Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity

Felix Waitzmann[1,2,*], Yue Kris Wu[1,2,*,#], Julijana Gjorgjieva[1,2,#]

[1]School of Life Sciences, Technical University of Munich, Freising, Germany;
[2]Max Planck Institute for Brain Research, Frankfurt, Germany
[*] These authors contributed equally
[#] Correspondence: kris.wu@tum.de, gjorgjieva@tum.de

Abstract

Cortical dynamics and computations are strongly influenced by diverse GABAergic interneurons, including those expressing parvalbumin (PV), somatostatin (SST), and vasoactive intestinal peptide (VIP). Together with excitatory (E) neurons, they form a canonical microcircuit and exhibit counterintuitive nonlinear phenomena. One instance of such phenomena is response reversal, whereby SST neurons show opposite responses to top-down modulation via VIP depending on the presence of bottom-up sensory input, indicating that the network may function in different regimes under different stimulation conditions. Combining analytical and computational approaches, we demonstrate that model networks with multiple interneuron subtypes and experimentally identified inhibitory short-term plasticity mechanisms can implement response reversal. Surprisingly, despite not directly affecting SST and VIP activity, PV-to-E short-term depression has a decisive impact on SST response reversal. We unveil the link between response reversal and two dynamical signatures of the network operating regime, inhibition stabilization and the paradoxical effect. We find that response reversal coincides with a change in the interneuron subtype responsible for network stabilization. In summary, our work suggests a role of short-term plasticity mechanisms in generating nonlinear phenomena in networks with multiple interneuron subtypes and makes several experimentally testable predictions.

Introduction

Inhibitory neurons in the cortex are highly diverse in anatomy, electrophysiology, and function (Pfeffer et al., 2013; Kepecs and Fishell, 2014; Jiang et al., 2015; Tremblay et al., 2016). In the mouse cortex, three major classes of interneurons expressing parvalbumin (PV), somatostatin (SST), and vasoactive intestinal peptide (VIP) make up more than 80% of GABAergic interneurons (Tremblay et al., 2016). Together with excitatory (E) neurons, they form a canonical microcircuit relevant for various cortical computations, including locomotion-induced gain modulation (Fu et al., 2014),

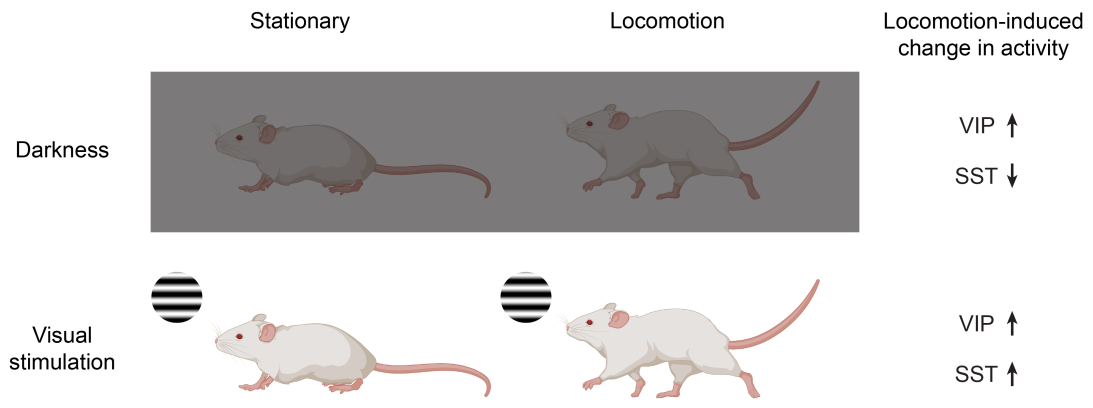

Fig. 1. Schematic diagrams illustrating that under different stimulus conditions, locomotion-induced modulatory input via VIP affects SST response oppositely. Top: in darkness, locomotion-induced top-down modulation increases VIP activity but decreases SST activity (Fu et al., 2014). Bottom: in the presence of visual stimulus, locomotion-induced top-down modulation increases both VIP and SST activity (Pakan et al., 2016; Dipoppa et al., 2018).

selective attention (Zhang et al., 2014), context-dependent modulation (Kuchibhotla et al., 2017; Keller et al., 2020), predictive processing (Keller et al., 2012; Attinger et al., 2017), novelty detection (Garrett et al., 2020, 2023), flexible routing of information flow (Yang et al., 2016; Wang and Yang, 2018), regulating global coherence (Veit et al., 2017, 2022), and gating of synaptic plasticity (Canto-Bustos et al., 2022). Interactions between different cell types in the canonical microcircuit can give rise to counterintuitive nonlinear phenomena. More specifically, in darkness, locomotion-induced top-down modulation via VIP decreases activity of SST neurons in layer 2/3 of mouse primary visual cortex (Fu et al., 2014; Fig. 1). In contrast, when animals receive visual stimuli, locomotion leads to an increase in SST activity (Pakan et al., 2016; Dipoppa et al., 2018; Fig. 1). This phenomenon in which the same locomotion-induced top-down modulation via VIP affects SST response oppositely depending on the visual stimulation condition is known as *response reversal* (Garcia del Molino et al., 2017).

Previous computational work has shown that networks with nonlinear neuronal input-output functions can generate response reversal (Garcia del Molino et al., 2017). However, cortical neurons exhibit highly irregular spiking (Shadlen and Newsome, 1998) and heterogeneous firing rates (Roxin et al., 2011; Buzsáki and Mizuseki, 2014) that are hallmarks of tightly balanced networks in which population-averaged responses are linear in the input (van Vreeswijk and Sompolinsky, 1998). This raises the possibility that the nonlinear population responses may result from other factors like dynamically changing synapses. On a perceptually and behaviorally relevant timescale from milliseconds to seconds, synapses are subject to short-term plasticity (STP) (Zucker and Regehr, 2002; Markram et al., 2015). Different types of synapses can experience different degrees of short-term depression (STD) or short-term facilitation (STF) (Zucker and Regehr, 2002). In particular, inhibitory synapses exhibit more pronounced short-term dynamics than excitatory

synapses (Campagnola et al., 2022). Furthermore, synapses from different interneuron subtypes can undergo different short-term plastic changes (Campagnola et al., 2022). Little is known about how these experimentally identified short-term plasticity mechanisms shape network dynamics and computations in recurrent neural circuits of multiple interneuron subtypes.

Response reversal of SST induced by the same top-down modulation may suggest that the network operates in different regimes under different stimulation conditions. Mounting evidence suggests that cortical networks operate in an inhibition-stabilized regime, in which feedback inhibition generated by the network is imperative to stabilize excitatory activity (Tsodyks et al., 1997; Sanzeni et al., 2020). In networks with one excitatory and one inhibitory population and fixed connectivity, an identifying characteristic of inhibition stabilization is that increasing (decreasing) excitatory input to the inhibitory population decreases (increases) inhibitory firing, known as the paradoxical effect (Tsodyks et al., 1997; Li et al., 2019; Miller and Palmigiano, 2020). Yet, it is unclear whether response reversal can be linked to inhibition stabilization and whether there exists a relationship between response reversal and the paradoxical effect. In addition, how short-term plasticity shapes inhibition stabilization in networks with multiple interneuron subtypes, particularly, how specific interneuron subtypes contribute to network stabilization (which we refer to as *interneuron-specific stabilization*) is unknown.

Here, we use analytical calculations and numerical simulations to demonstrate that inhibitory short-term plasticity enables response reversal without requiring neuronal nonlinearities. We find that despite not directly affecting SST and VIP activity, PV-to-E STD has a crucial influence on response reversal. We further reveal the relationship between response reversal, the paradoxical effect, and the inhibition stabilization property of the network. Interestingly, when the SST response to top-down modulation switches from suppression to enhancement, the network undergoes an interneuron-specific change in stabilization and becomes stabilized by SST alone instead of PV. In summary, our model suggests that inhibitory short-term plasticity enables the network to perform nonlinear computations and makes several experimentally testable predictions.

Results

To study how response reversal emerges in canonical cortical circuits, we used rate-based population models consisting of one excitatory (E) and three different inhibitory (PV, SST, VIP) populations with network connectivity constrained by previous experimental studies (Fig. 2A; Pfeffer et al., 2013). This type of model allows for a trade-off between including sufficient biological detail, while enabling mathematical analysis, and have previously been used with great success to study cortical computations (Murphy and Miller, 2009; Litwin-Kumar et al., 2016; Garcia del Molino et al., 2017; Mahrach et al., 2020; Richter and Gjorgjieva, 2022). Consistent with experimental work (Sanzeni et al., 2020), the network connectivity was chosen so that the network operates

in an inhibition-stabilized regime defined as the regime where feedback inhibition generated by the network is needed to stabilize the recurrent excitation (Tsodyks et al., 1997). As proposed by influential modeling work on cortical dynamics (van Vreeswijk and Sompolinsky, 1996, 1998), the network's population-averaged responses can be approximated as a rectified linear function of the input currents (Fig. 2A, inset). To account for activity-dependent changes in network connectivity on a perceptually and behaviorally relevant timescale, we modeled short-term plasticity based on recent experimental work from the Allen Institute (Campagnola et al., 2022), we incorporated the four most pronounced short-term plasticity mechanisms: PV-to-E short-term depression (STD), PV-to-PV STD, PV-to-SST STD, and SST-to-VIP short-term facilitation (STF) (Fig. 2A; Fig. S1; Methods). Since all the prominent synapses undergoing short-term plasticity are inhibitory, we refer to the plasticity mechanisms as *inhibitory short-term plasticity* (iSTP). The dynamics of the network with iSTP can be described as follows:

$$\tau_E \frac{dr_E}{dt} = -r_E + [J_{EE}\, r_E - x_{EP}\, J_{EP}\, r_P - J_{ES}\, r_S + g_E + \alpha]_+ \,, \tag{1}$$

$$\tau_P \frac{dr_P}{dt} = -r_P + [J_{PE}\, r_E - x_{PP}\, J_{PP}\, r_P - J_{PS}\, r_S + g_P + \alpha]_+ \,, \tag{2}$$

$$\tau_S \frac{dr_S}{dt} = -r_S + [J_{SE}\, r_E - J_{SV}\, r_V + g_S]_+ \,, \tag{3}$$

$$\tau_V \frac{dr_V}{dt} = -r_V + [J_{VE}\, r_E - x_{VP}\, J_{VP}\, r_P - u_{VS}\, J_{VS}\, r_S + g_V + c]_+ \,, \tag{4}$$

for the rates r_i of excitatory, PV, SST, and VIP populations with $i \in \{E, P, S, V\}$. τ_i represents the corresponding time constant of the rate dynamics, J_{ij} denotes the synaptic strength from population j to population i, and g_i is the individual background input. Importantly, we distinguish between bottom-up input to E and PV to represent different stimulation conditions, denoted as α, and top-down input to VIP mimicking locomotion-induced top-down modulation, denoted as c (Fig. 2A).

Short-term plasticity mechanisms are implemented based on the Tsodyks and Markram model (Tsodyks et al., 1998):

$$\frac{dx_{EP}}{dt} = \frac{1 - x_{EP}}{\tau_x} - U_d x_{EP} r_P, \tag{5}$$

$$\frac{dx_{PP}}{dt} = \frac{1 - x_{PP}}{\tau_x} - U_d x_{PP} r_P, \tag{6}$$

$$\frac{dx_{VP}}{dt} = \frac{1 - x_{VP}}{\tau_x} - U_d x_{VP} r_P, \tag{7}$$

$$\frac{du_{VS}}{dt} = \frac{1 - u_{VS}}{\tau_u} + U_f (U_{\max} - u_{VS}) r_S, \tag{8}$$

where x_{ij} is a short-term depression variable limited to the interval (0,1] for the synaptic connection from population j to population i. Biophysically, the short-term depression variable x represents the fraction of vesicles available for release. τ_x is the time constant of STD, and U_d is the depression

factor controlling the degree of depression induced by the presynaptic activity. Similarly, u_{ij} is a short-term facilitation variable constrained to the interval [1, U_{max}) for the synaptic connection from population j to population i. Unlike short-term depression, the short-term facilitation variable u biophysically represents the ability to release neurotransmitters. τ_u is the time constant of STF, U_f is the facilitation factor controlling the degree of facilitation induced by the presynaptic activity, and U_{max} is the maximal facilitation value.

iSTP enables response reversal of SST

To represent different stimulus conditions (e.g., darkness vs. visual stimulation), we varied the bottom-up input α to E and PV. Increasing α leads to increased baseline activity in all populations (Fig. S2). We modeled the effect of locomotion-induced top-down modulation on network activity by increasing the input to VIP by a positive value c. In our network model with iSTP, at different values of α, top-down modulation has an opposite effect on SST activity (Fig. 2B, C). For a low α, corresponding to low baseline activity in the absence of bottom-up input, top-down modulation via additional excitatory input to VIP decreases SST activity (Fig. 2B). In contrast, for a high α corresponding to high baseline activity, the same top-down modulation leads to an increase in SST activity and, thus, response reversal (Fig. 2C). Our modeling results suggest that under different stimulus conditions regulated by bottom-up inputs, identical top-down modulation reversely affects the change of SST activity, consistent with previous experiments (Fu et al., 2014; Pakan et al., 2016; Dipoppa et al., 2018).

To highlight the role of iSTP in generating response reversal of SST activity, we further simulated the same network while disabling iSTP (Fig. S3). In contrast to networks with iSTP, the change of SST activity remains largely unchanged for different values of α when iSTP is disabled (Fig. 2D).

Interestingly, for a high α, despite increased activity of all inhibitory populations, excitatory activity increased at the steady state (Fig. 2C). This observation appears to differ from the image of the classical disinhibition mechanism, in which the reduction of inhibition causes elevated excitatory activity. We confirm this by plotting the amount of inhibition the excitatory population receives (Fig. 2E-G). Surprisingly, even for a low α, despite decreased SST activity, top-down modulation via VIP increases the total inhibition to the excitatory population at the steady state (Fig. 2E). Rather than the decrease in the total inhibition, the increase in the recurrent excitation contributes to the elevated excitatory activity (Fig. 2E-G). Importantly, a co-increase in the excitation and inhibition of the excitatory population is a distinctive feature in inhibition-stabilized networks (Litwin-Kumar et al., 2016; Miller and Palmigiano, 2020; Wu and Gjorgjieva, 2023). In contrast, non-inhibition-stabilized networks do not experience an elevation in the total inhibitory inputs to the excitatory population (Fig. S4).

Taken together, our numerical simulation results reveal that experimentally identified forms of iSTP

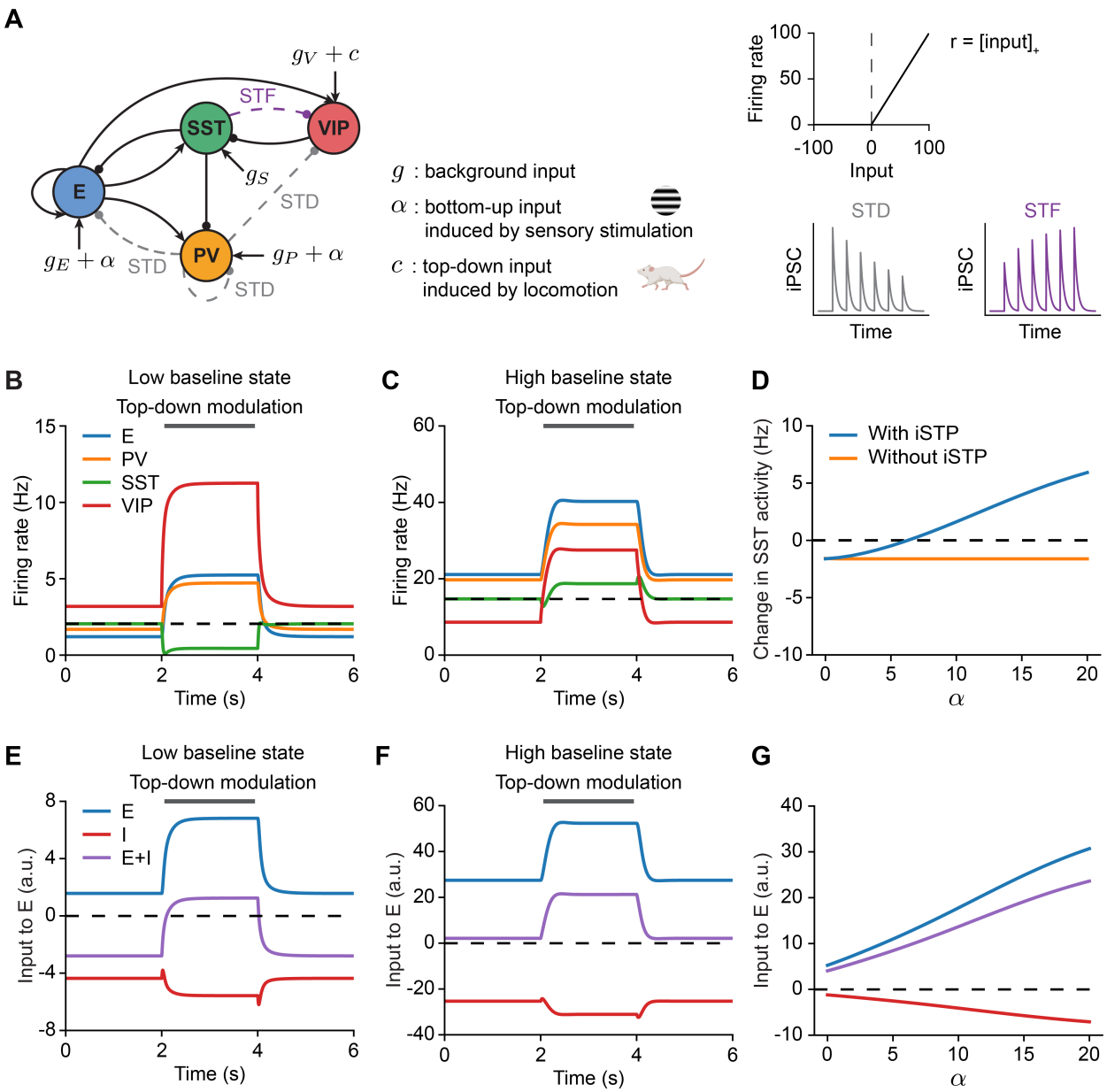

Fig. 2. Inhibitory short-term plasticity enables response reversal of SST induced by top-down modulation via VIP. (**A**) Schematic of network model with one excitatory (E, blue) population and three distinct inhibitory populations including PV (orange), SST (green), and VIP (red). Short-term depressing (STD) and short-term facilitating (STF) connections are indicated by the dashed lines. Each population receives a background input *g*. E and PV receive bottom-up input α depending on sensory stimulation, and VIP receives top-down input *c* during locomotion. Top right: rectified linear input-output function; Bottom right: cartoons showing the evolution of inhibitory postsynaptic current (iPSC) in response to a step current input under STD (gray) and STF (purple) regulations. (**B**) Network responses to top-down modulation without any bottom-up input (α = 0) corresponding to darkness without sensory stimulation. Top-down modulation via VIP is applied during the interval from 2 to 4 s (gray bar). Different colors denote the activity of different populations. The dashed line represents the initial activity level of SST. (**C**) Same as B but at α = 15 corresponding to sensory stimulation. (**D**) Change in SST response induced by top-down modulation to VIP as a function of bottom-up input α in networks with iSTP (blue) and without iSTP (orange). (**E**) Input to the E population at α = 0. Top-down modulation via VIP is applied during the interval from 2 to 4 s (gray bar). Different colors indicate different sources: input from the E population (blue), input from the I populations (red), and the sum of the inputs from the E and I populations (purple). (**F**) Same as E but at α = 15. (**G**) Change in different sources of inputs to the E population measured between baseline and at steady state during top-down modulation as a function of bottom-up input α.

enable our model networks to perform response reversal, a nonlinear computation observed in cortical circuits.

Theoretical analysis

To better understand how iSTP enables our model network to perform response reversal of SST activity, we sought to mathematically analyze how top-down modulation affects SST activity. Locomotion-induced top-down modulation can be considered a form of perturbation to the VIP population. Investigating how top-down modulation inversely affects SST activity under different stimulus conditions can therefore be mathematically formulated as how perturbations to VIP affect SST activity under varying levels of α. To this end, we extended previous studies on static networks (Garcia del Molino et al., 2017; Palmigiano et al., 2023) and developed a general theoretical framework for networks with iSTP. More specifically, we derived how the steady state response of any population changes with a perturbation of the external input to a given population while including iSTP (see Methods). Using this approach, we formulated the change of SST activity induced by top-down modulation via the change of the input to VIP, $\mathbf{R}_{SV}$, as follows:

$$\mathbf{R}_{SV} = D\,\mathbf{K}_{SV}\delta g_V. \tag{9}$$

Here, D is a positive quantity for any stable network (see Methods, Fig. S5A), δg_V represents the perturbation of the VIP population's input which is a positive number c in our network setting and $\mathbf{K}_{SV}$ is the response factor which is defined by:

$$\begin{aligned}\mathbf{K}_{SV} = {} & (x^*_{PP} + x^{*'}_{PP} r_P - 1)(J_{EE} - 1)J_{PP}J_{SV} - (x^*_{EP} + x^{*'}_{EP} r_P - 1)J_{SV}J_{PE}J_{EP} \\ & + (J_{EE} - 1)(J_{PP}J_{SV} + J_{SV}) - J_{SV}J_{PE}J_{EP},\end{aligned} \tag{10}$$

with x^*_{ij} the short-term plasticity variable from population j to population i at steady state before perturbation, and $x^{*'}_{ij}$ the change in the plasticity variable due to the perturbation (see Methods).

A negative (positive) $\mathbf{R}_{SV}$ denotes a decrease (increase) in SST activity caused by top-down modulation. For response reversal, $\mathbf{R}_{SV}$ must switch its sign for different values of α, corresponding to different stimulus conditions. More specifically, when animals perceive no visual stimulus in darkness, in other words, when α is low, top-down modulation via VIP decreases SST activity. Therefore, $\mathbf{R}_{SV}$ is expected to be negative for a low α. In contrast, when animals receive a visual stimulus, namely, when α is high, top-down modulation via VIP increases SST activity. Thus, $\mathbf{R}_{SV}$ is expected to be positive for a high α.

In agreement with our simulation results (Fig. 2B, C), we observed that $\mathbf{R}_{SV}$ changes its sign when calculated at different values of α (Fig. 3). As our theoretical framework is based on the linearization of the network dynamics around the steady state and higher order terms are ignored (see Methods), the computed $\mathbf{R}_{SV}$ agrees well with the numerical simulation results for small per-

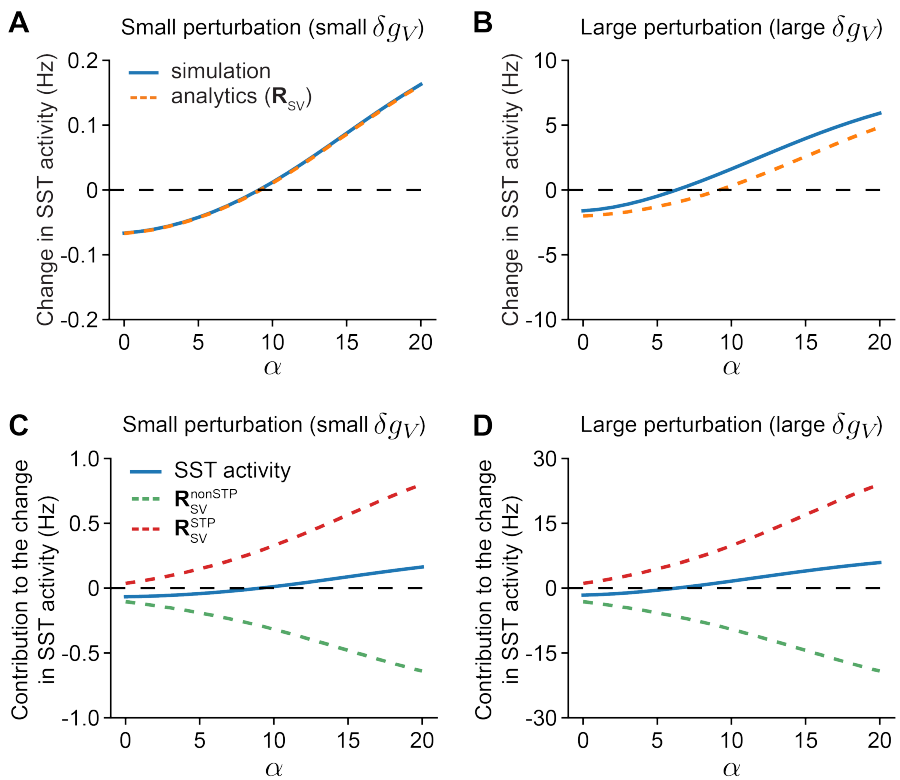

Fig. 3. Comparison between analytical predictions and numerical simulations on the change in SST population response induced by the perturbation to VIP as a function of different stimulus conditions regulated by α. (**A**) Analytical prediction (orange) of the change in SST population response induced by the perturbation to VIP ($\mathbf{R}_{SV}$) matches closely with numerical simulation (blue) for a small perturbation (δg_V = 0.1). (**B**) Same as A but with a large perturbation (δg_V = 3). (**C**) Analytical contributions of the STP-dependent term $\mathbf{R}_{SV}^{\text{STP}}$ (dashed red) and the STP-independent term $\mathbf{R}_{SV}^{\text{nonSTP}}$ (dashed green) to the change in SST activity (solid blue) as a function of bottom-up input α. Here, δg_V = 0.1. (**D**) Same as C but for δg_V = 3.

turbations (Fig. 3A) and diverges for large perturbations (Fig. 3B). Yet, it qualitatively captures the key aspect of modeling behaviors: the sign switch of the change in SST activity induced by top-down modulation with different values of α.

As shown in Eq. (9), since D and δg_V are positive, $\mathbf{K}_{SV}$ is the only term that can change the sign of $\mathbf{R}_{SV}$. To further investigate the influence of the iSTP mechanisms on response reversal, we rewrote $\mathbf{K}_{SV}$ as a sum of a short-term plasticity-dependent and a short-term plasticity-independent term:

$$\mathbf{K}_{SV} = \mathbf{K}_{SV}^{\text{STP}} + \mathbf{K}_{SV}^{\text{nonSTP}}, \tag{11}$$

where

$$\mathbf{K}_{SV}^{\text{STP}} = (x_{PP}^{*} + x_{PP}^{*'} r_P - 1)(J_{EE} - 1) J_{PP} J_{SV} - (x_{EP}^{*} + x_{EP}^{*'} r_P - 1) J_{SV} J_{PE} J_{EP}, \tag{12}$$

and

$$\mathbf{K}_{SV}^{\text{nonSTP}} = (J_{EE} - 1)(J_{PP} J_{SV} + J_{SV}) - J_{SV} J_{PE} J_{EP}. \tag{13}$$

Analogously to $\mathbf{K}_{SV}$, we have:

$$\mathbf{R}_{SV} = \mathbf{R}_{SV}^{\text{STP}} + \mathbf{R}_{SV}^{\text{nonSTP}} = D\,\mathbf{K}_{SV}^{\text{STP}}\delta g_V + D\,\mathbf{K}_{SV}^{\text{nonSTP}}\delta g_V. \tag{14}$$

As the short-term plasticity-independent term $\mathbf{K}_{SV}^{\text{nonSTP}}$ is governed by the absolute values of the static network weights, it is constant over the entire range of change in bottom-up input. In other words, $\mathbf{K}_{SV}^{\text{nonSTP}}$ does not change with α. Note that because of Eq. (9) and since D is always positive but subject to change in magnitude (Fig. S5A), $\mathbf{R}_{SV}^{\text{nonSTP}}$ changes in magnitude as well (Figure 3B). To match recent experimental findings that the network is inhibition stabilized when animals receive no stimulus in darkness (Sanzeni et al., 2020), we set J_{EE} to be larger than 1 (see Methods). In this case, the short-term plasticity-independent term $\mathbf{K}_{SV}^{\text{nonSTP}}$ is always negative, which implies that $\mathbf{R}_{SV}^{\text{nonSTP}}$ is also always negative. Thus, the short-term plasticity-dependent term $\mathbf{K}_{SV}^{\text{STP}}$, and as a result, $\mathbf{R}_{SV}^{\text{STP}}$ too, is the only part that can influence the sign of $\mathbf{R}_{SV}$ and enable the network to perform response reversal of SST activity (Fig. 3C, D).

In conclusion, our theoretical framework enables us to analyze how perturbations affect the activity of individual populations and hence reveals how iSTP enables response reversal of SST activity.

PV-to-E STD plays a key role in the generation of response reversal

Next, we sought to dissect the role of individual iSTP mechanisms in response reversal. To this end, we separated the short-term plasticity-dependent term $\mathbf{K}_{SV}^{\text{STP}}$ (Eq. 12) into a PV-to-PV STD-dependent part $\mathbf{K}_{SV}^{\text{PPD}}$ and a PV-to-E STD-dependent part $\mathbf{K}_{SV}^{\text{PED}}$ as follows:

$$\mathbf{K}_{SV}^{\text{STP}} = \underbrace{(x_{PP}^{*} + x_{PP}^{*'} r_P - 1)(J_{EE} - 1)J_{PP}J_{SV}}_{\mathbf{K}_{SV}^{\text{PPD}}} \underbrace{-(x_{EP}^{*} + x_{EP}^{*'} r_P - 1)J_{SV}J_{PE}J_{EP}}_{\mathbf{K}_{SV}^{\text{PED}}}. \tag{15}$$

Since both $x_{PP}^{*} + x_{PP}^{*'} r_P - 1$ and $x_{EP}^{*} + x_{EP}^{*'} r_P - 1$ are always negative (see Methods, Fig. S5B), the PV-to-PV STD-dependent part $\mathbf{K}_{SV}^{\text{PPD}}$ is always negative and the PV-to-E STD-dependent part $\mathbf{K}_{SV}^{\text{PED}}$ is always positive. Importantly, $\mathbf{K}_{SV}^{\text{PPD}}$ decreases with increasing bottom-up input α, whereas $\mathbf{K}_{SV}^{\text{PED}}$ increases with increasing bottom-up input α (see Methods, Fig. S5C).

Similarly, we can write:

$$\mathbf{R}_{SV}^{\text{STP}} = \mathbf{R}_{SV}^{\text{PPD}} + \mathbf{R}_{SV}^{\text{PED}} = D\,\mathbf{K}_{SV}^{\text{PPD}}\delta g_V + D\,\mathbf{K}_{SV}^{\text{PED}}\delta g_V. \tag{16}$$

$\mathbf{R}_{SV}^{\text{PPD}}$ and $\mathbf{R}_{SV}^{\text{PED}}$ show similar changes as $\mathbf{K}_{SV}^{\text{PPD}}$ and $\mathbf{K}_{SV}^{\text{PED}}$, respectively (Fig. 4A). Therefore, when bottom-up input α increases from a low to a high level (e.g., switching from darkness to visual stimulation condition), to display response reversal $\mathbf{R}_{SV}^{\text{STP}}$ must overcome in magnitude the negative $\mathbf{R}_{SV}^{\text{nonSTP}}$, resulting in an overall switch of $\mathbf{R}_{SV}$ from negative to positive. The increasing PV-to-E STD-dependent term rather than the decreasing PV-to-PV STD-dependent term is imperative for this switch (Figure 4A). Likewise, when bottom-up input α decreases from a high to a low level,

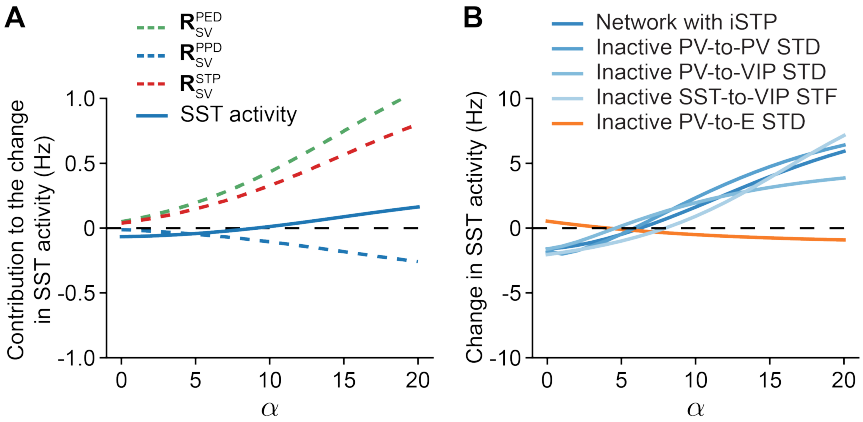

Fig. 4. PV-to-E STD plays a crucial role in the response reversal of SST. (**A**) Analytical contributions of the PV-to-E STD-dependent term $\mathbf{R}_{SV}^{\mathrm{PED}}$ (dashed green), the PV-to-PV STD-dependent term $\mathbf{R}_{SV}^{\mathrm{PPD}}$ (dashed blue), and the overall STD-dependent term $\mathbf{R}_{SV}^{\mathrm{STP}}$ (dashed red) to the change in SST activity (solid blue) as a function of bottom-up input α. Here, δg_V = 0.1. (**B**) Change in SST response induced by top-down modulation to VIP as a function of bottom-up input α for different network configurations marked with different colors. Here, δg_V = 3.

to have response reversal, $\mathbf{R}_{SV}^{\mathrm{STP}}$ has to be smaller than $\mathbf{R}_{SV}^{\mathrm{nonSTP}}$, resulting in a switch of $\mathbf{R}_{SV}$ from positive to negative. The decreasing PV-to-E STD-dependent term rather than the increasing PV-to-PV STD-dependent term is necessary for the switch. As no terms directly associated with PV-to-SST STD and SST-to-VIP STF appear in Eq. (15), our analysis shows that these two mechanisms are unimportant for the generation of response reversal.

We performed the same simulations as with the intact network to confirm our analysis while inactivating specific iSTP mechanisms. The inactivation of particular mechanisms was implemented by freezing the respective plasticity variables at their baseline values when bottom-up input is high, ensuring that the steady-state activities of all populations are positive at the baseline and during the top-down modulation period. We then varied the bottom-up inputs from high to low and found that SST response reversal still occurs despite inactivating PV-to-PV STD, PV-to-SST STD, or SST-to-VIP STF. Such networks show similar patterns to networks with intact iSTP (Fig. 4B). In contrast, when PV-to-E STD is inactivated, the change of SST activity manifests an opposite trend from that in networks with intact iSTP (Fig. 4B).

Taken together, our analysis and numerical simulations unveil that PV-to-E STD is the determining mechanism for generating response reversal. In contrast, the effects of PV-to-PV STD, PV-to-VIP STD, and SST-to-VIP STF on response reversal are negligible.

Relationship between response reversal and the paradoxical effect

Locomotion-induced top-down modulation excites VIP and effectively inhibits SST due to the mutually inhibitory connections between VIP and SST. However, when animals receive visual stimuli

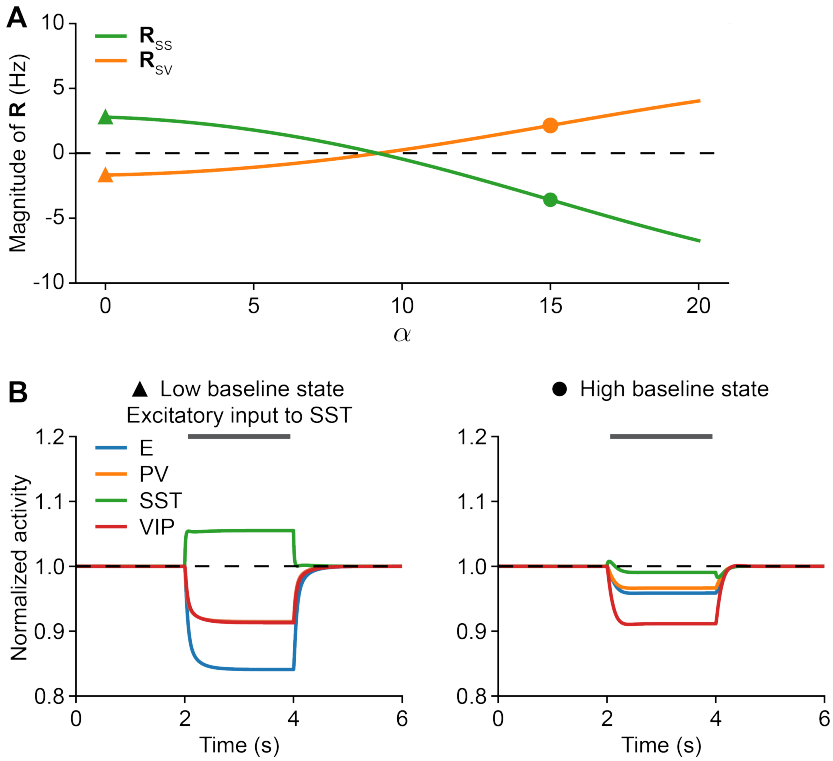

Fig. 5. Relationship between response reversal and paradoxical response of SST. (**A**) Analytical predictions of the change in SST response induced by an excitatory perturbation (δg_S) to SST, $\mathbf{R}_{SS}$ in green, and change in SST response induced by an excitatory perturbation (δg_V) to VIP, $\mathbf{R}_{SV}$ in orange, as a function of bottom-up input α. Here, $\delta g_V = \delta g_S = 3$. (**B**) Left: Normalized activity when injecting additional excitatory current into SST for networks at a low baseline state corresponding to $\alpha = 0$ marked with triangular in A. SST does not show a paradoxical response. Right: Same as left but for networks at a high baseline state corresponding to $\alpha = 15$ marked with a dot in A.

at a high baseline activity state (high α), VIP inhibition by top-down modulation increases the activity of SST, which is reminiscent of the paradoxical effect (Tsodyks et al., 1997; Ozeki et al., 2009). We thus sought to identify the relationship between response reversal and the paradoxical effect. To this end, we derived the change of SST activity induced by a change in the input to SST itself, $\mathbf{R}_{SS}$, as follows:

$$\mathbf{R}_{SS} = D\,\mathbf{K}_{SS}\delta g_S, \tag{17}$$

where

$$\begin{aligned}\mathbf{K}_{SS} &= -\Big[(x^*_{PP} + x^{*\prime}_{PP} r_P - 1)(J_{EE} - 1)J_{PP} - (x^*_{EP} + x^{*\prime}_{EP} r_P - 1)J_{PE}J_{EP} \\ &\quad + (J_{EE} - 1)(J_{PP} + 1) - J_{PE}J_{EP}\Big] \\ &= -\mathbf{K}_{SV}/J_{SV},\end{aligned} \tag{18}$$

and δg_S represents the change of input to SST. Furthermore,

$$\mathbf{R}_{SS} = -\frac{\delta g_S}{J_{SV}\delta g_V}\mathbf{R}_{SV}. \tag{19}$$

When δg_S is positive, to obtain a paradoxical response of SST, i.e., to have a negative $\mathbf{R}_{SS}$, $\mathbf{K}_{SS}$ has to be negative. As $\mathbf{K}_{SS}$ is equal to $-\mathbf{K}_{SV}/J_{SV}$, for low α corresponding to the darkness condition ($\mathbf{K}_{SV}$ and $\mathbf{R}_{SV}$ are negative), $\mathbf{K}_{SS}$ and $\mathbf{R}_{SS}$ are positive, hence, no paradoxical response is observed (Fig. 5). In contrast, for high α corresponding to the visual stimulation condition ($\mathbf{K}_{SV}$ and $\mathbf{R}_{SV}$ are positive), $\mathbf{K}_{SS}$ and $\mathbf{R}_{SS}$ are negative, therefore, SST exhibits a paradoxical response (Fig. 5).

Therefore, we have mathematically proven a correspondence between response reversal and the paradoxical response of SST. More specifically, the SST population will not show a paradoxical response when top-down modulation via VIP decreases SST activity but will respond paradoxically when top-down modulation via VIP increases SST activity.

Relationship between response reversal, the paradoxical effect, and interneuron-specific stabilization

The paradoxical effect is a defining characteristic of inhibition stabilization in networks with fixed connectivity (Tsodyks et al., 1997). We, therefore, sought to investigate the relationship between response reversal and inhibition stabilization. Identifying the relationship may shed light on how response reversal relates to other cortical functions as inhibition-stabilized networks can perform a variety of computations (Sadeh and Clopath, 2021). As the network in our study consists of three different inhibitory populations, the network can, in principle, be stabilized by any specific interneuron subtype. In addition to identifying inhibition stabilization, we particularly aimed to determine the type of interneuron which stabilizes the model networks in the case of different stimulation conditions.

To this end, we computed the leading eigenvalue of the Jacobian of individual subnetworks with the corresponding firing rates and short-term plasticity dynamics while excluding specific interneuron subtypes. Such eigenvalues can be used to determine the stability of the subnetwork. A negative leading eigenvalue implies that the fixed point is stable and a transient perturbation to the system does not result in a deviation from the original fixed point. In contrast, a positive leading eigenvalue means that the fixed point is unstable and a transient perturbation causes a deviation from the original fixed point. We found that the leading eigenvalue of the Jacobian of the E subnetwork is positive for all values of α, suggesting that the E subnetwork is unstable and the network is inhibition stabilized for all stimulation conditions (see Methods, Fig. S6). Furthermore, VIP does not stabilize the network, as the leading eigenvalue of the Jacobian of the E-VIP subnetwork is always positive (Fig. S6). By computing the leading eigenvalue of the Jacobian of the E-PV-VIP subnetwork, we found that the eigenvalue switches from negative to positive once the response of SST to top-down modulation is reversed (Fig. 6A), indicating that the network becomes stabilized by SST rather than PV when top-down modulation via VIP increases SST activity. This result further suggests that in the regime in which top-down modulation via VIP decreases SST activity,

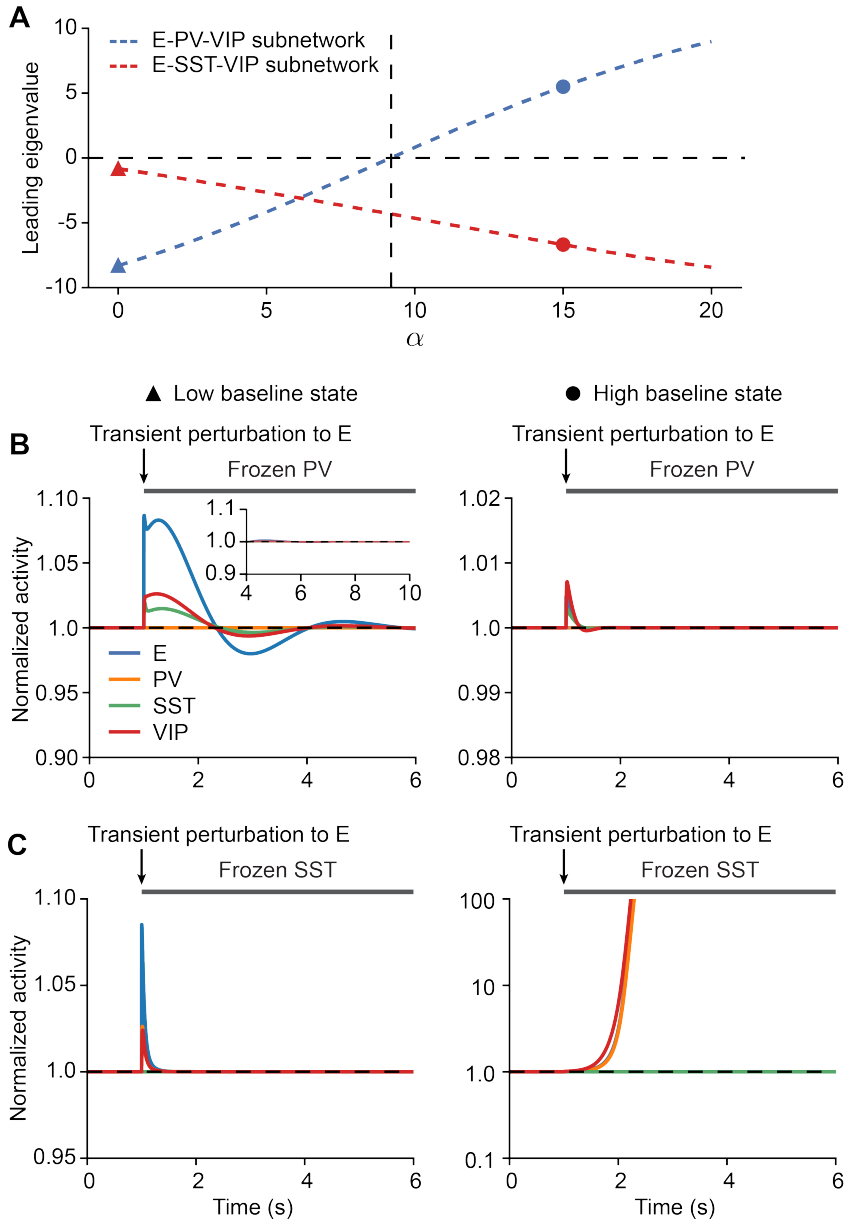

Fig. 6. The network undergoes a change in interneuron-specific stabilization with increasing bottom-up input. (**A**) Leading eigenvalues of the E-PV-VIP subnetwork (dashed blue) and the E-SST-VIP subnetwork (dashed red) as a function of bottom-up input α. The response reversal boundary is indicated by the vertical dashed line. (**B**) Left: Normalized activity when injecting additional excitatory current into E while freezing PV for networks at a low baseline state corresponding to α = 0 marked with a triangle in A. A small transient excitatory perturbation to excitatory population activity is introduced at the time marked with arrows. The periods in which PV is frozen are marked with the gray bar. Right: Same as left but for networks at a high baseline state corresponding to α = 15 marked with a dot in A. (**C**) Similar to B but with frozen SST.

the network can be stabilized by PV. To determine whether PV is the only interneuron subtype stabilizing the network in that regime, we calculated the leading eigenvalue of the Jacobian of the E-SST-VIP subnetwork. We found that this eigenvalue is always negative (Fig. 6A), suggesting

that SST can serve the stabilization role in that regime as well as PV.

We confirmed these results by injecting a transient excitatory perturbation into the excitatory population while clamping the activity of either PV or SST. We found that despite clamping PV activity, the network is stable to perturbations over the entire range of α and reaches the same fixed point as before the perturbation (Fig. 6B). In contrast, when SST activity is clamped, while the network is stable to perturbations at low α, transient perturbation leads to unstable dynamics at high α (Fig. 6C). Moreover, the condition for the paradoxical response of SST is in agreement with the stability condition for the E-PV-VIP subnetwork (see Methods). More specifically, when SST responds paradoxically, the E-PV-VIP subnetwork becomes unstable, suggesting that SST rather than PV stabilizes the network. Based on the relationship between response reversal and the paradoxical effect shown in the previous section, we can conclude that response reversal is only possible when SST, rather than PV, stabilizes the network, as the E-PV-VIP subnetwork is unstable in this regime.

Consistent with this change in the interneuron-specific stabilization property, we found that the excitatory population receives inhibition which shifts from being PV dominant to SST dominant once response reversal occurs with increasing α (Fig. S7). More specifically, at the low baseline state, the excitatory population receives more inhibition from PV than SST (Fig. S7A). Top-down modulation via VIP leads to increases in the overall inhibition and the inhibition from PV at the steady state but a decrease in the inhibition from SST (Fig. S7A). In contrast, at the high baseline state, the excitatory population receives more inhibition from SST than PV (Fig. S7B). Top-down modulation increases the overall inhibition at the steady state as well as the inhibition from both PV and SST (Fig. S7B). This increase in total inhibition at the steady state observed during top-down modulation is a unique characteristic of inhibition-stabilized networks with strong recurrent excitatory connections compared to non-inhibition-stabilization networks (Fig. S4; Litwin-Kumar et al., 2016; Miller and Palmigiano, 2020; Wu and Gjorgjieva, 2023). In inhibition-stabilized networks, top-down modulation induces a transient disinhibition enabling the growth of recurrent excitation and increasing excitation and inhibition to the excitatory population at the steady state.

Taken together, these results suggest that increasing bottom-up input, representing a switch in stimulation condition, leads to a change in the impact of top-down modulation on SST activity from suppression to enhancement. Once top-down modulation increases SST activity, the network exhibits a paradoxical response of SST and becomes stabilized by SST rather than PV (Figs. 5A, 6A, 7).

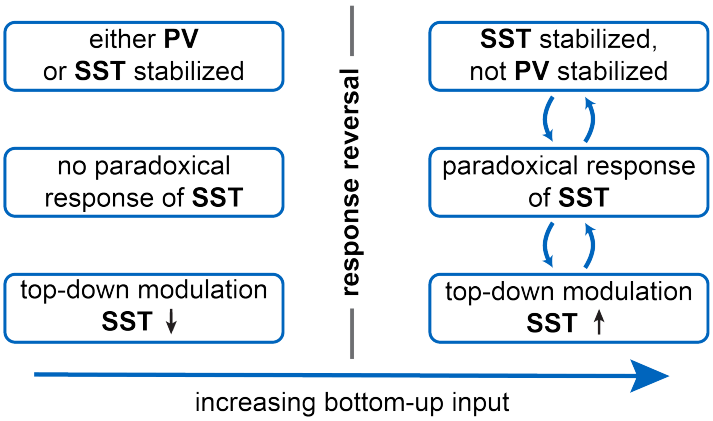

Fig. 7. The relationship of response reversal, inhibition stabilization, and paradoxical response. With increasing bottom-up input, the effect of top-down modulation on SST switches from suppression to enhancement. When the response of SST induced by top-down modulation is reversed, the network exhibits the paradoxical effect in the response of SST and becomes stabilized by SST instead of PV.

Discussion

In this paper, we investigated how experimentally measured inhibitory short-term plasticity (iSTP) mechanisms enable model networks with one excitatory and three types of interneuron populations to perform a nonlinear computation known as response reversal. Using analytical calculations and numerical simulations, we identified that PV-to-E short-term depression (STD) is the iSTP mechanism critical for generating response reversal. We further clarified the relationship between response reversal, the paradoxical response of SST, and the interneuron-specific stabilization property of the network, making important links between well-known operating regimes of cortical network dynamics.

In this study, we made several assumptions that enabled us to analytically understand response reversal. First, we studied responses in the presence of bottom-up and top-down inputs relative to a baseline state, assuming that the network activity has reached a fixed point, and we did not consider scenarios like multistability (Hertäg and Sprekeler, 2019; Pietras et al., 2022) or oscillations (Veit et al., 2022). While multistability and oscillations have been observed in the brain (Wang, 2001; Buzsáki and Draguhn, 2004), the single stable fixed point assumed here is considered to be a realistic approximation of the awake sensory cortex (Miller, 2016).

With regards to the modeled short-term plasticity, our model only included PV-to-E STD, PV-to-PV STD, PV-to-VIP STD, and SST-to-VIP STF. In addition to these four short-term plasticity mechanisms, substantial PV-to-SST STD has also been reported (Campagnola et al., 2022). However, experimental studies demonstrated negligible inhibition from PV to SST (Pfeffer et al., 2013), and hence we did not consider PV-to-SST STD.

Furthermore, our work models the neural input-output function as a rectified linear function, a

characteristic feature of tightly balanced networks (van Vreeswijk and Sompolinsky, 1996, 1998). Indeed, without iSTP, our model network behaves like a linear network when all populations have positive activity. In addition to iSTP proposed in our study, several other factors can induce nonlinearities in the population response and, therefore, could contribute to the studied response reversal. Recent studies have suggested that cortical networks may operate in a loosely balanced regime, resulting in a supralinear input-output function (Ahmadian et al., 2013; Hennequin et al., 2018; Ahmadian and Miller, 2021; Ekelmans et al., 2023). Response reversal can also be generated by such a nonlinear input-output function (Garcia del Molino et al., 2017).

Fourth, we modeled neurons of the same type as a homogeneous population governed by the same dynamics. In contrast, even within the same cell type, biological neurons have highly heterogeneous time constants and firing thresholds (Allen Institute for Brain Science, 2019; Cembrowski and Spruston, 2019). Such heterogeneity can theoretically also give rise to nonlinear population responses (Landau et al., 2016; Vegué and Roxin, 2019). Moreover, biological neurons possess complex morphologies (Jiang et al., 2015; Peng et al., 2021) and manifest nonlinear dendritic integrations (Poirazi et al., 2003; London and Häusser, 2005; Larkum et al., 2009; Tzilivaki et al., 2019). This suggests that the complete set of underlying mechanisms behind response reversal is even richer and remains to be examined experimentally.

Our study makes several predictions. First, during top-down modulation, along with decreased SST activity, we also observed that the inhibition to the excitatory population increased at the steady state. Top-down modulation via VIP induces transient disinhibition, facilitating the growth of recurrent excitation and resulting in increased excitatory activity. This increased recurrent excitation is balanced by the concurrent increase in inhibition, which is a characteristic of inhibition-stabilized networks. This prediction can be tested experimentally by measuring excitatory and inhibitory currents to the excitatory neurons during top-down modulation.

Second, due to iSTP, locomotion-induced top-down modulation via VIP can reversely regulate SST response under different stimulus conditions. Although PV-to-E STD does not directly affect SST and VIP activity, surprisingly, our analysis suggests that PV-to-E STD is the determining mechanism underlying the generation of response reversal. Theoretical studies have demonstrated that inhibitory-to-inhibitory connections have the dominant impact on cortical dynamics, memory capacity, and working memory maintenance (Mongillo et al., 2018; Kim and Sejnowski, 2021). Here, our work suggests that the dynamics of inhibitory to excitatory synapses can be more important than those of inhibitory to inhibitory synapses to generate certain nonlinear phenomena.

Third, our theory reveals a correspondence between response reversal and a paradoxical response of SST in the presence of iSTP. More specifically, when the bottom-up stimulation condition switches from darkness to visual input, the impact of locomotion-induced top-down modulation via VIP on SST activity changes from suppression to enhancement. Once the change of SST activity induced by top-down modulation is facilitating, the network exhibits a paradoxical response

of SST. This correspondence can therefore be tested directly in future optogenetic experiments to see whether injecting excitatory (inhibitory) currents into the SST population indeed decreases (increases) its activity.

Fourth, our analysis shows that response reversal is tightly linked to the interneuron-specific stabilization property of the network. In darkness, when top-down modulation decreases SST activity, both PV and SST stabilize the network. In contrast, in the presence of visual stimuli, top-down modulation increases SST activity, and the network becomes stabilized by SST rather than PV. Consistent with recent studies (Sanzeni et al., 2020), the network is inhibition-stabilized in all bottom-up stimulation conditions, even in darkness without visual stimuli. However, contrary to recent studies suggesting that PV is positioned to stabilize network activity (Bos et al., 2020), our work suggested that the specific inhibitory cell type stabilizing the network can change dynamically depending on the stimulation condition. Such an interneuron-specific change in the stabilization property can be tested experimentally by probing the paradoxical effects.

Last, for the network connectivity and the set of short-term plasticity mechanisms considered here, our results show that when SST stabilizes the network, the network exhibits a paradoxical response of SST. Several studies have investigated the relationship between inhibition stabilization and the paradoxical effect in networks with multiple interneuron subtypes (Litwin-Kumar et al., 2016; Miller and Palmigiano, 2020; Mahrach et al., 2020; Richter and Gjorgjieva, 2022; Palmigiano et al., 2023), in networks with short-term plasticity while ignoring different cell types (Sanzeni et al., 2020; Wu and Zenke, 2021), as well as in networks with multiple interneuron subtypes and short-term plasticity while ignoring cell type specificity (Wu and Gjorgjieva, 2023). Here, we revealed the relationship between interneuron-specific stabilization and the paradoxical effect in networks with multiple interneuron subtypes in the presence of a set of short-term plasticity mechanisms.

Taken together, our work sheds light on how experimentally identified iSTP mechanisms can generate response reversal, reveals the roles of individual iSTP mechanisms in response reversal, and uncovers the relationship between response reversal, the paradoxical effect, and interneuron-specific stabilization properties.

Methods

Response matrix

To investigate how the input to one particular population affects the response of any given population in the presence of short-term plasticity, we developed a general theoretical framework using linear perturbation theory. Using the separation of time scales for the rate dynamics and the short-term plasticity dynamics, we can write the system of equations introduced before (Equations (1) to (4)) in matrix form while replacing the short-term plasticity variables with their steady-state val-

ues, as follows,

$$\mathbf{T}\frac{d}{dt}\mathbf{r} = -\mathbf{r} + \mathbf{f}(\mathbf{P} \circ \mathbf{J}\mathbf{r} + \mathbf{g}), \tag{20}$$

where $\mathbf{T}$ is a diagonal matrix of time constants of the firing rate dynamics, $\mathbf{r}$ a vector of firing rates of the different populations, $\mathbf{f}(\mathbf{x})$ a vector of the rectified linear input-output function of the respective populations, $\mathbf{P}$ a matrix of the short-term plasticity variables, $\mathbf{J}$ the connectivity matrix, and $\mathbf{g}$ a vector of inputs to different populations. $\circ$ denotes the element-wise product.

$$\mathbf{T} = \begin{bmatrix} \tau_E & 0 & 0 & 0 \\ 0 & \tau_P & 0 & 0 \\ 0 & 0 & \tau_S & 0 \\ 0 & 0 & 0 & \tau_V \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} r_E \\ r_P \\ r_S \\ r_V \end{bmatrix}, \quad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_E(x_1) \\ f_P(x_2) \\ f_S(x_3) \\ f_V(x_4) \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 1 & x^*_{EP} & 1 & 1 \\ 1 & x^*_{PP} & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & x^*_{VP} & u^*_{VS} & 1 \end{bmatrix},$$

$$\mathbf{J} = \begin{bmatrix} J_{EE} & -J_{EP} & -J_{ES} & 0 \\ J_{PE} & -J_{PP} & -J_{PS} & 0 \\ J_{SE} & 0 & 0 & -J_{SV} \\ J_{VE} & -J_{VP} & -J_{VS} & 0 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_E + \alpha \\ g_P + \alpha \\ g_S \\ g_V \end{bmatrix}. \tag{21}$$

The steady states of the short-term plasticity variables are obtained by setting the Eqs. (5), (6), (7), (8) to 0 and given by

$$x^*_{EP} = x^*_{PP} = x^*_{VP} = \frac{1}{1 + U_d r_P \tau_x}, \tag{22}$$

$$u^*_{VS} = \frac{1 + U_f U_{\max} r_S \tau_u}{1 + U_f r_S \tau_u}. \tag{23}$$

Note that since the steady states of the short-term plasticity variables are determined by the presynaptic activity, x^*_{EP}, x^*_{PP}, and x^*_{VP} are the same. Furthermore, if short-term plasticity is not present on the synapses from j to i, the corresponding element $\mathbf{P}_{ij}$ is 1.

By linearizing about the fixed point, we obtain the following equation:

$$\mathbf{T}\frac{d}{dt}\delta\mathbf{r} = -\delta\mathbf{r} + \mathbf{F}(\mathbf{P} \circ \mathbf{J})\delta\mathbf{r} + \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\delta\mathbf{r} + \mathbf{F}\delta\mathbf{g}. \tag{24}$$

Here, $\delta\mathbf{r}$ is a vector containing the deviations of firing rates from their fixed point values. $\mathbf{F}$ is a diagonal matrix containing the derivatives of the input-output functions with respect to the firing rate of different populations. $\mathbf{P}'$ is a matrix containing the derivative of the short-term plasticity variables with respect to the corresponding presynaptic firing rate, evaluated at the fixed point. $\mathrm{diag}(\mathbf{r})$ is a diagonal matrix containing the firing rates of different populations. And $\delta\mathbf{g}$ is a vector containing the changes/perturbations of external inputs to different populations.

$$\delta\mathbf{r} = \begin{bmatrix} \delta r_E \\ \delta r_P \\ \delta r_S \\ \delta r_V \end{bmatrix}, \ \mathbf{F} = \begin{bmatrix} f'_E & 0 & 0 & 0 \\ 0 & f'_P & 0 & 0 \\ 0 & 0 & f'_S & 0 \\ 0 & 0 & 0 & f'_V \end{bmatrix}, \ \mathbf{P}' = \begin{bmatrix} 0 & x^{*\prime}_{EP} & 0 & 0 \\ 0 & x^{*\prime}_{PP} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & x^{*\prime}_{VP} & u^{*\prime}_{VS} & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{U_d \tau_x}{(1+U_d \tau_x r_P)^2} & 0 & 0 \\ 0 & -\frac{U_d \tau_x}{(1+U_d \tau_x r_P)^2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -\frac{U_d \tau_x}{(1+U_d \tau_x r_P)^2} & \frac{U_f (U_{max}-1)\tau_u}{(1+U_f \tau_u r_S)^2} & 0 \end{bmatrix},$$

$$\mathrm{diag}(\mathbf{r}) = \begin{bmatrix} r_E & 0 & 0 & 0 \\ 0 & r_P & 0 & 0 \\ 0 & 0 & r_S & 0 \\ 0 & 0 & 0 & r_V \end{bmatrix}, \ \delta\mathbf{g} = \begin{bmatrix} \delta g_E \\ \delta g_P \\ \delta g_S \\ \delta g_V \end{bmatrix}. \tag{25}$$

The fixed point solution of Eq. (24) quantifies the change in population rates $\delta\mathbf{r}$ to an input perturbation $\delta\mathbf{g}$

$$\begin{aligned} \delta\mathbf{r} &= \Big(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\Big)^{-1}\mathbf{F}\delta\mathbf{g} \\ &= \frac{1}{\det\left(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\right)}\mathrm{adj}\left(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\right)\mathbf{F}\delta\mathbf{g}, \end{aligned} \tag{26}$$

where $\mathbf{1}$ denotes the identity matrix. det and adj represent the determinant and the adjugate of the matrix, respectively.

By replacing $\delta\mathbf{g}$ with a diagonal matrix $\delta\mathbf{G}$ whose diagonal elements are elements of $\delta\mathbf{g}$, we can obtain a response matrix $\mathbf{R}$ as follows

$$\mathbf{R} = \frac{1}{\det\left(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\right)}\mathrm{adj}\left(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\right)\mathbf{F}\delta\mathbf{G} \tag{27}$$

with

$$\delta\mathbf{G} = \begin{bmatrix} \delta g_E & 0 & 0 & 0 \\ 0 & \delta g_P & 0 & 0 \\ 0 & 0 & \delta g_S & 0 \\ 0 & 0 & 0 & \delta g_V \end{bmatrix}. \tag{28}$$

Importantly, the element $\mathbf{R}_{ij}$ provides the change in the steady-state rate response of population i caused by an input perturbation $\delta\mathbf{G}_{jj}$ to population j.

We further define a scalar D and a response factor matrix $\mathbf{K}$ as follows

$$D = \frac{1}{\det(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r})))} \tag{29}$$

and

$$\mathbf{K} = \mathrm{adj}\left(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\right)\mathbf{F}. \tag{30}$$

So the response matrix $\mathbf{R}$ can be expressed as

$$\mathbf{R} = D\,\mathbf{K}\delta\mathbf{G}. \tag{31}$$

Note that if the network is stable, all eigenvalues of $\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))$ have negative real parts. Therefore, $\det(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))$, the product of the eigenvalues, is always positive (negative) if the dimension of the system is even (odd). Thus, in our case, D is always positive (Fig. S5A).

To investigate how top-down modulation via VIP affects SST response in networks with iSTP, we can apply the theoretical framework introduced above and write the change of SST activity as a function of the change of the input to VIP:

$$\mathbf{R}_{SV} = D\,\mathbf{K}_{SV}\delta g_V \tag{32}$$

with

$$\begin{aligned}\mathbf{K}_{SV} = &\left((x^*_{PP} + x^{*'}_{PP} r_P) J_{EE} J_{PP} J_{SV} - (x^*_{EP} + x^{*'}_{EP} r_P) J_{SV} J_{PE} J_{EP}\right) f'_E f'_P f'_S f'_V \\ &- (x^*_{PP} + x^{*'}_{PP} r_P) J_{PP} J_{SV} f'_P f'_S f'_V + J_{EE} J_{SV} f'_E f'_S f'_V - J_{SV} f'_S f'_V.\end{aligned} \tag{33}$$

To have response reversal, we need the change of SST activity induced by top-down modulation via VIP to switch from negative to positive as bottom-up input increases. In other words, we need a sign change of $\mathbf{R}_{SV}$ from negative to positive as the bottom-up input increases. Since D and δg_V are positive, $\mathbf{K}_{SV}$ is the only term that can switch the sign of $\mathbf{R}_{SV}$.

In the regime where all cell populations have positive firing rates, the derivatives of the rectified linear input-output functions are 1. Therefore, $\mathbf{K}_{SV}$ can be simplified to:

$$\begin{aligned}\mathbf{K}_{SV} = &((x^*_{PP} + x^{*'}_{PP} r_P) J_{EE} J_{PP} J_{SV} - (x^*_{EP} + x^{*'}_{EP} r_P) J_{SV} J_{PE} J_{EP}) \\ &- (x^*_{PP} + x^{*'}_{PP} r_P) J_{PP} J_{SV} + J_{EE} J_{SV} - J_{SV}.\end{aligned} \tag{34}$$

We can further separate $\mathbf{K}_{SV}$ into an STP-dependent part $\mathbf{K}^{\text{STP}}_{SV}$ and a non-STP-dependent part $\mathbf{K}^{\text{nonSTP}}_{SV}$:

$$\mathbf{K}_{SV} = \mathbf{K}^{\text{STP}}_{SV} + \mathbf{K}^{\text{nonSTP}}_{SV} \tag{35}$$

with

$$\mathbf{K}^{\text{STP}}_{SV} = (x^*_{PP} + x^{*'}_{PP} r_P - 1)(J_{EE} - 1) J_{PP} J_{SV} - (x^*_{EP} + x^{*'}_{EP} r_P - 1) J_{SV} J_{PE} J_{EP} \tag{36}$$

and

$$\mathbf{K}^{\text{nonSTP}}_{SV} = (J_{EE} - 1)(J_{PP} J_{SV} + J_{SV}) - J_{SV} J_{PE} J_{EP}. \tag{37}$$

The STP-dependent part $\mathbf{K}^{\text{STP}}_{SV}$ can be expressed as a sum of a PV-to-E STD-dependent term $\mathbf{K}^{\text{PED}}_{SV}$ and a PV-to-PV STD-dependent term $\mathbf{K}^{\text{PPD}}_{SV}$:

$$\mathbf{K}^{\text{STP}}_{SV} = \mathbf{K}^{\text{PPD}}_{SV} + \mathbf{K}^{\text{PED}}_{SV} \tag{38}$$

with

$$\mathbf{K}^{\text{PPD}}_{SV} = -(1 - x^*_{PP} - x^{*'}_{PP} r_P)(J_{EE} - 1) J_{PP} J_{SV} \tag{39}$$

and

$$\mathbf{K}^{\text{PED}}_{SV} = (1 - x^*_{EP} - x^{*'}_{EP} r_P) J_{SV} J_{PE} J_{EP}. \tag{40}$$

Note that

$$
\begin{aligned}
1 - x_{EP}^{*} - x_{EP}^{*'} r_P &= 1 - x_{PP}^{*} - x_{PP}^{*'} r_P \\
&= 1 - x_{EP}^{*} - \frac{U_d \tau_x}{(1 + U_d \tau_x r_P)^2} r_P \\
&= \frac{(1 + U_d \tau_x r_P)^2}{(1 + U_d \tau_x r_P)^2} - \frac{1 + U_d \tau_x r_P}{(1 + U_d \tau_x r_P)^2} - \frac{U_d \tau_x r_P}{(1 + U_d \tau_x r_P)^2} \\
&= \frac{(U_d \tau_x r_P)^2}{(1 + U_d \tau_x r_P)^2} \\
&> 0.
\end{aligned} \tag{41}
$$

Therefore, the PV-to-E STD-dependent term $\mathbf{K}_{SV}^{\mathrm{PED}}$ is always positive. Consistent with recent experimental studies (Sanzeni et al., 2020), the network is initialized in an inhibition-stabilized regime when the animal receives no stimulus in darkness, in other words, J_{EE} is greater than 1 (see section Interneuron-specific stabilization property), resulting in the PV-to-PV STD-dependent term $\mathbf{K}_{SV}^{\mathrm{PPD}}$ being always negative.

Furthermore,

$$
\begin{aligned}
\frac{d}{dr_P}(1 - x_{EP}^{*} - x_{EP}^{*'} r_P) &= \frac{d}{dr_P}(1 - x_{PP}^{*} - x_{PP}^{*'} r_P) \\
&= \frac{d}{dr_P} \frac{(U_d \tau_x r_P)^2}{(1 + U_d \tau_x r_P)^2} \\
&= \frac{2 U_d^2 \tau_x^2 r_P}{(1 + U_d \tau_x r_P)^3} \\
&> 0,
\end{aligned} \tag{42}
$$

resulting in a continuous increase (decrease) in $\mathbf{K}_{SV}^{\mathrm{PED}}$ and a continuous decrease (increase) in $\mathbf{K}_{SV}^{\mathrm{PPD}}$ with increasing (decreasing) rate of PV. Since PV activity r_P increases (decreases) with greater (smaller) bottom-up input α (Fig. S2), $\mathbf{K}_{SV}^{\mathrm{PED}}$ increases (decreases) while $\mathbf{K}_{SV}^{\mathrm{PPD}}$ decreases (increases) with greater (smaller) bottom-up input α (Fig. S5C). To have response reversal, $\mathbf{K}_{SV}$ needs to switch from negative to positive with increasing α. Therefore, $\mathbf{K}_{SV}^{\mathrm{PED}}$, rather than $\mathbf{K}_{SV}^{\mathrm{PPD}}$, is the critical short-term plasticity mechanism for the generation of response reversal.

Interneuron-specific stabilization property

The network stabilization is determined by the leading eigenvalue of the Jacobian of the system. The Jacobian of the network with STP is given by

$$\mathbf{M} = \begin{bmatrix} \frac{J_{EE}-1}{\tau_E} & -\frac{x^*_{EP}J_{EP}}{\tau_E} & -\frac{J_{ES}}{\tau_E} & -\frac{J_{EV}}{\tau_E} & -\frac{J_{EP}r_P}{\tau_E} & 0 & 0 & 0 \\ \frac{J_{PE}}{\tau_P} & \frac{-1-x^*_{PP}J_{PP}}{\tau_P} & -\frac{J_{PS}}{\tau_P} & -\frac{J_{PV}}{\tau_P} & 0 & -\frac{J_{PP}r_P}{\tau_P} & 0 & 0 \\ \frac{J_{SE}}{\tau_S} & -\frac{J_{SP}}{\tau_S} & \frac{-J_{SS}-1}{\tau_S} & -\frac{J_{SV}}{\tau_S} & 0 & 0 & 0 & 0 \\ \frac{J_{VE}}{\tau_V} & -\frac{x^*_{VP}J_{VP}}{\tau_V} & -\frac{u^*_{VS}J_{VS}}{\tau_V} & \frac{-J_{VV}-1}{\tau_V} & 0 & 0 & -\frac{J_{VP}r_P}{\tau_V} & \frac{-J_{VS}r_S}{\tau_V} \\ 0 & -U_d x^*_{EP} & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 & 0 & 0 \\ 0 & -U_d x^*_{PP} & 0 & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 & 0 \\ 0 & -U_d x^*_{VP} & 0 & 0 & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 \\ 0 & 0 & U_f(U_{\max} - u_{VS}) & 0 & 0 & 0 & 0 & -\frac{1}{\tau_u} - U_f r_S \end{bmatrix}. \tag{43}$$

A positive leading eigenvalue implies that the fixed point is unstable. In other words, transient perturbation to the system leads to a deviation from the original fixed point. In contrast, a negative leading eigenvalue suggests that the fixed point is stable. Namely, the system will return to the original fixed point after transient perturbation.

To identify the inhibition stabilization property of the network, we computed the leading eigenvalue of the Jacobian for the E subnetwork. In this case, all inhibitory populations and their related STP variable are omitted. The leading eigenvalue of the Jacobian of the E subnetwork is $\frac{J_{EE}-1}{\tau_E}$. Therefore, when J_{EE} is larger than 1, the excitatory subnetwork is unstable and the network is inhibition stabilized. In alignment with recent experimental studies showing that in darkness, when animals receive no stimulus, the network is inhibition stabilized (Sanzeni et al., 2020), we thus set J_{EE} to be larger than 1.

To identify the interneuron-specific stabilization property of the network, we compute the leading eigenvalue of the subnetwork in which a specific interneuron cell type and the corresponding STP mechanisms are omitted. If the leading eigenvalue is positive (negative), the subnetwork without this interneuron population is unstable (stable), suggesting the omitted interneuron subtype is necessary (unnecessary) for stabilization. Furthermore, the interneuron-specific stabilization property can be probed by freezing the respective population and transiently perturbing the excitatory population.

To reveal the relationship between response reversal of SST, the paradoxical effect, and the interneuron-specific stabilization, we compute the Jacobian for the E-PV-VIP subnetwork, the SST population and its related STP variable are omitted. The Jacobian of the E-PV-VIP subnetwork is

given by

$$\mathbf{M}_{E,PV,VIP} = \begin{bmatrix} \frac{J_{EE}-1}{\tau_E} & -\frac{x^*_{EP}J_{EP}}{\tau_E} & -\frac{J_{EV}}{\tau_E} & -\frac{J_{EP}r_P}{\tau_E} & 0 & 0 \\ \frac{J_{PE}}{\tau_P} & \frac{-1-x^*_{PP}J_{PP}}{\tau_P} & -\frac{J_{PV}}{\tau_P} & 0 & -\frac{J_{PP}r_P}{\tau_P} & 0 \\ \frac{J_{VE}}{\tau_V} & -\frac{x^*_{VP}J_{VP}}{\tau_V} & \frac{-J_{VV}-1}{\tau_V} & 0 & 0 & -\frac{J_{VP}r_P}{\tau_V} \\ 0 & -U_d x^*_{EP} & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 & 0 \\ 0 & -U_d x^*_{PP} & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 \\ 0 & -U_d x^*_{VP} & 0 & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P \end{bmatrix}. \tag{44}$$

Therefore, the determinant of the Jacobian is given by

$$\begin{aligned} \det(\mathbf{M}_{E,PV,VIP}) &= (\frac{J_{EE}-1}{\tau_E})(\frac{-1-x^*_{PP}J_{PP}}{\tau_P})(\frac{-1}{\tau_V})(-\frac{1}{\tau_x} - U_d r_P)(-\frac{1}{\tau_x} - U_d r_P)(-\frac{1}{\tau_x} - U_d r_P) \\ &\quad - (\frac{J_{EE}-1}{\tau_E})(-\frac{J_{PP}r_P}{\tau_P})(\frac{-1}{\tau_V})(-\frac{1}{\tau_x} - U_d r_P)(-U_d x^*_{PP})(-\frac{1}{\tau_x} - U_d r_P) \\ &\quad - (-\frac{x^*_{EP}J_{EP}}{\tau_E})(\frac{J_{PE}}{\tau_P})(\frac{-1}{\tau_V})(-\frac{1}{\tau_x} - U_d r_P)(-\frac{1}{\tau_x} - U_d r_P)(-\frac{1}{\tau_x} - U_d r_P) \\ &\quad + (-\frac{J_{EP}r_P}{\tau_E})(\frac{J_{PE}}{\tau_P})(\frac{-1}{\tau_V})(-\frac{1}{\tau_x} - U_d r_P)(-\frac{1}{\tau_x} - U_d r_P)(-\frac{1}{\tau_x} - U_d r_P) \\ &= -\frac{1}{\tau_E\tau_P\tau_V}(\frac{1}{\tau_x x^*_{PP}})^3 \Big[(x^*_{PP} + x^{*\prime}_{PP}r_P - 1)(J_{EE}-1)J_{PP} - (x^*_{EP} + x^{*\prime}_{EP}r_P - 1)J_{EP}J_{PE} \\ &\quad + (J_{EE}-1)(J_{PP}+1) - J_{PE}J_{EP}\Big] \\ &= \frac{1}{\tau_E\tau_P\tau_V}(\frac{1}{\tau_x x^*_{PP}})^3 \mathbf{K}_{SS} \\ &= -\frac{1}{\tau_E\tau_P\tau_V}(\frac{1}{\tau_x x^*_{PP}})^3 \frac{\mathbf{K}_{SV}}{J_{SV}}. \end{aligned} \tag{45}$$

Since $\det(\mathbf{M}_{E,PV,VIP})$ is the product of the eigenvalues of the Jacobian for the E-PV-VIP subnetwork, a negative $\det(\mathbf{M}_{E,PV,VIP})$ implies that the leading eigenvalue is positive, suggesting that the network without SST is unstable at the fixed point. Furthermore, this fixed point cannot be stabilized by PV, instead, SST takes charge of stabilizing the network. At the same time, a negative $\det(\mathbf{M}_{E,PV,VIP})$ also implies a negative $\mathbf{K}_{SS}$ and a positive $\mathbf{K}_{SV}$, suggesting that SST responds paradoxically, and top-down modulation via VIP increases SST activity (Fig. S8).

Simulations

Simulations were performed in Python. All differential equations were implemented by Euler integration with a time step of 0.1 ms. The simulation duration was 9 seconds for each experiment, stimulation was applied in the interval of 5 to 7 seconds. Figures depict 6 seconds of network activity following 3 seconds of relaxation after initialization. Bottom-up input α was modeled in the interval $[0, 20]$ with a step size of 0.5. Short-term plasticity variables were initially set to 1. All simulation parameters are listed in Tab. S1.

Acknowledgments

We thank Elizabeth Herbert, Shreya Lakhera, and Fabio Veneto for commenting on the manuscript, and the entire “Computation in Neural Circuits Group” for discussions. This work was supported by the European Research Council under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 804824 to J.G.), the Deutsche Forschungsgemeinschaft in the Collaborative Research Centre 1080 (project C7 to J.G.), the Max Planck Society, and a grant from the Technical University of Munich (TUM Innovation Network Neurotech) (to J.G.). Y.K. W. is supported by the Add-on Fellowship of the Joachim Herz Foundation. We acknowledge the use of BioRender to generate Figure 1.

Data Availability

The code used for model simulations is available at GitHub https://github.com/comp-neural-circuits/top-down-modulation-with-iSTP. All simulation parameters are listed in the Supplemental Table.

References

Ahmadian Y, Miller KD. What Is the Dynamical Regime of Cerebral Cortex? Neuron 2021 Nov;109(21):3373–3391. doi: https://doi.org/10.1016/j.neuron.2021.07.031.

Ahmadian Y, Rubin DB, Miller KD. Analysis of the Stabilized Supralinear Network. Neural Computation 2013 08;25(8):1994–2037. doi: https://doi.org/10.1162/NECO_a_00472.

Allen Institute for Brain Science. Synaptic Physiology - Intralaminar Connectivity and Synaptic Dynamics [dataset] 2019;Available from: https://portal.brain-map.org/explore/connectivity/synaptic-physiology.

Attinger A, Wang B, Keller GB. Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex. Cell 2017 Jun;169(7):1291–1302.e14. doi: https://doi.org/10.1016/j.cell.2017.05.023.

Bos H, Oswald AM, Doiron B. Untangling Stability and Gain Modulation in Cortical Circuits with Multiple Interneuron Classes. bioRxiv 2020 Jun; doi: https://doi.org/10.1101/2020.06.15.148114.

Buzsáki G, Draguhn A. Neuronal Oscillations in Cortical Networks. Science 2004 Jun;304(5679):1926–1929. doi: https://doi.org/10.1126/science.1099745.

Buzsáki G, Mizuseki K. The log-dynamic brain: How skewed distributions affect network operations. Nature Reviews Neuroscience 2014;15(4):264–278. doi: https://doi.org/10.1038/nrn3687.

Campagnola L, Seeman SC, Chartrand T, Kim L, Hoggarth A, Gamlin C, et al. Local Connectivity and Synaptic Dynamics in Mouse and Human Neocortex. Science 2022 Mar;375(6585):eabj5861. doi: https://doi.org/10.1126/science.abj5861.

Canto-Bustos M, Friason FK, Bassi C, Oswald AMM. Disinhibitory Circuitry Gates Associative Synaptic Plasticity in Olfactory Cortex. Journal of Neuroscience 2022;42(14):2942–2950. doi: https://doi.org/10.1523/JNEUROSCI.1369-21.2021.

Cembrowski MS, Spruston N. Heterogeneity within Classical Cell Types is the Rule: Lessons from Hippocampal Pyramidal Neurons. Nature Reviews Neuroscience 2019;20(4):193–204. doi: https://doi.org/10.1038/s41583-019-0125-5.

Dipoppa M, Ranson A, Krumin M, Pachitariu M, Carandini M, Harris KD. Vision and Locomotion Shape the Interactions between Neuron Types in Mouse Visual Cortex. Neuron 2018 May;98(3):602–615. doi: https://doi.org/10.1016/j.neuron.2018.03.037.

Ekelmans P, Kraynyukovas N, Tchumatchenko T. Targeting Operational Regimes of Interest in Recurrent Neural Networks. PLOS Computational Biology 2023 May;19(5):e1011097. doi: https://doi.org/10.1371/journal.pcbi.1011097.

Fu Y, Tucciarone JM, Espinosa JS, Sheng N, Darcy DP, Nicoll RA, et al. A Cortical Circuit for Gain Control by Behavioral State. Cell 2014 Mar;156(6):1139–1152. doi: https://doi.org/10.1016/j.cell.2014.01.050.

Garcia del Molino LC, Yang GR, Mejias JF, Wang XJ. Paradoxical Response Reversal of Top-down Modulation in Cortical Circuits with Three Interneuron Types. eLife 2017 Dec;6:e29742. doi: https://doi.org/10.7554/eLife.29742.

Garrett M, Groblewski P, Piet A, Ollerenshaw D, Najafi F, Yavorska I, et al. Stimulus novelty uncovers coding diversity in visual cortical circuits. bioRxiv 2023;p. 2023.02.14.528085. doi: https://doi.org/10.1101/2023.02.14.528085.

Garrett M, Manavi S, Roll K, Ollerenshaw DR, Groblewski PA, Ponvert ND, et al. Experience Shapes Activity Dynamics and Stimulus Coding of VIP Inhibitory Cells. eLife 2020 feb;9:e50340. doi: https://doi.org/10.7554/eLife.50340.

Hennequin G, Ahmadian Y, Rubin DB, Lengyel M, Miller KD. The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. Neuron 2018;98(4):846–860.e5. doi: https://doi.org/10.1016/j.neuron.2018.04.017.

Hertäg L, Sprekeler H. Amplifying the Redistribution of Somato-Dendritic Inhibition by the Interplay of Three Interneuron Types. PLOS Computational Biology 2019 May;15(5):e1006999. doi: https://doi.org/10.1371/journal.pcbi.1006999.

Jiang X, Shen S, Cadwell CR, Berens P, Sinz F, Ecker AS, et al. Principles of Connectivity among Morphologically Defined Cell Types in Adult Neocortex. Science 2015 Nov;350(6264):aac9462. doi: https://doi.org/10.1126/science.aac9462.

Keller AJ, Dipoppa M, Roth MM, Caudill MS, Ingrosso A, Miller KD, et al. A Disinhibitory Circuit for Contextual Modulation in Primary Visual Cortex. Neuron 2020 Dec;108(6):1181–1193.e8. doi: https://doi.org/10.1016/j.neuron.2020.11.013.

Keller GB, Bonhoeffer T, Hübener M. Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. Neuron 2012 Jun;74(5):809–815. doi: https://doi.org/10.1016/j.neuron.2012.03.040.

Kepecs A, Fishell G. Interneuron Cell Types Are Fit to Function. Nature 2014 Jan;505(7483):318–326. doi: https://doi.org/10.1038/nature12983.

Kim R, Sejnowski TJ. Strong Inhibitory Signaling Underlies Stable Temporal Dynamics and Working Memory in Spiking Neural Networks. Nature Neuroscience 2021 Jan;24(1):129–139. doi: https://doi.org/10.1038/s41593-020-00753-w.

Kuchibhotla KV, Gill JV, Lindsay GW, Papadoyannis ES, Field RE, Sten TAH, et al. Parallel Processing by Cortical Inhibition Enables Context-Dependent Behavior. Nature neuroscience 2017;20(1):62–71. doi: https://doi.org/10.1038/nn.4436.

Landau ID, Egger R, Dercksen VJ, Oberlaender M, Sompolinsky H. The Impact of Structural Heterogeneity on Excitation-Inhibition Balance in Cortical Networks. Neuron 2016;92(5):1106–1121. doi: https://doi.org/10.1016/j.neuron.2016.10.027.

Larkum ME, Nevian T, Sandler M, Polsky A, Schiller J. Synaptic Integration in Tuft Dendrites of Layer 5 Pyramidal Neurons: A New Unifying Principle. Science 2009 Aug;325(5941):756–760. doi: https://doi.org/10.1126/science.1171958.

Li N, Chen S, Guo ZV, Chen H, Huo Y, Inagaki HK, et al. Spatiotemporal constraints on optogenetic inactivation in cortical circuits. eLife 2019;8:e48622. doi: https://doi.org/10.7554/eLife.48622.

Litwin-Kumar A, Rosenbaum R, Doiron B. Inhibitory Stabilization and Visual Coding in Cortical Circuits with Multiple Interneuron Subtypes. Journal of Neurophysiology 2016 Mar;115(3):1399–1409. doi: https://doi.org/10.1152/jn.00732.2015.

London M, Häusser M. Dendritic Computation. Annual Review of Neuroscience 2005;28(1):503–532. doi: https://doi.org/10.1146/annurev.neuro.28.061604.135703.

Mahrach A, Chen G, Li N, van Vreeswijk C, Hansel D. Mechanisms Underlying the Response of Mouse Cortical Networks to Optogenetic Manipulation. eLife 2020 Jan;9:e49967. doi: https://doi.org/10.7554/eLife.49967.

Markram H, Muller E, Ramaswamy S, Reimann MW, Abdellah M, Sanchez CA, et al. Reconstruction and Simulation of Neocortical Microcircuitry. Cell 2015 Oct;163(2):456–492. doi: https://doi.org/10.1016/j.cell.2015.09.029.

Miller KD. Canonical Computations of Cerebral Cortex. Current Opinion in Neurobiology 2016 Apr;37:75–84. doi: https://doi.org/10.1016/j.conb.2016.01.008.

Miller KD, Palmigiano A. Generalized Paradoxical Effects in Excitatory/Inhibitory Networks. bioRxiv 2020 Oct; doi: https://doi.org/10.1101/2020.10.13.336727.

Mongillo G, Rumpel S, Loewenstein Y. Inhibitory Connectivity Defines the Realm of Excitatory Plasticity. Nature Neuroscience 2018 Oct;21(10):1463–1470. doi: https://doi.org/10.1038/s41593-018-0226-x.

Murphy BK, Miller KD. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. Neuron 2009;61(4):635–648. doi: https://doi.org/10.1016/j.neuron.2009.02.005.

Ozeki H, Finn IM, Schaffer ES, Miller KD, Ferster D. Inhibitory Stabilization of the Cortical Network Underlies Visual Surround Suppression. Neuron 2009 May;62(4):578–592. doi: https://doi.org/10.1016/j.neuron.2009.03.028.

Pakan JM, Lowe SC, Dylda E, Keemink SW, Currie SP, Coutts CA, et al. Behavioral-State Modulation of Inhibition Is Context-Dependent and Cell Type Specific in Mouse Visual Cortex. eLife 2016 Aug;5:e14985. doi: https://doi.org/10.7554/eLife.14985.

Palmigiano A, Fumarola F, Mossing DP, Kraynyukova N, Adesnik H, Miller KD. Common Rules Underlying Optogenetic and Behavioral Modulation of Responses in Multi-Cell-Type V1 Circuit. bioRxiv 2023; doi: https://doi.org/10.1101/2020.11.11.378729.

Peng H, Xie P, Liu L, Kuang X, Wang Y, Qu L, et al. Morphological Diversity of Single Neurons in Molecularly Defined Cell Types. Nature 2021 Oct;598(7879):174–181. doi: https://doi.org/10.1038/s41586-021-03941-1.

Pfeffer CK, Xue M, He M, Huang ZJ, Scanziani M. Inhibition of Inhibition in Visual Cortex: The Logic of Connections between Molecularly Distinct Interneurons. Nature Neuroscience 2013 Aug;16(8):1068–1076. doi: https://doi.org/10.1038/nn.3446.

Pietras B, Schmutz V, Schwalger T. Mesoscopic description of hippocampal replay and metastability in spiking neural networks with short-term plasticity. PLoS Computational Biology 2022;18(12):1–46. doi: https://doi.org/10.1371/journal.pcbi.1010809.

Poirazi P, Brannon T, Mel BW. Pyramidal Neuron as Two-Layer Neural Network. Neuron 2003 Mar;37(6):989–999. doi: https://doi.org/10.1016/S0896-6273(03)00149-1.

Richter LMA, Gjorgjieva J. A Circuit Mechanism for Independent Modulation of Excitatory and Inhibitory Firing Rates after Sensory Deprivation. Proceedings of the National Academy of Sciences 2022 Aug;119(32):e2116895119. doi: https://doi.org/10.1073/pnas.2116895119.

Roxin A, Brune N, Hansel D, Mongillo G, van Vreeswijk C. On the distribution of firing rates in networks of cortical neurons. Journal of Neuroscience 2011;31(45):16217–16226. doi: https://doi.org/10.1523/JNEUROSCI.1677-11.2011.

Sadeh S, Clopath C. Inhibitory Stabilization and Cortical Computation. Nature Reviews Neuroscience 2021 Jan;22(1):21–37. doi: https://doi.org/10.1038/s41583-020-00390-z.

Sanzeni A, Akitake B, Goldbach HC, Leedy CE, Brunel N, Histed MH. Inhibition Stabilization Is a Widespread Property of Cortical Networks. eLife 2020 Jun;9:e54875. doi: https://doi.org/10.7554/eLife.54875.

Shadlen MN, Newsome WT. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. Journal of neuroscience 1998;18(10):3870–3896. doi: https://doi.org/10.1523/JNEUROSCI.18-10-03870.1998.

Tremblay R, Lee S, Rudy B. GABAergic Interneurons in the Neocortex: From Cellular Properties to Circuits. Neuron 2016 Jul;91(2):260–292. doi: https://doi.org/10.1016/j.neuron.2016.06.033.

Tsodyks M, Pawelzik K, Markram H. Neural Networks with Dynamic Synapses. Neural Computation 1998 May;10(4):821–835. doi: https://doi.org/10.1162/089976698300017502.

Tsodyks MV, Skaggs WE, Sejnowski TJ, McNaughton BL. Paradoxical Effects of External Modulation of Inhibitory Interneurons. The Journal of Neuroscience 1997 Jun;17(11):4382–4388. doi: https://doi.org/10.1523/JNEUROSCI.17-11-04382.1997.

Tzilivaki A, Kastellakis G, Poirazi P. Challenging the Point Neuron Dogma: FS Basket Cells as 2-Stage Nonlinear Integrators. Nature Communications 2019 Aug;10(1):3664. doi: https://doi.org/10.1038/s41467-019-11537-7.

van Vreeswijk C, Sompolinsky H. Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity. Science 1996 Dec;274(5293):1724–1726. doi: https://doi.org/10.1126/science.274.5293.1724.

van Vreeswijk C, Sompolinsky H. Chaotic Balanced State in a Model of Cortical Circuits. Neural Computation 1998 Aug;10(6):1321–1371. doi: https://doi.org/10.1162/089976698300017214.

Vegué M, Roxin A. Firing Rate Distributions in Spiking Networks with Heterogeneous Connectivity. Physical Review E 2019;100(2):022208. doi: https://doi.org/10.1103/PhysRevE.100.022208.

Veit J, Hakim R, Jadi MP, Sejnowski TJ, Adesnik H. Cortical Gamma Band Synchronization through Somatostatin Interneurons. Nature Neuroscience 2017 Jul;20(7):951–959. doi: https://doi.org/10.1038/nn.4562.

Veit J, Handy G, Mossing DP, Doiron B, Adesnik H. Cortical VIP Neurons Locally Control the Gain but Globally Control the Coherence of Gamma Band Rhythms. Neuron 2022 Nov; doi: https://doi.org/10.1016/j.neuron.2022.10.036.

Wang XJ. Synaptic Reverberation Underlying Mnemonic Persistent Activity. Trends in Neurosciences 2001 Aug;24(8):455–463. doi: https://doi.org/10.1016/s0166-2236(00)01868-3.

Wang XJ, Yang GR. A disinhibitory circuit motif and flexible information routing in the brain. Current Opinion in Neurobiology 2018;49:75–83. doi: https://doi.org/10.1016/j.conb.2018.01.002.

Wu YK, Gjorgjieva J. Inhibition Stabilization and Paradoxical Effects in Recurrent Neural Networks with Short-Term Plasticity. bioRxiv 2023 May; doi: https://doi.org/10.1101/2022.12.19.520986.

Wu YK, Zenke F. Nonlinear Transient Amplification in Recurrent Neural Networks with Short-Term Plasticity. eLife 2021 Dec;10:e71263. doi: https://doi.org/10.7554/eLife.71263.

Yang GR, Murray JD, Wang XJ. A dendritic disinhibitory circuit mechanism for pathway-specific gating. Nature Communications 2016;7(May). doi: https://doi.org/10.1038/ncomms12815.

Zhang S, Xu M, Kamigaki T, Hoang Do JP, Chang WC, Jenvay S, et al. Selective Attention. Long-range and Local Circuits for Top-down Modulation of Visual Cortex Processing. Science 2014 Aug;345(6197):660–665. doi: https://doi.org/10.1126/science.1254126.

Zucker RS, Regehr WG. Short-Term Synaptic Plasticity. Annual Review of Physiology 2002;64:355–405. doi: https://doi.org/10.1146/annurev.physiol.64.092501.114547.

Supplementary Material

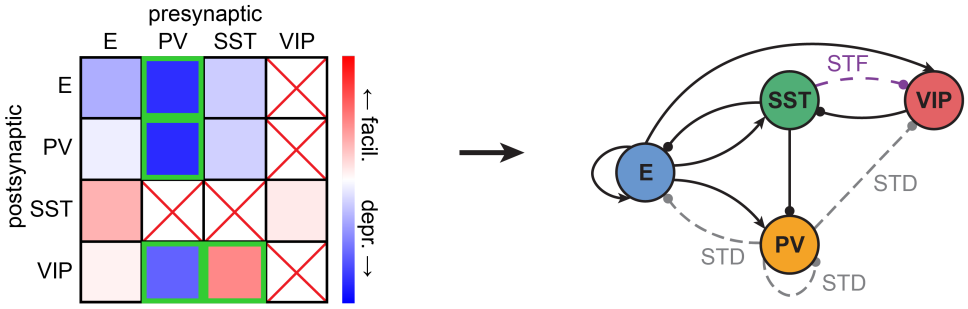

Fig. S1. Short-term plasticity mechanisms included in the network model. Left: Different degrees of short-term facilitation (STF, red) and depression (STD, blue) at different synapses measured by the Allen Institute (Campagnola et al., 2022). Green boxes indicate the four most pronounced mechanisms. Red crosses denote connections that are weak as reported in (Pfeffer et al., 2013), therefore, not considered in the model. Right: Network schematic including the four most pronounced STD (gray) and STF (purple) mechanisms.

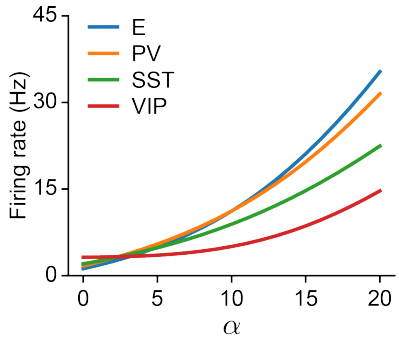

Fig. S2. Network activity as a function of bottom-up input α to E and PV. Different colors denote the activity of different populations.

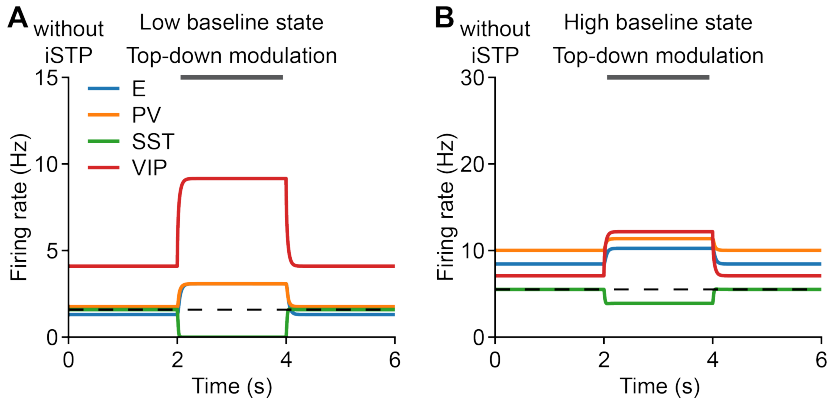

Fig. S3. In the absence of short-term plasticity mechanisms, top-down modulation via VIP cannot induce response reversal of SST. (**A**) Network responses to top-down modulation without iSTP at α = 0 corresponding to darkness without sensory stimulation. Top-down modulation via VIP is applied during the interval from 2 to 4 s (gray bar). The dashed line represents the initial activity level of SST. Here, δg_V = 3. (**B**) Same as A but at α = 15 corresponding sensory stimulation.

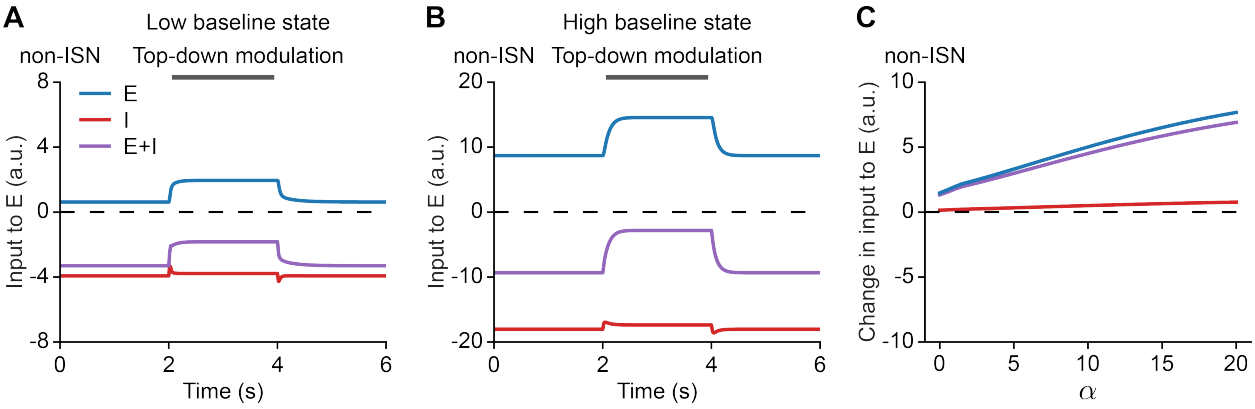

Fig. S4. Non-inhibition-stabilized networks (non-ISN) do not experience an elevation in the total inhibitory inputs to the excitatory population. (**A**) Input to the E population at α = 0 corresponding to darkness without sensory stimulation. Top-down modulation via VIP is applied during the interval from 2 to 4 s (gray bar). Different colors indicate different sources: input from the E population (blue), input from the I populations (red), and the sum of the inputs from the E and I populations (purple). (**B**) Same as A but at α = 15 corresponding to sensory stimulation. (**C**) Change in different sources of inputs to the E population measured between baseline and at steady state during top-down modulation as a function of bottom-up input α.

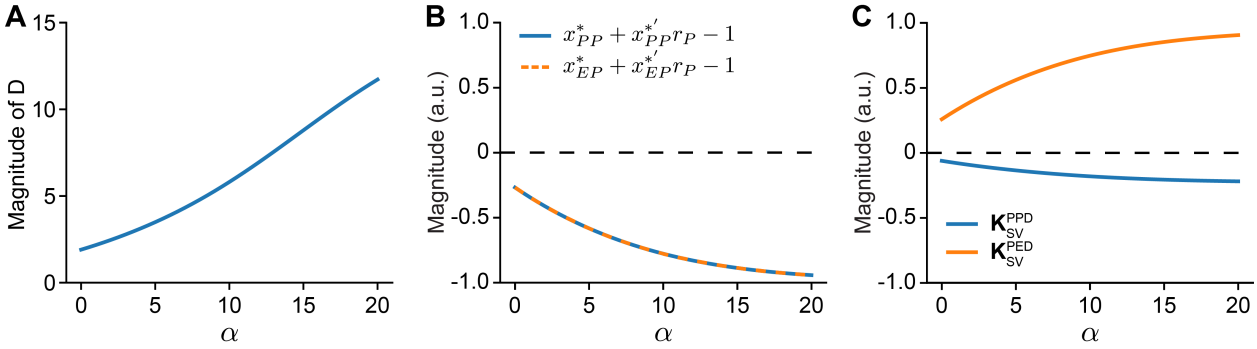

Fig. S5. Bottom-up input changes the magnitude of individual components. (**A**) Magnitude of *D* as a function of bottom-up input α. (**B**) Magnitude of PV-to-PV STD-related term (blue) and PV-to-E STD-related term (dashed orange) as a function of bottom-up input α. (**C**) PV-to-PV STD-dependent term $\mathbf{K}^{\text{PPD}}_{SV}$ (blue) and the PV-to-E STD-dependent term $\mathbf{K}^{\text{PED}}_{SV}$ (orange) as a function of bottom-up input α.

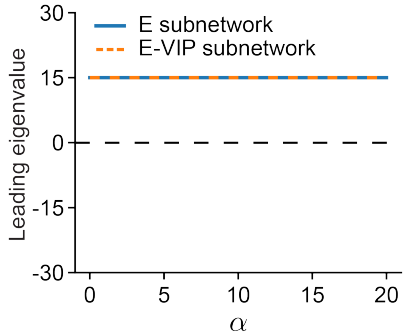

Fig. S6. Leading eigenvalue of the E subnetwork (blue) and the E-VIP subnetwork (dashed orange) subnetwork as a function of bottom-up input α.

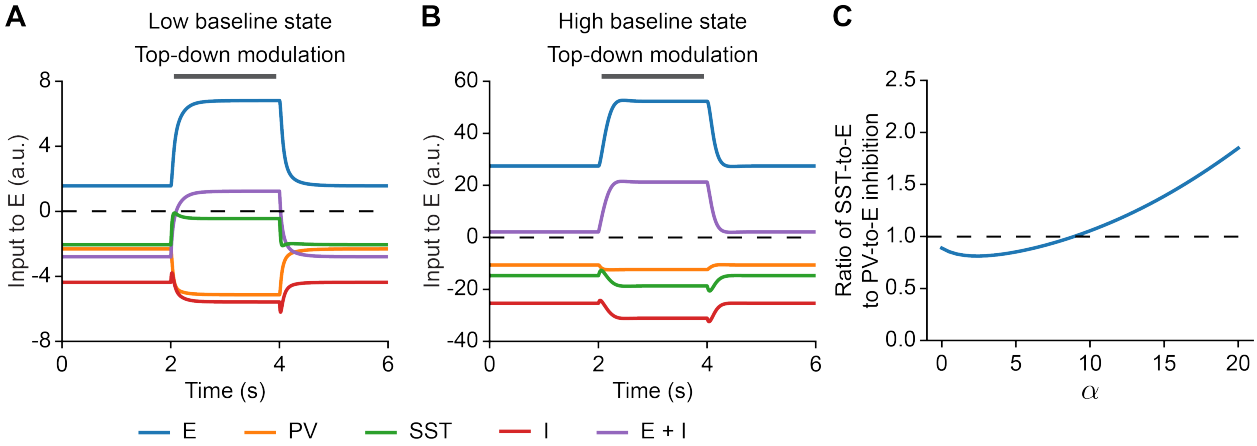

Fig. S7. Bottom-up inputs shift the inhibition received by the E population from PV-dominant to SST-dominant. (**A**) Input to the E population at α = 0 corresponding to darkness without sensory stimulation. Top-down modulation via VIP is applied during the interval from 2 to 4 s (gray bar). Different colors indicate different sources: input from the E population (blue), input from the PV population (orange), input from the SST population (green), I populations (red), and the sum of the inputs from the E and I populations (purple). (**B**) Same as A but at α = 15 corresponding to sensory stimulation. (**C**) Ratio of SST to PV inhibition received by the E population at baseline as a function of bottom-up input α.

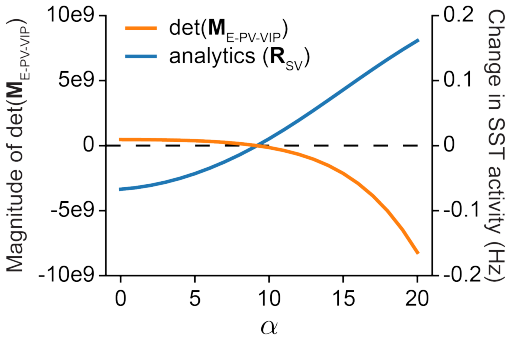

Fig. S8. Determinant of the E-PV-VIP subnetwork, det($\mathbf{M}_{\text{E-PV-VIP}}$) in orange, and change in SST response induced by top-down modulation via VIP (δg_V = 0.1), $\mathbf{R}_{SV}$ in blue, as a function of bottom-up input α.

Table S1: Parameters.

Network dynamics and network connectivity			
Symbol	**Value**	**Unit**	**Description**
τ_E	20	ms	time constant of E rate dynamics
τ_P	10	ms	time constant of PV rate dynamics
τ_S	10	ms	time constant of SST rate dynamics
τ_V	10	ms	time constant of VIP rate dynamics
J_{EE}	1.3	a.u.	connection strength from E to E
J_{EP}	1.6	a.u.	connection strength from PV to E
J_{ES}	1.0	a.u.	connection strength from SST to E
J_{PE}	1.0	a.u.	connection strength from E to PV
J_{PP}	1.3	a.u.	connection strength from PV to PV
J_{PS}	0.8	a.u.	connection strength from SST to PV
J_{SE}	0.8	a.u.	connection strength from E to SST
J_{SV}	0.6	a.u.	connection strength from VIP to SST
J_{VE}	1.1	a.u.	connection strength from E to VIP
J_{VP}	0.4	a.u.	connection strength from PV to VIP
J_{VS}	0.4	a.u.	connection strength from SST to VIP
Short-term plasticity			
τ_x	100	ms	time constant of short-term depression
U_d	1	a.u.	depression factor
τ_u	400	ms	time constant of short-term facilitation
U_f	1	a.u.	facilitation factor
$U_{\max}$	3	a.u.	maximum value of the facilitation variable
Inputs			
g_E	4	a.u.	background input to E
g_P	4	a.u.	background input to PV
g_S	3	a.u.	background input to SST
g_V	4	a.u.	background input to VIP
c	3	a.u.	top-down input to VIP