLoS computational biology *11*, e1004045.

Salinas, E. and Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. Neuron *27*, 15–21.

Schwartz, O. and Simoncelli, E.P. (2001). Natural signal statistics and sensory gain control. Nature neuroscience *4*, 819–825.

Seybold, B.A., Phillips, E.A.K., Schreiner, C.E., and Hasenstaub, A.R. (2015). Inhibitory Actions Unified by Network Integration. Neuron *87*, 1181–1192.

Silver, R.A. (2010). Neuronal arithmetic. Nature Reviews Neuroscience *11*, 474.

Stern, M., Bolding, K.A., Abbott, L.F., and Franks, K.M. (2018). A transformation from temporal to ensemble coding in a model of piriform cortex. eLife *7*, e34831.

Sutherland, C., Doiron, B., and Longtin, A. (2009). Feedback-induced gain control in stochastic spiking networks. Biological cybernetics *100*, 475–489.

Tetzlaff, T., Helias, M., Einevoll, G.T., and Diesmann, M. (2012). Decorrelation of Neural-Network Activity by Inhibitory Feedback. PLoS computational biology *8*, e1002596–29.

Thomson, A.M. (1997). Activity-dependent properties of synaptic transmission at two classes of connections made by rat neocortical pyramidal axons in vitro. The Journal of Physiology *502*, 131–147.

Tremblay, R., Lee, S., and Rudy, B. (2016). GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits. Neuron *91*, 260–292.

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. Neural computation *10*, 821–835.

Tsodyks, M.V., Skaggs, W.E., Sejnowski, T.J., and McNaughton, B.L. (1997). Paradoxical effects of external modulation of inhibitory interneurons. The Journal of neuroscience *17*, 4382–4388.

Urban-Ciecko, J. and Barth, A.L. (2016). Somatostatin-expressing neurons in cortical networks. Nature Publishing Group *17*, 401–409.

Urban-Ciecko, J., Fanselow, E.E., and Barth, A.L. (2015). Neocortical Somatostatin Neurons Reversibly Silence Excitatory Transmission via GABAb Receptors. Current Biology *25*, 1–11.

van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science *274*, 1724–1726.

Veit, J., Hakim, R., Jadi, M.P., Sejnowski, T.J., and Adesnik, H. (2017). Cortical gamma band synchronization through somatostatin interneurons. Nature neuroscience *20*, 951.

Vinje, W.E. and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. Science *287*, 1273–1276.

Wang, X.J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. Physiological reviews *90*, 1195–1268.

Wang, X.J., Tegnér, J., Constantinidis, C., and Goldman-Rakic, P. (2004). Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory. Proceedings of the National Academy of Sciences *101*, 1368–1373.

Wang, X.J. and Yang, G.R. (2018). A disinhibitory circuit motif and flexible information routing in the brain. Current opinion in neurobiology *49*, 75–83.

Wehr, M. and Zador, A.M. (2003). Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature *426*, 442.

Williford, T. and Maunsell, J.H.R. (2006). Effects of spatial attention on contrast response functions in macaque area V4. Journal of Neurophysiology *96*, 40–54.

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

Wilson, N.R., Runyan, C.A., Wang, F.L., and Sur, M. (2012). Division and subtraction by distinct cortical inhibitory networks in vivo. Nature *488*, 343.

Womelsdorf, T., Valiante, T.A., Sahin, N.T., Miller, K.J., and Tiesinga, P. (2014). Dynamic circuit motifs underlying rhythmic gain control, gating and integration. Nature neuroscience *17*, 1031.

Wood, K.C., Blackwell, J.M., and Geffen, M.N. (2017). Cortical inhibitory interneurons control sensory processing. Current opinion in neurobiology *46*, 200–207.

Xu, H., Jeong, H.Y., Tremblay, R., and Rudy, B. (2013). Neocortical Somatostatin-Expressing GABAergic Interneurons Disinhibit the Thalamorecipient Layer 4. Neuron *77*, 155–167.

Yang, G.R., Murray, J.D., and Wang, X.J. (2016). A dendritic disinhibitory circuit mechanism for pathway-specific gating. Nature communications *7*, 1–14.

Yavorska, I. and Wehr, M. (2016). Somatostatin-Expressing Inhibitory Interneurons in Cortical Circuits. Frontiers in Neural Circuits *10*, 226–18.

Zucker, R.S. and Regehr, W.G. (2002). Short-term synaptic plasticity. Annual review of physiology *64*, 355–405.