Technische Universität München
TUM School of Life Sciences

# Shaping cortical computations via short- and long-term synaptic plasticity

**Yue Wu**

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

**Vorsitz:** Prof. Dr. Mathias Wilhelm

**Prüfer*innen der Dissertation:**

1. Prof. Dr. Julijana Gjorgjieva
2. Dr. Srdjan Ostojic
3. Prof. Dr. Tilo Schwalger

Die Dissertation wurde am 31.10.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 02.01.2024 angenommen.

Yue Wu. *Shaping cortical computations via short- and long-term synaptic plasticity.* Dissertation, 2024.

# Abstract

Synapses and neuronal properties in the brain are not static; instead, they are constantly modified by various plasticity mechanisms operating at different timescales. Over the past decades, extensive experimental studies have identified and characterized these plasticity mechanisms. Yet, understanding their functional implications purely experimentally is challenging due to their complex interplay, and the complexity induced by the diversity of cell types.

In my dissertation, I take a complementary approach by combining theoretical analysis and computational modeling, and investigate how different plasticity mechanisms operating at different timescales shape cortical computations. In collaboration with the Gina Turrigiano lab, we studied how neural circuits use various homeostatic mechanisms to maintain stability in response to sensory perturbations. We found that functional correlations are subject to homeostatic regulation, and different homeostatic mechanisms can regulate distinct aspects of neural dynamics. Using analytical calculations, I demonstrated how neuronal nonlinearities and short-term plasticity affect inhibition stabilization, the paradoxical effect, and the relationship between the two. In collaboration with Dr. Friedemann Zenke, we investigated the underlying mechanism of stimulus-evoked transient neural dynamics in sensory systems, and examined the functional benefits of co-tuned inhibition and transient onset responses. In collaboration with Christoph Miehl, we reviewed the recent literature on inhibitory plasticity and its functional implications. In collaboration with Felix Waitzmann, we studied a nonlinear phenomenon in canonical cortical circuits, whereby depending on the presence of visual input, top-down modulation via VIP affects SST response oppositely, known as response reversal. We demonstrated that experimentally identified inhibitory short-term plasticity can generate response reversal of SST, and revealed the relationship between response reversal, cell-type-specific inhibition stabilization, and paradoxical effects.

Taken together, my work highlights the important role of plasticity mechanisms in shaping cortical computations.

# Zusammenfassung

Synapsen und neuronale Eigenschaften im Gehirn sind nicht statisch, sondern werden durch verschiedene Plastizitätsmechanismen, die auf unterschiedlichen Zeitskalen wirken, ständig verändert. In den letzten Jahrzehnten haben umfangreiche experimentelle Studien diese Plastizitätsmechanismen identifiziert und charakterisiert. Ihre funktionellen Auswirkungen rein experimentell zu verstehen, ist jedoch aufgrund ihres komplexen Zusammenspiels und der Komplexität, die durch die Vielfalt von Zelltypen entsteht, eine Herausforderung.

In meiner Dissertation verfolge ich einen komplementären Ansatz, indem ich theoretische Analysen und Computermodellierung kombiniere und untersuche, wie verschiedene Plastizitätsmechanismen, die auf unterschiedlichen Zeitskalen arbeiten, kortikale Berechnungen beeinflussen. In Zusammenarbeit mit dem Labor von Gina Turrigiano haben wir untersucht, wie neuronale Schaltkreise verschiedene homöostatische Mechanismen nutzen, um als Reaktion auf sensorische Störungen stabil zu bleiben. Wir fanden heraus, dass funktionale Korrelationen einer homöostatischen Regulierung unterliegen und dass verschiedene homöostatische Mechanismen unterschiedliche Aspekte der neuronalen Dynamik regulieren können. Mithilfe analytischer Berechnungen habe ich gezeigt, wie neuronale Nichtlinearitäten und kurzfristige Plastizität die Stabilisierung durch Inhibition, den paradoxen Effekt und die Beziehung zwischen beiden beeinflussen. In Zusammenarbeit mit Dr. Friedemann Zenke untersuchten wir den zugrundeliegenden Mechanismus der durch Reize ausgelösten transienten neuronalen Dynamik in sensorischen Systemen und untersuchten den funktionellen Nutzen von aufeinander abgestimmter Inhibition und der unmittelbar folgenden neuronalen Dynamik. In Zusammenarbeit mit Christoph Miehl haben wir die aktuelle Literatur zur inhibitorischen Plastizität und ihren funktionellen Implikationen untersucht. In Zusammenarbeit mit Felix Waitzmann untersuchten wir ein nichtlineares

Phänomen in kanonischen kortikalen Schaltkreisen, bei dem je nach Vorhandensein eines visuellen Inputs eine Top-down-Modulation über VIP das Verhalten der SST-Aktivität entgegengesetzt beeinflusst, was als Antwortumkehr bekannt ist. Wir haben gezeigt, dass experimentell identifizierte Kurzzeitplastizität der Inhibition eine Umkehrung der SST-Antwort bewirken kann. Darüber hinaus haben wir die Beziehung zwischen der Umkehrung der Antwort, der zelltypspezifischen Stabilisierung durch Inhibition und den paradoxen Effekten aufgezeigt. Insgesamt unterstreicht meine Arbeit die wichtige Rolle von Plastizitätsmechanismen bei der Gestaltung kortikaler Berechnungen.

# Contents

**Abstract** 3

**Zusammenfassung** 5

**1 Introduction** 11

Neuron . . . . . 11
Network connectivity and neural computations . . . . . 12
Short-term synaptic plasticity . . . . . 13
Long-term synaptic plasticity . . . . . 16
Metaplasticity . . . . . 19
Synaptic scaling . . . . . 19
Intrinsic plasticity . . . . . 20
Interneuron diversity and interneuron-specific plasticity . . . . . 22
Computational implications of plasticity mechanisms . . . . . 25

**2 Methods and mathematical framework** 29

Rate-based models and short-term plasticity mechanisms . . . . . 30
Spiking neural network models and plasticity mechanisms . . . . . 31

**3 Results** 35

3.1 Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics . . . . . 36
3.2 Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity . . . . . 37
3.3 Regulation of circuit organization and function through inhibitory synaptic plasticity . . . . . 38
3.4 Nonlinear transient amplification in recurrent neural networks with short-term plasticity . . . . . 39

3.5 Rapid and active stabilization of visual cortical firing rates across light-dark transitions . . . . . 40
3.6 Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity . . . . . 41

**4 Discussion** **43**

**Bibliography** **49**

**List of scientific communications** **61**

**Acknowledgements** **63**

**Appendix** **65**
I. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics . . . . . 65
II. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity . . . . . 78
III. Regulation of circuit organization and function through inhibitory synaptic plasticity . . . . . 88
IV. Nonlinear transient amplification in recurrent neural networks with short-term plasticity . . . . . 104
V. Rapid and active stabilization of visual cortical firing rates across light–dark transitions . . . . . 148
VI. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity . . . . . 159

# List of Figures


1.1 Schematic of a neuron . . . . . 12
1.2 Short-term depression . . . . . 15
1.3 Excitatory synaptic plasticity . . . . . 17
1.4 Inhibitory synaptic plasticity . . . . . 18
1.5 Metaplasticity . . . . . 20
1.6 Synaptic scaling . . . . . 21
1.7 Intrinsic plasticity . . . . . 21
1.8 Schematic of a canonical cortical circuit . . . . . 22
1.9 Short-term plasticity in the canonical cortical circuit with multiple interneuron subtypes . . . . . 23
1.10 Interneuron-specific synaptic plasticity . . . . . 24

# 1 Introduction

When looking at an apple, an image of the apple projects on the retina, and visual information from the retina is then relayed via the lateral geniculate nucleus of the thalamus to the primary visual cortex (Kandel, 2013; Seabrook *et al.,* 2017). When smelling fragrant flowers, odorants enter the nose and activate olfactory receptor neurons (Malnic *et al.,* 1999), after which olfactory information passed through the olfactory bulb to the olfactory cortex (Kandel, 2013; Uchida *et al.,* 2014). When listening to music, sound waves enter the ear, and auditory information is then passed via the auditory thalamus to the auditory cortex (Jones, 2012; Kandel, 2013; Lee, 2013). Different sensory information processed by different sensory cortices then propagates to other brain areas including the prefrontal cortex and the motor cortex generating decisions and behaviors (Ebbesen & Brecht, 2017; Euston *et al.,* 2012; Kandel, 2013). To perception and behavior, of particular importance is patterned neural activity in the brain generated by interacting units, called neurons.

## Neuron

Neurons are the fundamental units of the brain that receive, process, and transfer information via electrical and chemical signals. A neuron typically consists of three parts: a cell body or soma, housing the cell's nucleus and sustaining its vitality; dendrites, which are branching fibers responsible for collecting input from other cells and transmitting it to the soma; and an elongated fiber known as the axon, responsible for transmitting information from the cell body to other neurons, muscles, or glands (Fig. 1.1). Neurons transmit signals through specialized connections known as synapses, which can be categorized into two distinct types: chemical synapses and electrical synapses. At a chemical synapse, the electrical activity of the presynaptic neuron triggers the release of neurotransmitters that bind to receptors situated in the postsynaptic membrane. At an electrical synapse,

the presynaptic and postsynaptic cell membranes are connected by special channels called gap junctions that are capable of passing electric currents between neurons. Neurons can be broadly divided into excitatory neurons or inhibitory neurons. Excitatory neurons increase the action potential generation probability of postsynaptic neurons, whereas inhibitory neurons tend to decrease the likelihood of a postsynaptic action potential occurring. Excitatory neurons typically contain glutamatergic neurotransmitter, which binds to $\alpha$-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors and N-methyl-D-aspartate (NMDA) receptors in the postsynaptic membrane, whereas inhibitory neurons usually contain gamma-aminobutyric acid-ergic (GABAergic) neurotransmitter, which binds to postsynaptic $GABA_A$ receptors.

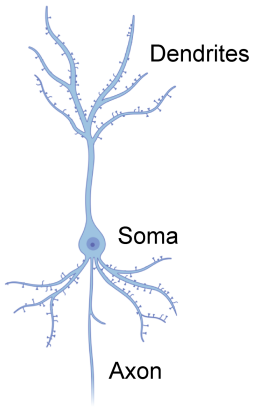


**Figure 1.1: Schematic of a neuron.** A neuron typically is composed of three parts: a cell body or soma, dendrites, and an axon. Figure created with BioRender.com.

## Network connectivity and neural computations

To generate appropriate neural activity that enables cognitive functions and flexible behaviors, connections between neurons have to be set up properly. Proper network connectivity is essential to achieve the balance of excitation and inhibition (E/I balance), and to generate asynchronous irregular activity, a hallmark of spontaneous cortical activity (Renart *et al.,* 2010; van Vreeswijk & Sompolinsky, 1996; van Vreeswijk & Sompolinsky, 1998). Networks with improper connectivity tend to exhibit abnormal neural activity associated with various brain disorders (Monday *et al.,* 2023).

Network connectivity exerts a great influence on the ability of neural circuits to

carry out different computations. For instance, strong excitatory to excitatory connection strength can generate persistent activity which is considered to underlie working memory (Wong & Wang, 2006). Multisynaptic connectivity motifs that mediate reciprocal inhibition between excitatory neurons with similar tuning enable the olfactory bulb to perform 'whitening', a fundamental computation that decorrelates activity patterns and facilitates pattern classification (Wanner & Friedrich, 2020).

The establishment of appropriate synaptic connectivity in neural circuits is rather complex. It involves activity-independent processes determined by genetic programs like axon guidance during early development (McLaughlin & O'Leary, 2005), as well as activity-dependent processes like activity-dependent plasticity (Thompson *et al.,* 2017). Multiple activity-dependent plasticity mechanisms operating in the brain greatly shape synaptic connections and neuronal properties at different timescales. On a timescale from hundreds of milliseconds to seconds, synapses can change due to short-term plasticity (Regehr, 2012; Tsodyks & Markram, 1997). On a timescale from minutes to hours, synapses are modified by long-term plasticity, which is considered as the neural substrate of learning and memory (Bi & Poo, 1998; D'amour & Froemke, 2015; Martin *et al.,* 2000; Sjöström *et al.,* 2001). On an even slower timescale from hours to days, there are slow homeostatic mechanisms like synaptic scaling and intrinsic plasticity (Desai *et al.,* 1999; Turrigiano *et al.,* 1998).

## Short-term synaptic plasticity

Repetitive presynaptic stimulation can lead to changes in synaptic strength occurring on a timescale from hundreds of milliseconds to seconds, termed short-term synaptic plasticity (Markram *et al.,* 2015; Regehr, 2012; Tsodyks & Markram, 1997). Experimentally, short-term synaptic plasticity is measured by paired-pulse ratio (PPR), which is defined as the ratio of the amplitude of the second response to that of the first. A PPR smaller than 1 indicates short-term depression, whereas a PPR larger than 1 suggests short-term facilitation. Mechanistically, short-term plasticity can involve both presynaptic and postsynaptic changes (Regehr, 2012). Since

studies on short-term plasticity primarily focused on the presynaptic factors, an overview of the major presynaptic factors is provided below.

## Short-term depression

Short-term depression can be attributed to several presynaptic factors including vesicle depletion, inactivation of release sites and calcium channels (Fioravante & Regehr, 2011; Regehr, 2012; Zucker & Regehr, 2002).

### Vesicle depletion

Typically, at small central nervous system synapses, there are hundreds of vesicles in a presynaptic terminal (Rizzoli & Betz, 2005). They reside in three different pools: the readily releasable pool, the recycling pool, and the reserve pool (Fig. 1.2; Rizzoli & Betz, 2005). The readily releasable pool consists of synaptic vesicles that are immediately available on presynaptic stimulation and typically about less than 5% of all vesicles. The recycling pool is a pool of vesicles which recycle upon moderate stimulation, and about 15% of the total vesicles. The reserve pool contains approximately 80% of all vesicles. The vesicles in the reserve pool are reluctant to release, and their release is triggered only upon intense stimulation. Vesicles in the readily releasable pool associate with a synaptic active zone that is the principal site of neurotransmitter release (Rizzoli & Betz, 2005). Depending on the size of the readily releasable pool of vesicles, however, usually only a small fraction of these vesicles are available for immediate release by an action potential (Rizzoli & Betz, 2005). Since the number of vesicles in the readily releasable pool is limited, if a large fraction of vesicles in the readily releasable pool is released by an action potential, fewer vesicles will get released by the subsequent action potential (Fig. 1.2). The depletion of vesicles caused by action potentials can lead to short-term synaptic depression until vesicles from a recycling pool replenish the readily releasable pool (Fioravante & Regehr, 2011; Zucker & Regehr, 2002).

### Inactivation of release sites

For the release of neurotransmitter, a synaptic vesicle first fuses with the plasma membrane of the presynaptic axon terminal and then releases its contents into the synaptic cleft, a process called exocytosis (Stevens, 2003). Synaptic vesicle fusion at a release site can inhibit the occurrence of subsequent fusion events at that site (Neher & Sakaba, 2008). The release site inactivation can take for seconds follow-

ing exocytosis and lead to short-term synaptic depression (Fioravante & Regehr, 2011; Neher & Sakaba, 2008). Furthermore, short-term depression is also affected by endocytosis, a process by which the plasma membrane at the presynaptic axon terminal invaginates to form synaptic vesicles (Hosoi *et al.,* 2009). More specifically, blocking endocytosis enhances short-term synaptic depression (Hosoi *et al.,* 2009; Hua *et al.,* 2013).

### Inactivation of calcium channels

At the calyx of Held, a brainstem giant synapse, inactivation of calcium channels results in a decrease in calcium influx, contributing to short-term depression (Forsythe *et al.,* 1998; Xu & Wu, 2005).

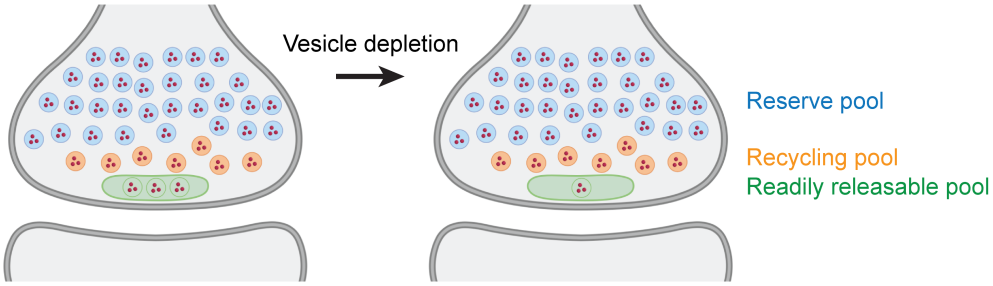


**Figure 1.2: Short-term depression.** Synaptic vesciles reside in three different pools: the reserve pool (blue), the recycling pool (orange), and the readily releasable pool (green). Presynaptic action potentials trigger the release of vesicles in the readily releasable pool, resulting in vesicle depletion and less vesicles in the readily releasable pool, therefore, short-term depression. Figure adapted from (Rizzoli & Betz, 2005).

## Short-term facilitation

Short-term facilitation is thought to depend on increased calcium at the presynaptic axon terminal caused by the presynaptic action potential. The increased calcium level will increase the release probability of neurotransmitters, therefore leading to short-term facilitation (Fioravante & Regehr, 2011; Regehr, 2012). More specifically, a presynaptic action potential initiates a local calcium signal that triggers the release of neurotransmitter. Despite being at a low level, calcium persists within the presynaptic bouton. If the residual calcium signal constitutes a substantial portion of the local calcium signal responsible for driving the release, the residual calcium can enhance the release probability of neurotransmitters and thus lead to short-term facilitation (Katz & Miledi, 1968; Regehr, 2012). Furthermore, short-term facili-

tation can arise due to the local saturation of rapid endogenous $Ca^{2+}$ buffers within the terminal during a series of action potentials, therefore leading to a gradual elevation in the $Ca^{2+}$ concentration at the release site (Blatow *et al.,* 2003; Rozov *et al.,* 2001).

## Long-term synaptic plasticity

On a timescale from minutes to hours, synaptic strength can undergo potentiation or depression as a result of long-term synaptic plasticity (Martin *et al.,* 2000). Long-term synaptic plasticity is widely considered to be the neural substrate of learning and memory (Lamprecht & LeDoux, 2004; Martin *et al.,* 2000). It was postulated by Donald Hebb that coactivation of presynaptic and postsynaptic neurons causes long-term plasticity (Hebb, 1949). Over the last three decades, extensive experimental studies have been conducted to quantify how plasticity at different types of synapses depends on different features of neural activity including firing rates and spike timing (Bi & Poo, 1998; Caporale & Dan, 2008; D'amour & Froemke, 2015; Kirkwood *et al.,* 1993; Sjöström *et al.,* 2001).

### Excitatory synaptic plasticity

Early studies have examined the dependency of synaptic plasticity at excitatory synapses on afferent stimulation rates, and have shown that low stimulation rates lead to depression whereas high stimulation rates cause potentiation (Fig. 1.3A; Dudek & Bear, 1992; Kirkwood *et al.,* 1993). Later, the impact of spike timing on synaptic plasticity has been investigated by pairing presynaptic spikes with postsynaptic spikes within a time window of tens of milliseconds (Bi & Poo, 1998). In hippocampus, long-term potentiation occurs when presynaptic spikes precede postsynaptic spikes, whereas long-term depression occurs when presynaptic spikes follow postsynaptic spikes (Fig. 1.3B; Bi & Poo, 1998). Similar spike timing-dependent plasticity (STDP) rules for excitatory synapses have also been observed in the cortex (Feldman, 2000; Markram *et al.,* 1997; Sjöström *et al.,* 2001). Biophysically, the induction of long-term excitatory plasticity normally includes the activation of NMDA receptors (Lüscher & Malenka, 2012). The expression of excitatory plasticity typically involves modifications of the number and the properties of AMPA

receptors (Diering & Huganir, 2018; Huganir & Nicoll, 2013; Makino & Malinow, 2009; Malinow & Malenka, 2002).

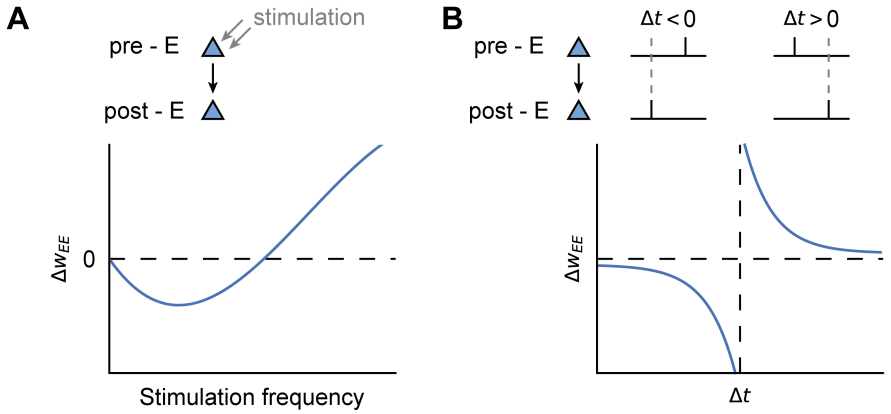


**Figure 1.3: Excitatory synaptic plasticity. A.** Change in excitatory synaptic strength ($\Delta w_{EE}$) as a function of afferent stimulation frequency. Low stimulation rates produce depression ($\Delta w_{EE} < 0$) whereas high stimulation rates cause potentiation ($\Delta w_{EE} > 0$). Figure adapted from (Kirkwood *et al.,* 1993). **B.** Change in excitatory synaptic strength as a function of the timing difference between pre- and postsynaptic spikes ($\Delta t$). Long-term potentiation ($\Delta w_{EE} > 0$) takes place when presynaptic spikes precede postsynaptic spikes ($\Delta t > 0$), whereas long-term depression ($\Delta w_{EE} < 0$) occurs when presynaptic spikes succeed postsynaptic spikes ($\Delta t < 0$). Figure adapted from (Bi & Poo, 1998).

## Inhibitory synaptic plasticity

Accumulating evidence has also demonstrated long-term plasticity at inhibitory synapses (Capogna *et al.,* 2021; Castillo *et al.,* 2011; McFarlan *et al.,* 2023). In contrast to excitatory synaptic plasticity, inhibitory synaptic plasticity is more diverse. Different types of inhibitory STDP curves have been reported in different brain regions (D'amour & Froemke, 2015; Haas *et al.,* 2006; Vickers *et al.,* 2018; Woodin *et al.,* 2003). For instance, irrespective of the temporal order of pre- and postsynaptic spikes, coincident pre- and postsynaptic activity leads to potentiation at inhibitory synapses onto layer 5 pyramidal neurons in the auditory cortex (Fig. 1.4A; D'amour & Froemke, 2015) but leads to depression at inhibitory synapses onto layer 4 principal neurons in the auditory cortex (Fig. 1.4B; Vickers *et al.,* 2018). In some other brain regions, the temporal order of the pre- and postsynaptic spikes is however crucial, and determines the sign of plasticity at inhibitory synapses onto excitatory neurons. For instance, in the entorhinal cortex, pre-before-post pairing produces potentiation whereas post-before-pre pairing induces depression (Fig. 1.4C; Haas *et al.,* 2006). In contrast, in the hippocampus, to obtain the same effects in the

change of inhibitory synapses, the orders of pre- and postsynaptic spikes are reversed (Fig. 1.4D; Woodin *et al.,* 2003). The induction of long-term inhibitory plasticity is commonly mediated by retrograde signaling following repetitive activation of nearby excitatory synapses (Capogna *et al.,* 2021). The expression of long-term inhibitory plasticity involves changes in the presynaptic GABA release, and modifications of the number and the function of postsynaptic $GABA_A$ receptors (Chiu *et al.,* 2019; Luscher *et al.,* 2011).

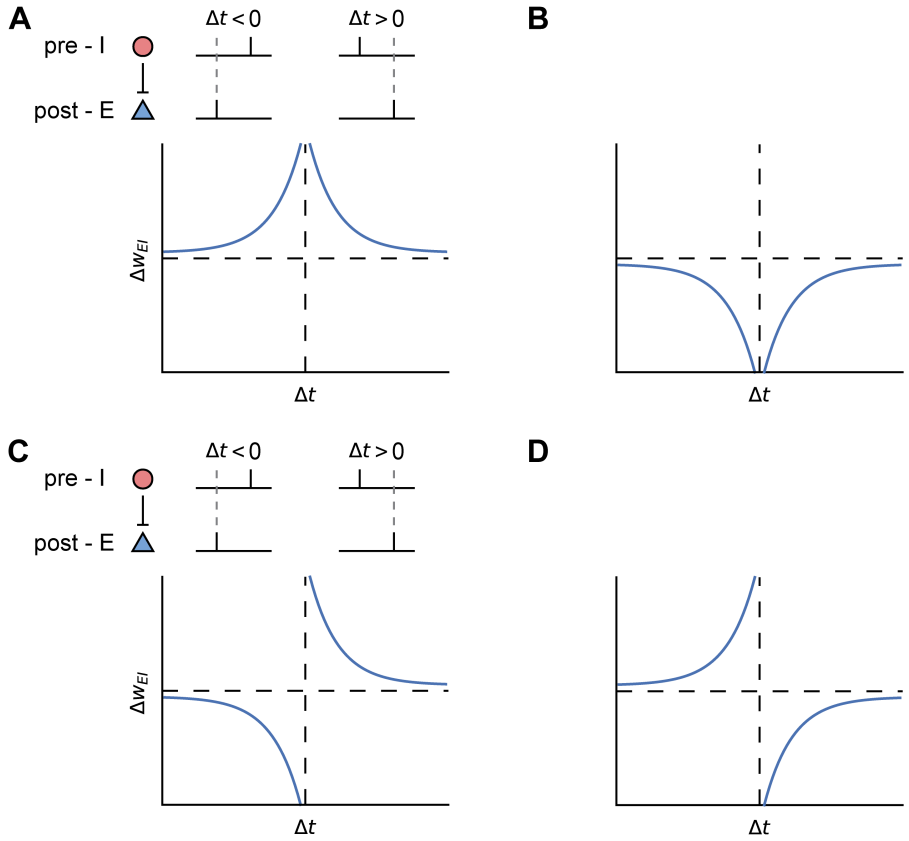


**Figure 1.4: Inhibitory synaptic plasticity.** Different spike timing dependent plasticity curves for inhibitory synapses. **A.** Symmetric Hebbian learning rule characterized by a symmetric function of the difference in spike times of pre- and postsynaptic neurons, and potentiation induced by coincident pre- and postsynaptic activity (D'amour & Froemke, 2015). **B.** Symmetric anti-Hebbian learning rule in which depression induced by coincident pre- and postsynaptic activity (Vickers *et al.,* 2018). **C.** Asymmetric Hebbian learning rule characterized by an asymmetric function of the difference in spike times of pre- and postsynaptic neurons, and potentiation induced by pre-before-post pairing (Haas *et al.,* 2006). **D.** Asymmetric anti-Hebbian learning rule in which depression induced by pre-before-post pairing (Woodin *et al.,* 2003). Figure adapted from (Wu *et al.,* 2022).

### Heterosynaptic plasticity

Pairing pre- and postsynaptic activity does not only induce plastic changes in synapses between paired neurons, but also plasticity of nearby synapses associated with unpaired presynaptic neurons, known as heterosynaptic plasticity (Chater & Goda, 2021; Chistiakova *et al.,* 2015; Field *et al.,* 2020; Lynch *et al.,* 1977; Oh *et al.,* 2015; Royer & Paré, 2003). Compared to homosynaptic Hebbian-type plasticity, heterosynaptic plasticity imposes an opposite effect on changes in synaptic weights, suggesting a homeostatic role of heterosynaptic plasticity in synaptic weight stabilization (Chistiakova *et al.,* 2015). Heterosynaptic plasticity has been observed at both excitatory and inhibitory synapses (Field *et al.,* 2020).

## Metaplasticity

Synaptic activity can change synaptic strengths due to synaptic plasticity. Synaptic plasticity can be also affected by prior synaptic activity, a phenomenon known as metaplasticity (Abraham, 2008; Abraham & Bear, 1996). For instance, the threshold for LTP and LTD at which stimulation frequency leads to no change in synaptic strength can vary depending on the history of synaptic activity. Direct experimental evidence has been obtained by comparing the learning curves in the visual cortex of rats in the dark-rearing condition with that in the control condition (Kirkwood *et al.,* 1996). In the dark-rearing condition, LTP is enhanced, LTD is suppressed, and the LTP/LTD threshold shifts to a lower frequency, indicating that a stimulation frequency normally induces LTD now can induce LTP instead (Fig. 1.5; Kirkwood *et al.,* 1996).

## Synaptic scaling

To maintain proper functioning of neural circuits when confronted with external perturbations, synapses can regulate neural dynamics by upscaling or downscaling synaptic strengths, known as synaptic scaling (Fig. 1.6; Turrigiano, 2011; Turrigiano *et al.,* 1998). In comparison with long-term plasticity, synaptic scaling is relatively slow and operating on a timescale from hours to days (Turrigiano *et al.,* 1998). In culture, abolishing neural activity using tetrodotoxin (TTX) for 48

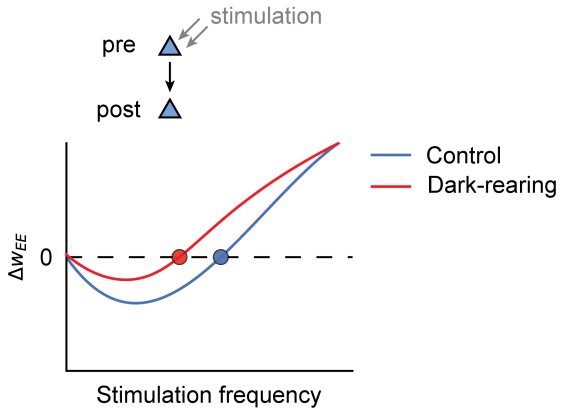


**Figure 1.5: Metaplasticity.** Metaplasticity leads to different plasticity curves in the control condition (blue) and in the dark-rearing condition (red). Compared to the control condition, the LTP/LTD threshold marked by circles in the dark-reading condition shifts to the left, making LTP induction easier and LTD induction harder. Figure adapted from (Kirkwood *et al.,* 1996).

hours leads to upscaling of excitatory synapses, therefore, an increase in the amplitude of miniature excitatory postsynaptic currents (mEPSCs) (Turrigiano *et al.,* 1998). Overexcited neural activity induced by applying bicucullin, a $GABA_A$ receptor antagonist, results in downscaling of excitatory synapses, thus, a decrease of mEPSCs (Turrigiano *et al.,* 1998). As synaptic scaling attempts to counteract deviations from normal activity level of neurons, it has been considered as one of the important homeostatic mechanisms (Desai *et al.,* 2002; Ibata *et al.,* 2008; Keck *et al.,* 2013). Furthermore, the scaling of synaptic strength is found to be multiplicative (Turrigiano *et al.,* 1998). In other words, originally stronger synapses undergo more substantial changes than initially weaker synapses. Multiplicative synaptic scaling can preserve the relative differences in synaptic strengths. Mechanistically, synaptic scaling is realized by regulating AMPA receptors at the postsynaptic sites (Turrigiano, 2008).

## Intrinsic plasticity

In addition to synaptic homeostatic mechanisms, neurons can bidirectionally adjust intrinsic excitability to regulate their activity (Daoudal & Debanne, 2003; Debanne *et al.,* 2019; Desai *et al.,* 1999). The change in excitability is measured experimentally by the frequency-current curve (f-I curve). To counteract the decreased activity, neurons can increase their excitability, resulting in a higher firing compared to the

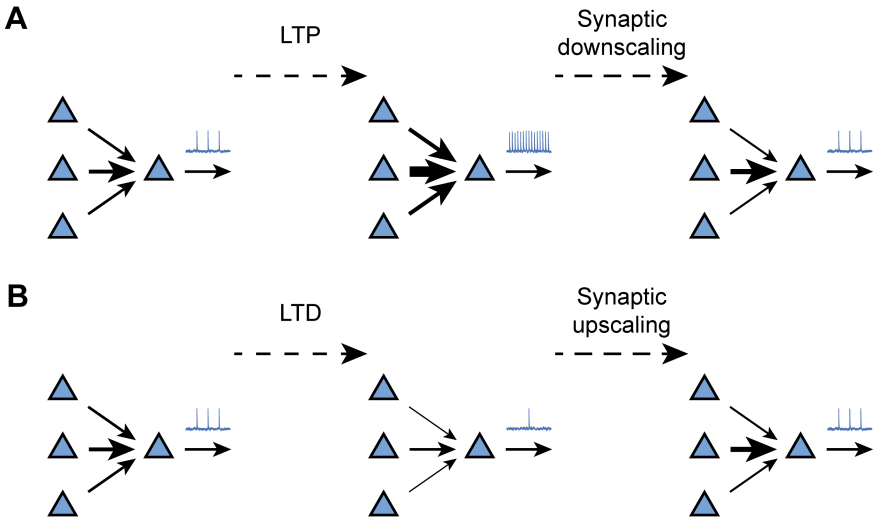


**Figure 1.6: Synaptic scaling. A.** Synaptic downscaling compensates for the overexcited activity induced by LTP. **B.** Synaptic upscaling counteracts the reduced activity induced by LTD. Note that synaptic strengths are scaled in a multiplicative manner, so that the relative strength of the synapses remains unchanged. Figure adapted from (Turrigiano, 2008).

control condition when injecting the same amount of current (Fig. 1.7; Desai *et al.,* 1999). To compensate for increased activity, neurons decrease their excitability. Consequently, the same amount of current leads to a lower firing than that in the control condition (Fig. 1.7; Fan *et al.,* 2005). Biophysically, intrinsic plasticity involves changes in axon initial segments (Grubb & Burrone, 2010; Jamann *et al.,* 2021).

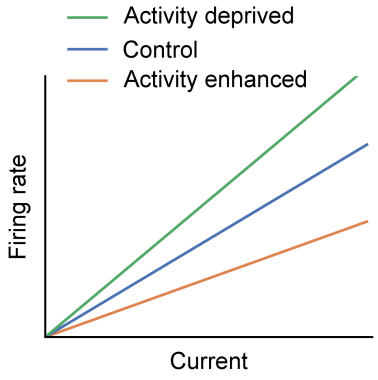


**Figure 1.7: Intrinsic plasticity.** Neuronal excitability is typically measured by the frequency-current curve (f-I curve). Increased excitability indicated by a higher firing rate in response to the same amount of current injection compensates for decreased activity, whereas decreased excitability counteracts overexcited activity. Figure adapted from (Desai *et al.,* 1999; Fan *et al.,* 2005).

## Interneuron diversity and interneuron-specific plasticity

The complexity of neural circuits does not only arise from various plasticity mechanisms interacting with each other, but also from the enormous diversity of cell types. In contrast to excitatory neurons, inhibitory interneurons are highly diverse in terms of anatomy, electrophysiology and functions (Jiang *et al.,* 2015; Tremblay *et al.,* 2016). In the mouse neocortex, 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). Distinct interneuron subtypes selectively innervate specific compartments of pyramidal cells. PV neurons tend to target perisomatic regions of pyramidal neurons, while SST neurons preferentially target distal dendritic regions of pyramidal neurons. (Tremblay *et al.,* 2016). Together with excitatory neurons, they form a canonical microcircuit. Previous experimental studies have identified common features of the connectivity structure of this circuit (Fig. 1.8). For instance, SST and VIP mutually inhibit each other, and inhibitory connections between SST, and between VIP are rare (Pfeffer *et al.,* 2013). Diverse interneuron subtypes are relevant for various computations and cognitive functions, such as locomotion-induced gain modulation (Fu *et al.,* 2014), selective attention (Zhang *et al.,* 2014), context-dependent modulation (Keller *et al.,* 2020; Kuchibhotla *et al.,* 2017), predictive processing (Attinger *et al.,* 2017; Keller & Mrsic-Flogel, 2018), novelty detection (Garrett *et al.,* 2020), regulating global coherence (Veit *et al.,* 2017, 2022), and gating of synaptic plasticity (Canto-Bustos *et al.,* 2022; Krabbe *et al.,* 2019; Williams & Holtmaat, 2019).

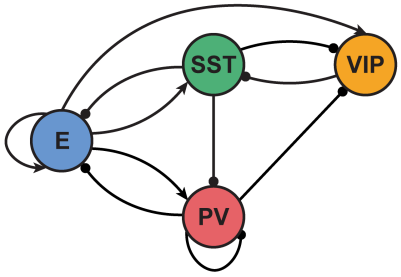


**Figure 1.8: Schematic of a canonical cortical circuit.** The canonical cortical circuit consists of one excitatory (E, blue) population and three distinct inhibitory populations including PV (orange), SST (green), and VIP (red). Network connectivity and figure adapted from (Pfeffer *et al.,* 2013).

## Interneuron-specific short-term plasticity

Synapses between different types of neurons can exhibit distinct short-term dynamics. Quantitative measurements of cell-type specific short-term plasticity have recently been conducted by the Allen Institute (Campagnola *et al.,* 2022). Several types of synapses exhibit significant short-term dynamics (Fig. 1.9). For instance, synapses from PV to E, from PV to PV, and from PV to VIP undergo considerable short-term depression, whereas synapses from SST to VIP display pronounced short-term facilitation.

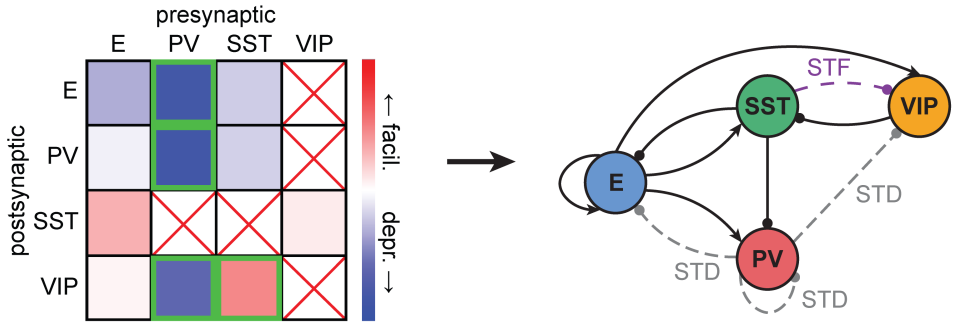


**Figure 1.9: Short-term plasticity in the canonical cortical circuit with multiple interneuron subtypes.** Left: Different degrees of short-term facilitation (STF, red) and depression (STD, blue) at different synapses measured by the Allen Institute (Campagnola *et al.,* 2022). Green boxes denote the four most pronounced short-term dynamic connections. Red crosses indicate connections that are weak as reported in (Pfeffer *et al.,* 2013). Right: Network schematic incorporating the four most pronounced STD (gray) and STF (purple) mechanisms. Figure adapted from (Campagnola *et al.,* 2022; Pfeffer *et al.,* 2013; Waitzmann *et al.,* 2023).

## Interneuron-specific long-term plasticity

Synapses from different inhibitory cell types also exhibit different long-term plasticity (Lagzi *et al.,* 2021; Song *et al.,* 2022; Udakis *et al.,* 2020; Yap *et al.,* 2021). Recent studies have examined spike timing-dependent plasticity of inhibitory synapses from different types of interneurons in hippocampus (Udakis *et al.,* 2020). By pairing single IPSCs with a burst of action potentials of excitatory neurons, it has been revealed that PV synapses onto excitatory neurons in CA1 only undergo long-term depression whereas SST synapses can undergo both long-term depression and long-term potentiation depending on the precise spike timing of pre- and postsynaptic spikes (Fig. 1.10A). More specifically, small intervals between presynaptic and postsynaptic spikes lead to long-term potentiation of SST synapses whereas

large intervals result in long-term depression. In the layer 2/3 of the orbitofrontal cortex, PV and SST exhibit different forms of spike timing-dependent plasticity (Fig. 1.10B; Lagzi *et al.,* 2021). Regardless of the temporal order of pre- and postsynaptic spikes, PV synapses to excitatory neurons get potentiated by small intervals between presynaptic and postsynaptic spikes and get depressed by large intervals. In contrast, for SST synapses to excitatory neurons, pre-before-post pairing produces potentiation whereas post-before-pre pairing induces depression.

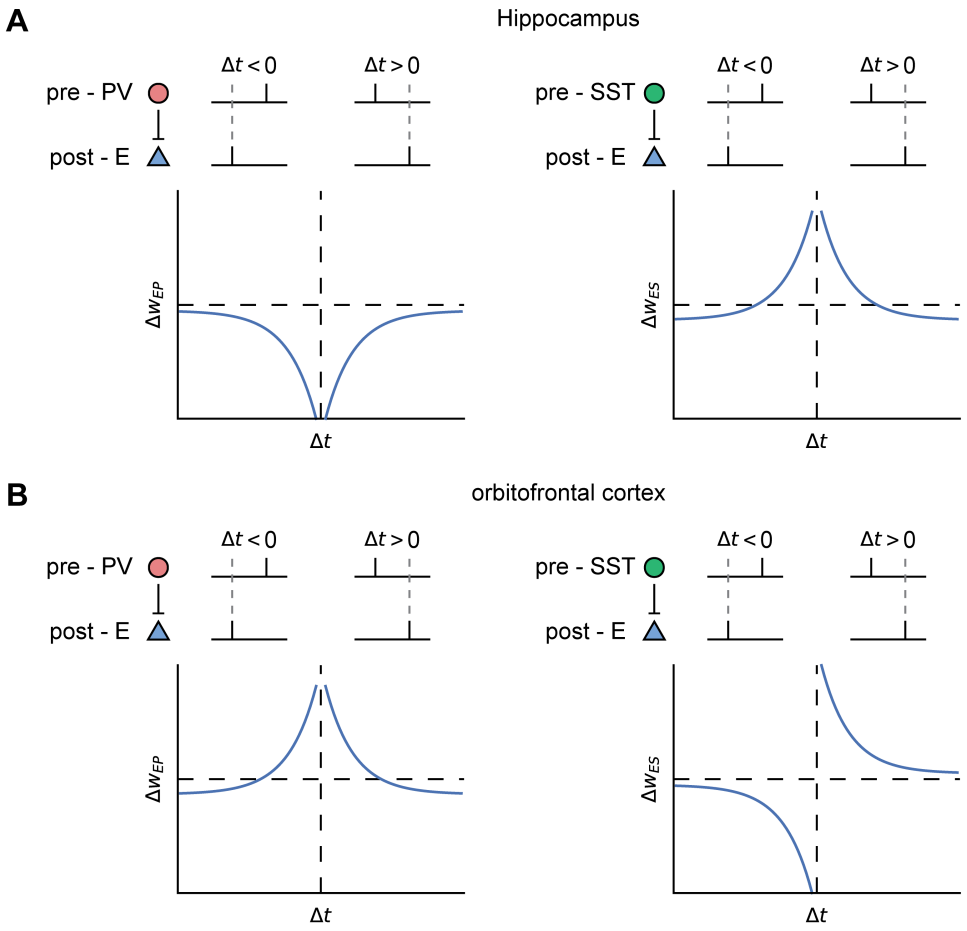


**Figure 1.10: Interneuron-specific synaptic plasticity.** Different spike timing dependent plasticity curves for different types of inhibitory synapses in different brain regions. **A.** Symmetric anti-Hebbian learning rule for PV-to-E synapses whereas symmetric Hebbian learning rule for SST-to-E synapses observed in hippocampus. Figure adapted from (Udakis *et al.,* 2020). **B.** Symmetric Hebbian learning rule for PV-to-E synapses whereas asymmetric Hebbian learning rule for SST-to-E synapses observed in the orbitofrontal cortex. Figure adapted from (Lagzi *et al.,* 2021).

## Computational implications of plasticity mechanisms

Based on experimental measurements of various plasticity mechanisms, theorists have built phenomenological models to investigate the functional implications of plasticity on neural computations (Abbott & Regehr, 2004; Abbott *et al.,* 1997; Clopath *et al.,* 2010; Vogels *et al.,* 2011).

Computational studies have shown that recurrent networks equipped with short-term depression are able to carry out complex computations on time varying inputs and generate temporally synchronous activity, resembling information processing in the auditory cortex (Loebel *et al.,* 2007; Loebel & Tsodyks, 2002). Inspired by the significant short-term facilitation of excitatory synapses observed in the prefrontal cortex (Wang *et al.,* 2006), theoretical work has suggested that working memory can be maintained via short-term synaptic facilitation (Mongillo *et al.,* 2008).

Multiple modeling studies have investigated how different long-term plasticity mechanisms interact with each other to form certain network structures like assemblies (Litwin-Kumar & Doiron, 2014; Miehl & Gjorgjieva, 2022; Zenke *et al.,* 2015) and chain-like structures (Maes *et al.,* 2021, 2020). Specific network structures enable networks to carry out particular computations. For instance, assemblies, a group of strongly interconnected neurons, are thought to underlie the encoding of memories (Buzsáki, 2010). Networks with chain-like structures can generate sequences resembling hippocampal replay that is important for memory consolidation (Carr *et al.,* 2011). Importantly, in plastic networks, inhibitory plasticity plays an important role in the establishment of E/I balance (Miehl & Gjorgjieva, 2022; Sprekeler, 2017; Vogels *et al.,* 2011).

Multiple theoretical studies have investigated plasticity mechanisms operating at a slow timescale. The classical model of metaplasticity is Bienenstock–Cooper–Munro (BCM) synaptic learning rule (Bienenstock *et al.,* 1982). The model has suggested that metaplasticity can serve as a homeostatic mechanism to stabilize weight dynamics. Several computational studies have examined the functional implications on synaptic scaling, and demonstrated that slow synaptic scaling mechanisms may be important for memory stabilization (Tetzlaff *et al.,* 2013), and the allocation of multiple memory representations without interference

(Auth *et al.,* 2020). In addition, intrinsic excitability mechanisms can contribute to the formation of neuronal ensembles (Alejandre-García *et al.,* 2022).

In this dissertation, I investigate how different plasticity mechanisms operating at different timescales shape cortical computations. In particular, I introduce four distinct projects on

1. homeostatic regulation of cortical circuit dynamics (Wu *et al.,* 2020),

2. the effect of short-term plasticity on inhibition stabilization and paradoxical effects in recurrent neural networks (Wu & Gjorgjieva, 2023),

3. rapid sensory processing in recurrent neural networks with short-term plasticity (Wu & Zenke, 2021),

4. the impact of inhibitory short-term plasticity on top-down modulation in canonical cortical circuits (Waitzmann *et al.,* 2023).

To highlight commonalities among these projects, I first introduce the mathematical frameworks applied throughout. Finally, I discuss several future research directions related to this topic. Taken together, the work in my dissertation provides mechanistic insight on how different plasticity mechanisms shape cortical computations by organizing network connectivity, and makes concrete experimentally testable predictions.

# 2 Methods and mathematical framework

In computational neuroscience, models can be built with different levels of detail to answer different kinds of research questions. For instance, single neuron models with complex morphologies and multiple ion channels allow us to investigate how single neurons integrate their inputs and the computational power of single neurons (Beniaguev *et al.,* 2021; Poirazi *et al.,* 2003). Hodgkin–Huxley neuron models enable us to study how different types of ion channels affect the firing properties of single neurons (Hodgkin & Huxley, 1952; Izhikevich, 2007). In my PhD, to investigate how different plasticity mechanisms shape neural dynamics and computations, I mainly built two types of models: rate-based population models and spiking neural network models. Despite ignoring biophysical details of single neurons, rate-based population models allow us to study how network connectivity modified by plasticity mechanisms affects the emergence of network behaviors in a broad parameter range and how different parameters affect the emergent features systematically. In contrast, spiking neural network models contain more realistic spiking neurons and allow us to study functional consequences of spike-timing dependent plasticity mechanisms and more direct comparisons with electrophysiological data.

## Rate-based models and short-term plasticity mechanisms

In rate-based population models, the dynamics of a network consisting of an excitatory (E) and an inhibitory (I) population are given by (Wilson & Cowan, 1972):

$$\tau_E \frac{dr_E}{dt} = -r_E + \left[J_{EE} r_E - J_{EI} r_I + g_E\right]_+^{\alpha_E}, \tag{2.1}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[J_{IE} r_E - J_{II} r_I + g_I\right]_+^{\alpha_I} \tag{2.2}$$

where $r_E$ and $r_I$ are the firing rates of the excitatory and inhibitory population; $\tau_E$ and $\tau_I$ represent the corresponding time constants; $J_{AB}$ denotes the connection strength from population $B$ to population $A$, where $A, B \in \{E, I\}$; $g_E$ and $g_I$ represent the external inputs to the respective populations; and $\alpha_E$ and $\alpha_I$ denote the exponents of the respective input-output functions. Note that $\alpha$ determines neuronal nonlinearities with $\alpha_E = \alpha_I = 1$ for threshold-linear networks and $\alpha_E = \alpha_I > 1$ for supralinear networks.

### Short-term plasticity

Modeling short-term plasticity was introduced by Tsodyks and Markram (Tsodyks & Markram, 1997). Based on the Tsodyks and Markram model, the dynamics of the network with short-term plasticity become as follows:

$$\tau_E \frac{dr_E}{dt} = -r_E + \left[p_{EE} J_{EE} r_E - p_{EI} J_{EI} r_I + g_E\right]_+^{\alpha_E}, \tag{2.3}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[p_{IE} J_{IE} r_E - p_{II} J_{II} r_I + g_I\right]_+^{\alpha_I} \tag{2.4}$$

where $p_{AB}$ is the short-term plasticity variable from population $B$ to population $A$.

For short-term depression (STD), we replaced $p_{AB}$ by $x_{AB}$ and expressed the STD dynamics as follows:

$$\frac{dx_{AB}}{dt} = \frac{1 - x_{AB}}{\tau_x} - U_d x_{AB} r_B, \tag{2.5}$$

where $x_{AB}$ represents a short-term depression variable that is constrained to the interval (0,1] for the connection from population B to population A. Biophysically, the short-term depression variable $x$ represents the fraction of vesicles available for

release, $\tau_x$ is the time constant of STD, and $U_d$ represents the depression factor that controls the degree of depression induced by the presynaptic activity.

For short-term facilitation (STF), we replaced $p_{AB}$ by $u_{AB}$ and described the STF dynamics as follows:

$$\frac{du_{AB}}{dt} = \frac{1 - u_{AB}}{\tau_u} + U_f(U_{max} - u_{AB})r_B, \tag{2.6}$$

where $u_{AB}$ represents a short-term facilitation variable that is limited to the interval $[1, U_{max})$ for the connection from population B to population A. Biophysically, the short-term facilitation variable $u$ represents the ability to release neurotransmitters, $\tau_u$ is the time constant of STF, $U_f$ represents the facilitation factor that controls the degree of facilitation induced by the presynaptic activity, and $U_{max}$ denotes the maximal facilitation value.

## Spiking neural network models and plasticity mechanisms

In spiking neural network models, single neurons are modeled as leaky integrate-and-fire with membrane potential of neuron $i$, $U_i$, given by (Dayan & Abbott, 2005; Gerstner *et al.,* 2014):

$$\tau^m \frac{dU_i}{dt} = (U^{\text{rest}} - U_i) + g_i^{\text{ext}}(t)(U^{\text{exc}} - U_i) + g_i^{\text{inh}}(t)(U^{\text{inh}} - U_i) \tag{2.7}$$

Here, $\tau^m$ is the membrane time constant, $U^{\text{rest}}$ is the resting potential, $U^{\text{exc}}$ is the excitatory reversal potential, $U^{\text{inh}}$ is the inhibitory reversal potential, $g_i^{\text{ext}}$ and $g_i^{\text{inh}}$ are the excitatory and inhibitory conductance, respectively. The neuron elicits a spike when its membrane potential reaches a spiking threshold which is typically around -50mV. After a spike, the membrane potential is reset to $U^{\text{rest}}$, and the neuron has a refractory period in which no spikes are permitted. Inhibitory neurons also follow the same integrate-and-fire formalism, but with a shorter membrane time constant.

In the model, excitatory synapses contain a fast AMPA component and a slow

NMDA component. Dynamics of excitatory conductances are given by:

$$\tau^{\text{ampa}} \frac{dg_i^{\text{ampa}}}{dt} = -g_i^{\text{ampa}} + \sum_{j \in \text{exc}} J_{ij} S_j(t) \tag{2.8}$$

$$\tau^{\text{nmda}} \frac{dg_i^{\text{nmda}}}{dt} = -g_i^{\text{nmda}} + g_i^{\text{ampa}} \tag{2.9}$$

$$g_i^{\text{exc}}(t) = \beta g_i^{\text{ampa}}(t) + (1 - \beta) g_i^{\text{nmda}}(t) \tag{2.10}$$

Here, $\tau^{\text{ampa}}$ is the AMPA decay time constant, $\tau^{\text{nmda}}$ is the NMDA decay time constant, and $J_{ij}$ is the synaptic strength from neuron $j$ to neuron $i$. $S_j(t)$ is the spike train of neuron $j$, which is defined as $S_j(t) = \sum_k \delta(t - t_j^k)$, where $\delta$ is the Dirac delta function $t_j^k$ are the spikes times $k$ of neuron j. Finally, $\beta$ is a weighting parameter.

Dynamics of inhibitory conductances are given by:

$$\tau^{\text{gaba}} \frac{dg_i^{\text{inh}}}{dt} = -g_i^{\text{inh}} + \sum_{j \in \text{inh}} J_{ij} S_j(t) \tag{2.11}$$

where $\tau^{\text{gaba}}$ is the GABA decay time constant.

## Excitatory plasticity

Classical pairwise STDP of excitatory synapses onto excitatory neurons can be easily formulated mathematically. However, standard pair-based STDP models fail to capture several frequency-dependent aspects of synaptic plasticity observed experimentally. These aspects include no potentiation for pre-post pairing at low stimulation frequency and increased potentiation with increasing stimulation frequency. To address these problems, a STDP rule based on triplet of spikes was proposed (Pfister & Gerstner, 2006). According to the triplet STDP rule, the dynamics of synaptic strength from excitatory neuron $j$ to excitatory neuron $i$ follow

$$\frac{dJ_{ij}}{dt} = -z_i^-(t) A^- S_j(t) + z_j^+(t) A^+ z_i^{\text{slow}}(t - \epsilon) S_i(t) \tag{2.12}$$

Here, $A^-$ and $A^+$ are the amplitude of the weight change induced by a post-pre pair or a post-pre-post triplet of spikes. $\epsilon$ is a small positive constant. The synaptic

traces $z_n^+(t)$, $z_n^-(t)$ and $z_n^{\text{slow}}(t)$ evolve according to

$$\frac{dz_n^m}{dt} = -\frac{z_n^m}{\tau^m} + S_n(t), m \in \{+, -, \text{slow}\}, n \in \{i, j\} \tag{2.13}$$

with different time constants $\tau^m$.

## Inhibitory plasticity

The STDP of inhibitory synapses onto excitatory neurons, known as inhibitory STDP (iSTDP), is formulated as follows (Vogels *et al.,* 2011):

$$\Delta w_{ij} = \eta^{\text{iSTDP}}(x_i - 2r_i^0\tau^{\text{iSTDP}})S_j(t) + \eta^{\text{iSTDP}}x_jS_i(t) \tag{2.14}$$

$$\frac{dx_n}{dt} = -\frac{x_n}{\tau^{\text{iSTDP}}} + S_n(t) \tag{2.15}$$

where $\Delta w_{ij}$ is the change in strength of inhibitory synapses onto excitatory neurons, $x_i$ and $x_j$ are the synaptic traces of the postsynaptic excitatory and presynaptic inhibitory neuron, $r_i^0$ is the target firing rate of excitatory neuron $i$, $\tau^{\text{iSTDP}}$ is the time constant of the synaptic trace, and $\eta^{\text{iSTDP}}$ denotes the learning rate of iSTDP.

## Heterosynaptic plasticity

To ensure that the sum of all incoming excitatory synaptic weights at each postsynaptic excitatory neuron is kept below a target (Fiete *et al.,* 2010), heterosynaptic plasticity is modelled as follows:

$$J_{ij}^{EE}(t) \leftarrow J_{ij}^{EE}(t) - \left(\sum_j J_{ij}^{EE}(t) - \gamma \sum_j J_{ij}^{EE}(0)\right)/N_i^E \tag{2.16}$$

where $N_i^E$ is the number of excitatory synaptic connections to excitatory neuron $i$. And $\gamma$ is a factor which makes the maximal $J_{ij}$ allowed by heterosynaptic plasticity approximately the same as the hard upper bound of $J^{EE}$. Heterosynaptic plasticity is implemented every 1 s, and only affects synaptic weights when the $\sum_j J_{ij}^{EE}(t)$ is larger than $\gamma \sum_j J_{ij}^{EE}(0)$.

## Metaplasticity

Previous studies have shown that the triplet STDP can be mapped to the BCM learning rule and the LTP/LTD threshold can change by varying the LTD amplitude $A^-$ (Pfister & Gerstner, 2006). Based on these findings, metaplasticity is implemented by having an adaptive LTD amplitude as follows:

$$A_i^- \leftarrow A_i^- \frac{x_i^{\text{est}}}{\tau^{\text{est}} r_i^0} \tag{2.17}$$

with

$$\frac{dx_i^{\text{est}}}{dt} = -\frac{x_i^{\text{est}}}{\tau^{\text{est}}} + S_i(t) \tag{2.18}$$

where $x_i^{\text{est}}$ is a firing rate estimator for excitatory neuron $i$, and $\tau^{\text{est}}$ is the time constant of the firing rate estimator. If the firing rate of a neuron is close to its target, $r_i^0$, then $\frac{x_i^{\text{est}}}{\tau^{\text{est}} r_i^0} \approx 1$. Metaplasticity is implemented every 30 s.

## Synaptic scaling

The change in synapse strength from excitatory neuron $j$ to excitatory neuron $i$ governed by synaptic scaling is given by (van Rossum *et al.,* 2000):

$$\tau^{ss} \frac{dJ_{ij}}{dt} = J_{ij} \left(1 - \frac{x_i^{\text{est}}}{\tau^{\text{est}} r_i^0}\right) \tag{2.19}$$

where $\tau^{ss}$ is the time constant of synaptic scaling.

## Intrinsic plasticity

Intrinsic plasticity is implemented by dynamically adjusting the firing threshold of a neuron according to its activity (Lazar *et al.,* 2009). The firing threshold of neuron $i$ regulated by intrinsic plasticity is given by:

$$\frac{dU_i^{\text{thr}}}{dt} = \eta^{\text{ip}} \left(\frac{x_i^{\text{est}}}{\tau^{\text{est}}} - r_i^0\right) \tag{2.20}$$

where $\eta^{\text{ip}}$ is the learning rate of intrinsic plasticity. Initial firing threshold is -50mV.

# 3 Results

During my PhD, using analytical and numerical methods, I investigated how short- and long-term plasticity shape cortical computations. I have contributed to five peer-reviewed journal articles and one preprint that is currently under review. I am the first or co-first author of five of these articles (Waitzmann *et al.,* 2023; Wu & Gjorgjieva, 2023; Wu *et al.,* 2020, 2022; Wu & Zenke, 2021), and I am a contributing author of the other one (Pacheco *et al.,* 2019). In the following, I provide a summary for each article, indicate my contribution, and reproduce the full text in the Appendix.

## 3.1 Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics

In Wu *et al.* (2020), we analyze extensive datasets of electrophysiological recordings over 9 days of the collective activity of multiple cells in the monocular region of primary visual cortex in freely behaving rodents during normal development 2-3 weeks after eye opening, and after monocular deprivation. We examine higher-order network properties during normal development and prolonged monocular deprivation. We further investigate how the network exploits various homeostatic mechanisms to restore normal dynamics following monocular deprivation by using a plastic recurrent spiking network model. We find that:

1. Functional correlations are subject to homeostatic regulations.
2. Different homeostatic mechanisms can regulate distinct aspects of cortical dynamics.
3. Synaptic scaling promotes the recovery of correlations and network structure.
4. Intrinsic plasticity greatly contributes to the rebound of firing rates.

The work was completed with experimental collaborators Dr. Keith Hengen and Prof. Gina Turrigiano, as well as my supervisor Prof. Dr. Julijana Gjorgjieva. My contributions to the article include designing research, performing research, creating new reagents/analytic tools, analyzing data, and writing the paper. The full article was published on Septmeber 11th, 2020 in PNAS and is reproduced in *Appendix I. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics* under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0.

## 3.2 Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity

In Wu & Gjorgjieva (2023), we investigate how neuronal nonlinearity and short-term plasticity affect inhibition stabilization, the paradoxical effect, and the relationship between the two. We find that:

1. Regardless of the neuronal nonlinearity, in networks with excitatory-to-excitatory short-term depression, inhibition stabilization does not necessarily imply the paradoxical effect, but the paradoxical effect implies inhibition stabilization.
2. In networks with static connectivity or networks with other short-term plasticity mechanisms instead of excitatory-to-excitatory short-term depression, inhibition stabilization and the paradoxical effect imply each other.
3. When neuronal nonlinearities and excitatory-to-excitatory short-term depression coexist, monotonically increasing excitatory activity can lead to nonmonotonic transitions between noninhibition-stabilization and inhibition-stabilization, as well as between nonparadoxically-responding and paradoxically-responding regimes.
4. Our results of the relationship between inhibition stabilization and the paradoxical effect can be generalized to more complex scenarios including networks with multiple interneuron subtypes and any monotonically increasing neuronal nonlinearities.

The work was completed with my supervisor Prof. Dr. Julijana Gjorgjieva. My contributions to the article include designing research, performing research, and writing the paper. The full article was published on July 12th, 2023 in Physical Review Research and is reproduced in *Appendix II. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity* under Creative Commons Attribution 4.0 International License.

## 3.3 Regulation of circuit organization and function through inhibitory synaptic plasticity

In Wu *et al.* (2022), we review the recent literature on inhibitory plasticity. We:

1. provide an overview of the evolution of inhibition over development.
2. review literature on how inhibitory plasticity controls excitation at different spatiotemporal scales.
3. review existing research related to how inhibition and inhibitory plasticity affect excitatory plasticity.
4. review studies on the role of inhibitory plasticity in the formation of structured networks and resulting computations.
5. review recent findings on interneuron-specific plasticity mechanisms and their functional implications.

The work was completed together with my co-author Christoph Miehl and my supervisor Prof. Dr. Julijana Gjorgjieva. My contributions to the article include conceptualization, visualization, and writing the paper. The full article was published on 28 October 2022 in Trends in Neurosciences and is reproduced in *Appendix III. Regulation of circuit organization and function through inhibitory synaptic plasticity* under Creative Commons Attribution 4.0 International License.

## 3.4 Nonlinear transient amplification in recurrent neural networks with short-term plasticity

In Wu & Zenke (2021), we investigate the underlying mechanism of stimulus-evoked transient neural dynamics in sensory systems, and examine the functional benefits of co-tuned inhibition and transient onset responses. We find that:

1. Transient neural dynamics can be generated via a nonlinear transient amplification mechanism, in which neuronal ensembles can rapidly, nonlinearly, and transiently amplify inputs by briefly switching from stable to unstable dynamics before being re-stabilized through short-term plasticity.
2. Co-tuned inhibition broadens the parameter regime in which nonlinear transient amplification is possible while avoiding persistent activity.
3. Transient onset responses are advantageous for neural computations like pattern completion and pattern separation.

The work was completed with Dr. Friedemann Zenke. My contributions to the article include formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review and editing. The full article was published on December 13th, 2021 in eLife and is reproduced in *Appendix IV. Nonlinear transient amplification in recurrent neural networks with short-term plasticity* under Creative Commons Attribution 4.0 International License.

## 3.5 Rapid and active stabilization of visual cortical firing rates across light-dark transitions

In Pacheco *et al.* (2019), we investigate how light-dark transitions affect firing on rapid timescales by analyzing datasets of electrophysiological recordings from neurons in primary visual cortex of freely behaving rodents. We find that:

1. Expected light-dark transitions have only a modest effect on the mean firing rates of neurons in primary visual cortex.
2. Functional correlations of neural activity are significantly stronger during the light than in darkness.
3. Unexpected light-dark transitions lead to a significant increase in the firing across the majority of neurons in primary visual cortex.

The work was completed with experimental collaborators Alejandro Torrado Pacheco, Elizabeth I. Tilden, Sophie M. Grutzner, Brian J. Lane, Dr. Keith Hengen and Prof. Gina Turrigiano, as well as my supervisor Prof. Dr. Julijana Gjorgjieva. My contributions to the article include analyzing data and writing the paper. The full article was published on Septmeber 11th, 2020 in PNAS and is reproduced in *Appendix V. Rapid and active stabilization of visual cortical firing rates across light–dark transitions* under the PNAS License.

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

In Waitzmann *et al.* (2023), we study the emergence of nonlinear phenomena in canonical cortical circuits consisting of multiple interneuron subtypes. We focuse on a counterintuitive nonlinear phenomenon, in which locomotion-induced top-down modulation via vasoactive intestinal peptide (VIP)-expressing interneurons affects the response of somatostatin (SST)-expressing interneurons in opposite directions depending on the sensory stimulation condition, is referred to as response reversal. We find that:

1. Response reversal can be generated through experimentally identified inhibitory short-term plasticity.
2. While not directly impacting SST and VIP activity, the short-term depression between parvalbumin (PV)-expressing interneurons to excitatory neurons plays a decisive role in generating response reversal.
3. Response reversal is tightly liked to interneuron-specific stabilization and the paradoxical effect.

The work was completed together with my co-author Felix Waitzmann and my supervisor Prof. Dr. Julijana Gjorgjieva. My contributions to the article include conceptualization, visualization, and writing the paper. The full article is currently under revision, is available on bioRxiv, and is reproduced in *Appendix VI. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity* under 

# 4 Discussion

Extensive discussions have been provided in discussion sections of the published articles. In this section below, I will mainly discuss aspects that have not been mentioned previously and highlight some interesting future directions.

**Cell-type-specific synaptic scaling mechanisms in associative learning**

In Wu *et al.,* 2020, I investigated the role of synaptic scaling in regulating network dynamics in the context of sensory perturbations. Recent experimental studies have examined how excitatory synaptic scaling contributes to associative learning (Wu *et al.,* 2021). By applying a conditioned taste aversion learning paradigm, the authors found that pairing a conditioned tastant with an unconditioned aversive stimulus leads to an overgeneralization of this aversion to other novel tastants at 4 hours after the pairing (Wu *et al.,* 2021). At 24 hours, this overgeneralization disappears and the specificity of associative memories emerges (Wu *et al.,* 2021). Blocking excitatory synaptic scaling prolongs the conditioned taste aversion-induced overgeneralization up to more than 24 hours, suggesting an important role of excitatory synaptic scaling in associative learning (Wu *et al.,* 2021). Interestingly, another recent study has revealed that inhibitory synapses onto excitatory neurons exhibit different scaling rules in a target-dependent manner (Prestigio *et al.,* 2021). More specifically, hyperactivity of excitatory neurons leads to upscaling of perisomatic inhibition but downscaling of dendritic inhibition (Prestigio *et al.,* 2021). How different synaptic scaling mechanisms interact with Hebbian plasticity and achieve memory specificity in associative learning is unclear. Together with my colleague, Ayça Kepçe, we are currently investigating this problem using a computational model. As PV neurons preferentially target perisomatic regions of excitatory neurons whereas SST neurons mainly target dendritic regions, we postulate that the target-dependent inhibitory synaptic scaling can be attributed to cell-type-specific synaptic scaling mechanisms. Namely, hyperactivity of excitatory neurons results

in upscaling of PV synapses but downscaling of SST synapses. Based on this assumption, we built a rate-based model consisting of two subnetworks. Each subnetwork contains one E, one PV, and one SST population. To mimic the experimental paradigm, the entire simulation in the computational model is separated into one conditioning period and one testing period. During the conditioning period, the E and PV population in the subnetwork 1 receive additional inputs corresponding to the stimulation of the conditioned stimulus in the experiments, and excitatory to excitatory weights evolve according to a three-factor Hebbian plasticity rule, in which the third factor serves as a control signal for Hebbian learning resembling the presence of the unconditioned aversive stimulus. While Hebbian learning is controlled by the third factor and thus only active during the conditioning period, synaptic scaling is active both during the conditioning period and the testing period. We tested overgeneralization by stimulating E and PV in the subnetwork 2 at three different time points (early, intermediate, and late) during the testing period, resembling 4h, 24h, and 48h after pairing in the experiments. We demonstrated that after conditioning, long-term potentiation induced by Hebbian plasticity results in overgeneralization to novel stimuli, in other words, elevated activity of the subnetwork 2 at the early stage of the testing period. The overgeneralization is subsequently eliminated by synaptic scaling mechanisms at the intermediate stage of the testing period resembling the disappearance of overgeneralization at 24h in the experiments. Blocking all synaptic scaling mechanisms leads to the persistence of overgeneralization and prevents the establishment of memory specificity over the entire testing period, suggesting that synaptic scaling is necessary for achieving memory specificity. Blocking excitatory synaptic scaling alone leads to a prolonged overgeneralization but memory specificity emerges at the late stage of the testing period, suggesting that inhibitory synaptic scaling mechanisms can rescue memory specificity in the absence of excitatory synaptic scaling. We further found that E-to-E scaling and PV-to-E scaling synergistically promote memory specificity, whereas SST-to-E scaling has an antagonistic effect. Furthermore, as other brain regions can exert top-down influence on SST via VIP that inhibits SST, we investigated the top-down modulation effect on memory specificity by changing the inputs to SST. We found that top-down inputs can greatly regulate the system towards either overgeneralization or memory specificity, suggesting a powerful control of internal states on associative memories. In summary, our work makes predictions on

the role of individual scaling mechanisms as well as their joint effects in associative learning.

### Computational implication of response reversal

In Waitzmann *et al.,* 2023, we showed that experimentally identified inhibitory short-term plasticity mechanisms can generate response reversal of SST. What does response reversal imply computationally? We unveiled the correspondence between the transition from a either PV or SST stabilized regime to a SST only stabilized regime and the occurrence of response reversal. And we found that in the network simulations, the excitatory population receives more inhibition from SST than PV once SST response is reversed. In the predictive processing framework, neurons integrate sensory inputs from the external world with internally generated predictive signals. Layer 2/3 in the cortex is recognized as a key site where predictive processing occurs (Keller & Mrsic-Flogel, 2018). In sensory cortices, excitatory neurons in layer 2/3 receive bottom-up sensory inputs at their basal dendrites from excitatory neurons in layer 4 (Petreanu *et al.,* 2009). Excitatory neurons in layer 2/3 receive top-down inputs at their distal dendrites from higher cortical areas (Zhang *et al.,* 2014). As PV neurons preferentially target perisomatic regions of excitatory neurons, whereas SST neurons predominantly target the distal dendritic regions of pyramidal neurons (Tremblay *et al.,* 2016), different sources of inhibition can gate different information flow. We therefore postulate that in darkness, weaker inhibition provided by SST than PV can permit top-down influence, thereby allowing the system to rely more heavily on the internal model in environments with high uncertainty, such as in darkness. In contrast, in the presence of visual stimulus, weaker inhibition provided by PV than SST favors bottom-up information over top-down predictive information. This bias towards bottom-up information allows the system to rely more effectively on the incoming sensory inputs, particularly in environments with low uncertainty. Therefore, the shift in interneuron-specific dominance occurring concurrently with the response reversal of SST could regulate information flow and might play an important role in predictive processing.

### Distinguish neuronal nonlinearities and synaptic nonlinearities

Previous computational studies have suggested that response reversal can be generated by supralinear-like neuronal nonlinearities (Garcia Del Molino *et al.,* 2017).

In Waitzmann *et al.,* 2023, we demonstrated that synaptic nonlinearities could underlie the generation of response reversal. How can we experimentally distinguish these two hypotheses given the fact that both of them exhibit the same phenomena? One possibility is that they might have different phase transitions in terms of interneuron-specific stabilization. In Garcia Del Molino *et al.,* 2017, the authors stated that neuronal nonlinearity-dependent response reversal is associated with a network transition from a non-inhibition stabilized regime to an inhibition stabilized regime. Despite not directly shown in Garcia Del Molino *et al.,* 2017, it is straightforward to prove that in the presence of neuronal nonlinearities, in the inhibition stabilized regime in which top-down modulation via VIP increases SST activity, the network is stabilized by SST only. In contrast, for synaptic nonlinearties-dependent response reversal, network can transition from a PV-only stabilized regime to a either PV or SST-stabilized regime to a SST-only stabilized regime. As a result, in the regime in which top-down modulation via VIP decreases SST activity, networks with synaptic nonlinearities might exhibit richer phase transitions than networks with neuronal nonlinearities in terms of interneuron-specific stabilization. These transitions can be experimentally tested by examining paradoxical effects using optogenetic tools. More specifically, in the PV-only stabilized regime, the network should display paradoxical response of PV. In the either PV or SST stabilized regime, the network would exhibit neither paradoxical response of PV nor of SST. In the SST-only stabilized regime, the network would show paradoxical response of SST. Therefore, the phase transitions in interneuron-specific stabilization and paradoxical effects could potentially help to distinguish these two hypotheses.

**Interplay between neuronal nonlinearities and synaptic nonlinearities**

In Wu *et al.,* 2020, we have shown how interesting computations can emerge due to the presence of both neuronal nonlinearities and synaptic nonlinearities in networks with excitatory and inhibitory populations. In Waitzmann *et al.,* 2023, we have demonstrated how synaptic nonlinearities can generate response reversal in networks with multiple interneuron subtypes in the absence of neuronal nonlinearities. Various computations in the brain involve multiple interneuron subtypes. At the timescale of perception, synaptic nonlinearities can greatly affect neural activity. It is therefore interesting for future research to investigate how neuronal nonlin-

earities and synaptic nonlinearities interact in networks with multiple interneuron subtypes, how the interplay between them co-shapes neural dynamics and affects computations involving diverse types of interneurons. In particular, do neuronal and synaptic nonlinearities work in an antagonistic or cooperative manner? How does the working mode of these nonlinearities depend on the network operating regimes? Which computations require the co-existence of both nonlinearities?

**Learning in networks with multiple interneuron subtypes**

To be able to perform different computations in the brain, the connectivity of neural circuits needs to be set up properly. Activity-dependent plasticity mechanisms play a crucial role in establishing appropriate network connectivity. With recent advances in experimental techniques, studies have started to reveal plasticity rules at different types of inhibitory synapses onto excitatory neurons. Of particular interest for future studies is characterizing and formulating plasticity rules between different types of synapses to provide a comprehensive plasticitome, investigating how these plasticity mechanisms interact with each other in networks with multiple interneuron subtypes, and how they shape connectivity structures and computations. It is worth noting that inhibitory plasticity rules of the same types of synapses identified molecularly on excitatory neurons can even differ across brain regions. In the interconnected brain network, changes in the connectivity of one brain region can alter local neural activity and thus affect other regions. To ensure proper neural activity in downstream brain areas that control behaviors, different brain regions have to coordinate with each other. Future studies are required to understand how neural circuits achieve this coordination in the presence of heterogeneity of plasticity rules.

**Alternative approaches for studying learning in biological systems**

Major classical computational studies use bottom-up approaches by incorporating phenomenological yet biologically plausible learning rules, and examine the functional implications of these learning rules. These learning rules are typically characterized *in vitro*, which can be very different from *in vivo*. In realistic settings, when animals are learning new tasks, multiple factors like attention and behavioral states are involved. These factors are typically associated with neuromodulators, which can dramatically change the learning rules. A few studies attempted

to infer learning rules from distributions of firing rates *in vivo* but only assuming that excitatory-to-excitatory connections are plastic (Lim *et al.,* 2015). In contrast, learning can involve complex cell-type-specific reorganization of inhibition (Poort *et al.,* 2022). To understand how biological circuits evolve during learning, it would be interesting to infer learning rules using *in vivo* data in networks with multiple interneuron subtypes. Furthermore, to narrow down possible learning rules, other constraints like capturing temporal dynamics of single neurons could be imposed.

In addition, recent works start to apply top-down approaches to study synaptic plasticity (Confavreux *et al.,* 2020; Jordan *et al.,* 2021). In the top-down approaches, synaptic plasticity rules are identified by minimizing a loss function that relates to the desired functionality of the network. How to define those functional loss functions and how to efficiently optimize them are active research topics. As a complement to bottom-up approaches, top-down approaches could be useful to explain how biological systems can achieve robust function in the presence of noisy plasticity rules, and how and why different biological systems use different plasticity rules to implement the same function.

**Final remarks**

By combining analytical and numerical methods, my PhD work provides insight into how different plasticity mechanisms operating over different time scales affect neural dynamics and computations. With the advances in neurotechnologies, we are now able to record many different types of neurons simultaneously for a long time while animals learn new tasks. All of these bring new challenges and opportunities for developing new theories and synthesizing data and knowledge to understand plasticity and neural computations.

# Bibliography


1. Abbott, L. F. & Regehr, W. G. Synaptic computation. *Nature* **431,** 796–803 (2004).
2. Abbott, L. F., Varela, J. A., Sen, K. & Nelson, S. B. Synaptic Depression and Cortical Gain Control. *Science* **275,** 221–224 (1997).
3. Abraham, W. C. Metaplasticity: Tuning synapses and networks for plasticity. *Nature Reviews Neuroscience* **9,** 387–399 (2008).
4. Abraham, W. C. & Bear, M. F. Metaplasticity: plasticity of synaptic plasticity. *Trends Neuroscience* **19,** 126–130 (1996).
5. Alejandre-García, T., Kim, S., Pérez-Ortega, J. & Yuste, R. Intrinsic excitability mechanisms of neuronal ensemble formation. *eLife* **11,** e77470 (2022).
6. Attinger, A., Wang, B. & Keller, G. B. Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex. *Cell* **169,** 1291–1302 (2017).
7. Auth, J. M., Nachstedt, T. & Tetzlaff, C. The Interplay of Synaptic Plasticity and Scaling Enables Self-Organized Formation and Allocation of Multiple Memory Representations. *Frontiers in Neural Circuits* **14,** 541728 (2020).
8. Beniaguev, D., Segev, I. & London, M. Single cortical neurons as deep artificial neural networks. *Neuron* **109,** 2727–2739.e3 (2021).
9. Bi, G.-g. & Poo, M.-M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience* **18,** 10464–72 (1998).
10. Bienenstock, E. L., Cooper, L. N. & Munro, P. W. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience* **2,** 32–48 (1982).
11. Blatow, M., Caputi, A., Burnashev, N., Monyer, H. & Rozov, A. Ca2+ buffer saturation underlies paired pulse facilitation in calbindin-D28k-containing terminals. *Neuron* **38,** 79–88 (2003).
12. Buzsáki, G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* **68,** 362–385 (2010).

13. Campagnola, L., Seeman, S. C., Chartrand, T., Kim, L., Hoggarth, A., Gamlin, C., Ito, S., Trinh, J., Davoudian, P., Radaelli, C., Kim, M.-H., Hage, T., Braun, T., Alfiler, L., Andrade, J., Bohn, P., Dalley, R., Henry, A., Kebede, S., Mukora, A., Sandman, D., Williams, G., Larsen, R., Teeter, C., Daigle, T. L., Berry, K., Dotson, N., Enstrom, R., Gorham, M., Hupp, M., Lee, S. D., Ngo, K., Nicovich, R., Potekhina, L., Ransford, S., Gary, A., Goldy, J., McMillen, D., Pham, T., Tieu, M., Siverts, L., Walker, M., Farrell, C., Schroedter, M., Slaughterbeck, C., Cobb, C., Ellenbogen, R., Gwinn, R. P., Keene, C. D., Ko, A. L., Ojemann, J. G., Silbergeld, D. L., Carey, D., Casper, T., Crichton, K., Clark, M., Dee, N., Ellingwood, L., Gloe, J., Kroll, M., Sulc, J., Tung, H., Wadhwani, K., Brouner, K., Egdorf, T., Maxwell, M., McGraw, M., Pom, C. A., Ruiz, A., Bomben, J., Feng, D., Hejazinia, N., Shi, S., Szafer, A., Wakeman, W., Phillips, J., Bernard, A., Esposito, L., D'Orazi, F. D., Sunkin, S., Smith, K., Tasic, B., Arkhipov, A., Sorensen, S., Lein, E., Koch, C., Murphy, G., Zeng, H. & Jarsky, T. Local Connectivity and Synaptic Dynamics in Mouse and Human Neocortex. *Science* **375** (2022).

14. Canto-Bustos, M., Friason, F. K., Bassi, C. & Oswald, A.-M. M. Disinhibitory circuitry gates associative synaptic plasticity in olfactory cortex. *The Journal of Neuroscience* **42,** 2942–2950 (2022).

15. Capogna, M., Castillo, P. E. & Maffei, A. The ins and outs of inhibitory synaptic plasticity: Neuron types, molecular mechanisms and functional roles. *European Journal of Neuroscience* **54,** 6882–6901 (2021).

16. Caporale, N. & Dan, Y. Spike Timing-Dependent Plasticity: A Hebbian Learning Rule. *Annual Review of Neuroscience* **31,** 25–46 (2008).

17. Carr, M. F., Jadhav, S. P. & Frank, L. M. Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nature Neuroscience* **14,** 147–153 (2011).

18. Castillo, P. E., Chiu, C. Q. & Carroll, R. C. Long-term plasticity at inhibitory synapses. *Current Opinion in Neurobiology* **21,** 328–338 (2011).

19. Chater, T. E. & Goda, Y. My Neighbour Hetero-deconstructing the mechanisms underlying heterosynaptic plasticity. *Current Opinion in Neurobiology* **67,** 106–114 (2021).

20. Chistiakova, M., Bannon, N. M., Chen, J.-Y., Bazhenov, M. & Volgushev, M. Homeostatic role of heterosynaptic plasticity: models and experiments. *Frontiers in Computational Neuroscience* **9** (2015).

21. Chiu, C. Q., Barberis, A. & Higley, M. J. Preserving the balance: diverse forms of long-term GABAergic synaptic plasticity. *Nature Reviews Neuroscience* **20,** 272–281 (2019).

22. Clopath, C., Büsing, L., Vasilaki, E. & Gerstner, W. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neuroscience* **13,** 344–352 (2010).

23. Confavreux, B., Zenke, F., Agnes, E., Lillicrap, T. & Vogels, T. A meta-learning approach to (re) discover plasticity rules that carve a desired function into a neural network. *Advances in Neural Information Processing Systems* **33,** 16398–16408 (2020).

24. D'amour, J. A. & Froemke, R. C. Inhibitory and Excitatory Spike-Timing-Dependent Plasticity in the Auditory Cortex. *Neuron* **86,** 514–528 (2015).

25. Daoudal, G. & Debanne, D. Long-term plasticity of intrinsic excitability: learning rules and mechanisms. *Learning Memory* **10,** 456–465 (2003).

26. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience* (2005).

27. Debanne, D., Inglebert, Y. & Russier, M. Plasticity of intrinsic neuronal excitability. *Current Opinion in Neurobiology* **54,** 73–82 (2019).

28. Desai, N. S., Cudmore, R. H., Nelson, S. B. & Turrigiano, G. G. Critical periods for experience-dependent synaptic scaling in visual cortex. *Nature Neuroscience* **5,** 783–789 (2002).

29. Desai, N. S., Rutherford, L. C. & Turrigiano, G. G. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience* **2,** 515–520 (1999).

30. Diering, G. H. & Huganir, R. L. The AMPA Receptor Code of Synaptic Plasticity. *Neuron* **100,** 314–329 (2018).

31. Dudek, S. M. & Bear, M. F. Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *Proceedings of the National Academy of Sciences* **89,** 4363–4367 (1992).

32. Ebbesen, C. L. & Brecht, M. Motor cortex—to act or not to act? *Nature Reviews Neuroscience* **18,** 694–705 (2017).

33. Euston, D. R., Gruber, A. J. & McNaughton, B. L. The role of medial prefrontal cortex in memory and decision making. *Neuron* **76,** 1057–1070 (2012).

34. Fan, Y., Fricker, D., Brager, D. H., Chen, X., Lu, H. C., Chitwood, R. A. & Johnston, D. Activity-dependent decrease of excitability in rat hippocampal neurons through increases in Ih. *Nature Neuroscience* **8,** 1542–1551 (2005).

35. Feldman, D. E. Timing-Based LTP and LTD at Vertical Inputs to Layer II/III Pyramidal Cells in Rat Barrel Cortex. *Neuron* **27,** 45–56 (2000).

36. Field, R. E., D'amour, J. A., Tremblay, R., Miehl, C., Rudy, B., Gjorgjieva, J. & Froemke, R. C. Heterosynaptic Plasticity Determines the Set Point for Cortical Excitatory-Inhibitory Balance. *Neuron* **106,** 842–854 (2020).

37. Fiete, I. R., Senn, W., Wang, C. Z. H. & Hahnloser, R. H. R. Spike-Time-Dependent Plasticity and Heterosynaptic Competition Organize Networks to Produce Long Scale-Free Sequences of Neural Activity. *Neuron* **65,** 563–576 (2010).

38. Fioravante, D. & Regehr, W. G. Short-term forms of presynaptic plasticity. *Current Opinion in Neurobiology* **21,** 269–274 (2011).

39. Forsythe, I. D., Tsujimoto, T., Barnes-Davies, M., Cuttle, M. F. & Takahashi, T. Inactivation of presynaptic calcium current contributes to synaptic depression at a fast central synapse. *Neuron* **20,** 797–807 (1998).

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

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

42. Garrett, M., Manavi, S., Roll, K., Ollerenshaw, D. R., Groblewski, P. A., Ponvert, N. D., Kiggins, J. T., Casal, L., Mace, K., Williford, A., Leon, A., Jia, X., Ledochowitsch, P., Buice, M. A., Wakeman, W., Mihalas, S. & Olsen, S. R. Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *eLife* **9,** e50340 (2020).

43. Gerstner, W., Kistler, W. M., Naud, R. & Paninski, L. *Neuronal dynamics: From single neurons to networks and models of cognition* (Cambridge University Press, 2014).

44. Grubb, M. S. & Burrone, J. Activity-dependent relocation of the axon initial segment fine-tunes neuronal excitability. *Nature* **465,** 1070–1074 (2010).

45. Haas, J. S., Nowotny, T. & Abarbanel, H. D. I. Spike-Timing-Dependent Plasticity of Inhibitory Synapses in the Entorhinal Cortex. *Journal of Neurophysiology* **96,** 3305–3313 (2006).

46. Hebb, D. O. *The organization of behavior; a neuropsychological theory* (Wiley, 1949).

47. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology* **117,** 500 (1952).

48. Hosoi, N., Holt, M. & Sakaba, T. Calcium Dependence of Exo- and Endocytotic Coupling at a Glutamatergic Synapse. *Neuron* **63,** 216–229 (2009).

49. Hua, Y., Woehler, A., Kahms, M., Haucke, V., Neher, E. & Klingauf, J. Blocking endocytosis enhances short-term synaptic depression under conditions of normal availability of vesicles. *Neuron* **80,** 343–349 (2013).

50. Huganir, R. L. & Nicoll, R. A. AMPARs and synaptic plasticity: The last 25 years. *Neuron* **80,** 704–717 (2013).

51. Ibata, K., Sun, Q. & Turrigiano, G. G. Report Rapid Synaptic Scaling Induced by Changes in Postsynaptic Firing. *Neuron* **57,** 819–826 (2008).

52. Izhikevich, E. M. *Dynamical systems in neuroscience* (MIT press, 2007).

53. Jamann, N., Dannehl, D., Lehmann, N., Wagener, R., Thielemann, C., Schultz, C., Staiger, J., Kole, M. H. & Engelhardt, M. Sensory input drives rapid homeostatic scaling of the axon initial segment in mouse barrel cortex. *Nature Communications* **12,** 1–14 (2021).

54. Jiang, X., Shen, S., Cadwell, C. R., Berens, P., Sinz, F., Ecker, A. S., Patel, S. & Tolias, A. S. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science* **350,** aac9462 (2015).

55. Jones, E. G. *The thalamus* (Springer Science & Business Media, 2012).

56. Jordan, J., Schmidt, M., Senn, W. & Petrovici, M. A. Evolving interpretable plasticity for spiking networks. *Elife* **10,** e66273 (2021).

57. Kandel, E. R. *Principles of neural science* (2013).

58. Katz, B. & Miledi, R. The role of calcium in neuromuscular facilitation. *The Journal of physiology* **195,** 481–492 (1968).

59. Keck, T., Keller, G. B., Jacobsen, R. I., Eysel, U. T., Bonhoeffer, T. & Hübener, M. Synaptic scaling and homeostatic plasticity in the mouse visual cortex in vivo. *Neuron* **80,** 327–334 (2013).

60. Keller, A. J., Dipoppa, M., Roth, M. M., Caudill, M. S., Ingrosso, A., Miller, K. D. & Scanziani, M. A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron* **108,** 1181–1193 (2020).

61. Keller, G. B. & Mrsic-Flogel, T. D. Predictive processing: a canonical cortical computation. *Neuron* **100,** 424–435 (2018).

62. Kirkwood, A., Dudek, S. M., Gold, J. T., Aizenman, C. D. & Bear, M. F. Common forms of synaptic plasticity in the hippocampus and neocortex in vitro. *Science* **260,** 1518–1521 (1993).

63. Kirkwood, A., Rioult, M. G. & Bear, M. F. Experience-dependent modification of synaptic plasticity in visual cortex. *Nature* **381,** 526–528 (1996).

64. Krabbe, S., Paradiso, E., D'Aquin, S., Bitterman, Y., Courtin, J., Xu, C., Yonehara, K., Markovic, M., Müller, C., Eichlisberger, T., Gründemann, J., Ferraguti, F. & Lüthi, A. Adaptive disinhibitory gating by VIP interneurons permits associative learning. *Nature Neuroscience* **22,** 1834–1843 (2019).

65. Kuchibhotla, K. V., Gill, J. V., Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. H., Miller, K. D. & Froemke, R. C. Parallel processing by cortical inhibition enables context-dependent behavior. *Nature neuroscience* **20,** 62–71 (2017).

66. Lagzi, F., Canto-Bustos, M., Oswald, A.-M. M. & Doiron, B. Assembly formation is stabilized by Parvalbumin neurons and accelerated by Somatostatin neurons. *bioRxiv* (2021).

67. Lamprecht, R. & LeDoux, J. Structural plasticity and memory. *Nature Reviews Neuroscience* **5,** 45–54 (2004).

68. Lazar, A., Pipa, G. & Triesch, J. SORN: A Self Organizing Recurrent Neural Network. *Frontiers in Computational Neuroscience* **3** (2009).

69. Lee, C. C. Thalamic and cortical pathways supporting auditory processing. *Brain and language* **126,** 22–28 (2013).

70. Lim, S., Mckee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L. & Brunel, N. Inferring learning rules from distributions of firing rates in cortical neurons. *Nature Neuroscience* **18,** 1804–1810 (2015).

71. Litwin-Kumar, A. & Doiron, B. Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature Communications* **5,** 5319 (2014).

72. Loebel, A., Nelken, I. & Tsodyks, M. Processing of sounds by population spikes in a model of primary auditory cortex. *Frontiers in Neuroscience* **1,** 197–209 (2007).

73. Loebel, A. & Tsodyks, M. Computation by ensemble synchronization in recurrent networks with synaptic depression. *J Comput Neurosci* **13,** 111–124 (2002).

74. Luscher, B., Fuchs, T. & Kilpatrick, C. L. GABAA Receptor Trafficking-Mediated Plasticity of Inhibitory Synapses. *Neuron* **70,** 385–409 (2011).

75. Lüscher, C. & Malenka, R. C. NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). *Cold Spring Harbor perspectives in biology* **4,** a005710 (2012).

76. Lynch, G. S., Dunwiddie, T. & Gribkoff, V. Heterosynaptic depression: a postsynaptic correlate of long-term potentiation. *Nature* **266,** 737–739 (1977).

77. Maes, A., Barahona, M. & Clopath, C. Learning compositional sequences with multiple time scales through a hierarchical network of spiking neurons. *PLoS Computational Biology* **17,** e1008866 (2021).

78. Maes, A., Barahona, M. & Clopath, C. Learning spatiotemporal signals using a recurrent spiking network that discretizes time. *PLoS Computational Biology* **16,** e1007606 (2020).

79. Makino, H. & Malinow, R. AMPA Receptor Incorporation into Synapses during LTP: The Role of Lateral Movement and Exocytosis. *Neuron* **64,** 381–390 (2009).

80. Malinow, R. & Malenka, R. C. AMPA receptor trafficking and synaptic plasticity. *Annual Review of Neuroscience* **25,** 103–126 (2002).

81. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96,** 713–723 (1999).

82. Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. *Science* **275,** 213–215 (1997).

83. Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., Kahou, G. A. A., Berger, T. K., Bilgili, A., Buncic, N., Chalimourda, A., Chindemi, G., Courcol, J. D., Delalondre, F., Delattre, V., Druckmann, S., Dumusc, R., Dynes, J., Eilemann, S., Gal, E., Gevaert, M. E., Ghobril, J. P., Gidon, A., Graham, J. W., Gupta, A., Haenel, V., Hay, E., Heinis, T., Hernando, J. B., Hines, M., Kanari, L., Keller, D., Kenyon, J., Khazen, G., Kim, Y., King, J. G., Kisvarday, Z., Kumbhar, P., Lasserre, S., Le Bé, J. V., Magalhães, B. R., Merchán-Pérez, A., Meystre, J., Morrice, B. R., Muller, J., Muñoz-Céspedes, A., Muralidhar, S., Muthurasa, K., Nachbaur, D., Newton, T. H., Nolte, M., Ovcharenko, A., Palacios, J., Pastor, L., Perin, R., Ranjan, R., Riachi, I., Rodríguez, J. R., Riquelme, J. L., Rössert, C., Sfyrakis, K., Shi, Y., Shillcock, J. C., Silberberg, G., Silva, R., Tauheed, F., Telefont, M., Toledo-Rodriguez, M., Tränkler, T., Van Geit, W., Díaz, J. V., Walker, R., Wang, Y., Zaninetta, S. M., Defelipe, J., Hill, S. L., Segev, I. & Schürmann, F. Reconstruction and Simulation of Neocortical Microcircuitry. *Cell* **163,** 456–492 (2015).

84. Martin, S. J., Grimwood, P. D. & Morris, R. G. M. Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review of Neuroscience* **23,** 649–711 (2000).

85. McFarlan, A. R., Chou, C. Y., Watanabe, A., Cherepacha, N., Haddad, M., Owens, H. & Sjöström, P. J. The plasticitome of cortical interneurons. *Nature Reviews Neuroscience* **24,** 80–97 (2023).

86. McLaughlin, T. & O'Leary, D. D. Molecular gradients and development of retinotopic maps. *Annual Review of Neuroscience* **28,** 327–355 (2005).

87. Miehl, C. & Gjorgjieva, J. Stability and learning in excitatory synapses by nonlinear inhibitory plasticity. *PLoS Computational Biology* **18,** e1010682 (2022).

88. Monday, H. R., Wang, H. C. & Feldman, D. Circuit-level theories for sensory dysfunction in autism: Convergence across mouse models? *Frontiers in Neurology* **14,** 1254297 (2023).

89. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic Theory of Working Memory. *Science* **319,** 1543–1546 (2008).

90. Neher, E. & Sakaba, T. Multiple Roles of Calcium Ions in the Regulation of Neurotransmitter Release. *Neuron* **59,** 861–872 (2008).

91. Oh, W. C., Parajuli, L. K. & Zito, K. Heterosynaptic Structural Plasticity on Local Dendritic Segments of Hippocampal CA1 Neurons. *Cell Reports* **10,** 162–169 (2015).

92. Pacheco, A. T., Tilden, E. I., Grutzner, S. M., Lane, B. J., Wu, Y., Hengen, K. B., Gjorgjieva, J. & Turrigiano, G. G. Rapid and active stabilization of visual cortical firing rates across light-dark transitions. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 18068–18077 (2019).

93. Petreanu, L., Mao, T., Sternson, S. M. & Svoboda, K. The subcellular organization of neocortical excitatory connections. *Nature* **457,** 1142–1145 (2009).

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

95. Pfister, J.-P. & Gerstner, W. Triplets of Spikes in a Model of Spike Timing-Dependent Plasticity. *Journal of Neuroscience* **26,** 9673–9682 (2006).

96. Poirazi, P., Brannon, T. & Mel, B. W. Pyramidal neuron as two-layer neural network. *Neuron* **37,** 989–999 (2003).

97. Poort, J., Wilmes, K. A., Blot, A., Chadwick, A., Sahani, M., Clopath, C., Mrsic-Flogel, T. D., Hofer, S. B. & Khan, A. G. Learning and attention increase visual response selectivity through distinct mechanisms. *Neuron* **110,** 686–697 (2022).

98. Prestigio, C., Ferrante, D., Marte, A., Romei, A., Lignani, G., Onofri, F., Valente, P., Benfenati, F. & Baldelli, P. REST/NRSF drives homeostatic plasticity of inhibitory synapses in a target-dependent fashion. *Elife* **10,** e69058 (2021).

99. Regehr, W. G. Short-term presynaptic plasticity. *Cold Spring Harbor Perspectives in Biology* **4,** 1–19 (2012).

100. Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A. & Harris, K. D. The asynchronous state in cortical circuits. *Science* **327,** 587–590 (2010).

101. Rizzoli, S. O. & Betz, W. J. Synaptic vesicle pools. *Nature Reviews Neuroscience* **6,** 57–69 (2005).

102. Royer, S. & Paré, D. Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature* **422,** 518–522 (2003).

103. Rozov, A., Burnashev, N., Sakmann, B. & Neher, E. Transmitter release modulation by intracellular Ca2+ buffers in facilitating and depressing nerve terminals of pyramidal cells in layer 2/3 of the rat neocortex indicates a target cell-specific difference in presynaptic calcium dynamics. *The Journal of physiology* **531,** 807–826 (2001).

104. Seabrook, T. A., Burbridge, T. J., Crair, M. C. & Huberman, A. D. Architecture, function, and assembly of the mouse visual system. *Annual review of neuroscience* **40,** 499–538 (2017).

105. Sjöström, P. J., Turrigiano, G. G. & Nelson, S. B. Rate, Timing, and Cooperativity Jointly Determine Cortical Synaptic Plasticity. *Neuron* **32,** 1149–1164 (2001).

106. Song, S. C., Shen, B., Machold, R., Rudy, B., Glimcher, P. W., Louie, K. & Froemke, R. C. Input-Specific Inhibitory Plasticity Improves Decision Accuracy Under Noise. *bioRxiv* (2022).

107. Sprekeler, H. Functional consequences of inhibitory plasticity: homeostasis, the excitation-inhibition balance and beyond. *Current Opinion in Neurobiology* **43,** 198–203 (2017).

108. Stevens, C. F. Neurotransmitter release at central synapses. *Neuron* **40,** 381–388 (2003).

109. Tetzlaff, C., Kolodziejski, C., Timme, M., Tsodyks, M. & Wörgötter, F. Synaptic Scaling Enables Dynamically Distinct Short- and Long-Term Memory Formation. *PLoS Computational Biology* **9,** e1003307 (2013).

110. Thompson, A., Gribizis, A., Chen, C. & Crair, M. C. Activity-dependent development of visual receptive fields. *Current Opinion in Neurobiology* **42,** 136–143 (2017).

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

112. Tsodyks, M. V. & Markram, H. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 719–723 (1997).

113. Turrigiano, G. G. The Self-Tuning Neuron: Synaptic Scaling of Excitatory Synapses. *Cell* **135,** 422–435 (2008).

114. Turrigiano, G. G. Too Many Cooks? Intrinsic and Synaptic Homeostatic Mechanisms in Cortical Circuit Refinement. *Annual Review of Neuroscience* **34,** 89–103 (2011).

115. Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C. & Nelson, S. B. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* **391,** 892–896 (1998).

116. Uchida, N., Poo, C. & Haddad, R. Coding and transformations in the olfactory system. *Annual review of neuroscience* **37,** 363–385 (2014).

117. Udakis, M., Pedrosa, V., Chamberlain, S. E. L., Clopath, C. & Mellor, J. R. Interneuron-specific plasticity at parvalbumin and somatostatin inhibitory synapses onto CA1 pyramidal neurons shapes hippocampal output. *Nature Communications* **11** (2020).

118. Van Rossum, M. C. W., Bi, G.-Q. & Turrigiano, G. G. Stable Hebbian Learning from Spike Timing-Dependent Plasticity. *The Journal of Neuroscience* **20,** 8812–8821 (2000).

119. Van Vreeswijk, C. & Sompolinsky, H. Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity. *Science* **274,** 1724–1726 (1996).

120. Van Vreeswijk, C. & Sompolinsky, H. Chaotic Balanced State in a Model of Cortical Circuits. *Neural Computation* **10,** 1321–1371 (1998).

121. Veit, J., Hakim, R., Jadi, M. P., Sejnowski, T. J. & Adesnik, H. Cortical gamma band synchronization through somatostatin interneurons. *Nature Neuroscience* **20,** 951–959 (2017).

122. Veit, J., Handy, G., Mossing, D. P., Doiron, B. & Adesnik, H. Cortical VIP neurons locally control the gain but globally control the coherence of gamma band rhythms. *Neuron* (2022).

123. Vickers, E. D., Clark, C., Osypenko, D., Fratzl, A., Kochubey, O., Bettler, B. & Schneggenburger, R. Parvalbumin-Interneuron Output Synapses Show Spike-Timing-Dependent Plasticity that Contributes to Auditory Map Remodeling. *Neuron* **99,** 720–735 (2018).

124. Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory Plasticity Balances Excitation and Inhibition in Sensory Pathways and Memory Networks. *Science* **334,** 1569–1573 (2011).

125. Waitzmann, F., Wu, Y. K. & Gjorgjieva, J. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity. *bioRxiv* (2023).

126. Wang, Y., Markram, H., Goodman, P. H., Berger, T. K., Ma, J. & Goldman-Rakic, P. S. Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nature Neuroscience* **9,** 534–542 (2006).

127. Wanner, A. A. & Friedrich, R. W. Whitening of odor representations by the wiring diagram of the olfactory bulb. *Nature Neuroscience* (2020).

128. Williams, L. E. & Holtmaat, A. Higher-Order Thalamocortical Inputs Gate Synaptic Long-Term Potentiation via Disinhibition. *Neuron* **101,** 91–102 (2019).

129. Wilson, H. R. & Cowan, J. D. Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal* **12,** 1–24 (1972).

130. Wong, K. F. & Wang, X. J. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience* **26,** 1314–1328 (2006).

131. Woodin, M. A., Ganguly, K. & Poo, M.-M. Coincident Pre- and Postsynaptic Activity Modifies GABAergic Synapses by Postsynaptic Changes in Cl- Transporter Activity. *Neuron* **39,** 807–820 (2003).

132. Wu, C.-H., Ramos, R., Katz, D. B. & Turrigiano, G. G. Homeostatic synaptic scaling establishes the specificity of an associative memory. *Current biology* **31,** 2274–2285 (2021).

133. Wu, Y. K. & Gjorgjieva, J. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity. *Physical Review Research* **5,** 033023 (2023).

134. Wu, Y. K., Hengen, K. B., Turrigiano, G. G. & Gjorgjieva, J. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 24514–24525 (2020).

135. Wu, Y. K., Miehl, C. & Gjorgjieva, J. Regulation of circuit organization and function through inhibitory synaptic plasticity. *Trends in Neurosciences* **45,** 884–898 (2022).

136. Wu, Y. K. & Zenke, F. Nonlinear transient amplification in recurrent neural networks with short-term plasticity. *eLife* **10,** e71263 (2021).

137. Xu, J. & Wu, L. G. The decrease in the presynaptic calcium current is a major cause of short-term depression at a calyx-type synapse. *Neuron* **46,** 633–645 (2005).

138. Yap, E. L., Pettit, N. L., Davis, C. P., Nagy, M. A., Harmin, D. A., Golden, E., Dagliyan, O., Lin, C., Rudolph, S., Sharma, N., Griffith, E. C., Harvey, C. D. & Greenberg, M. E. Bidirectional perisomatic inhibitory plasticity of a Fos neuronal network. *Nature* **590,** 115–121 (2021).

139. Zenke, F., Agnes, E. J. & Gerstner, W. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature Communications* **6,** 6922 (2015).

140. Zhang, S., Xu, M., Kamigaki, T., Hoang Do, J. P., Chang, W.-C., Jenvay, S., Miyamichi, K., Luo, L. & Dan, Y. Long-range and local circuits for top-down modulation of visual cortex processing. *Science* **345,** 660–665 (2014).

141. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annual Review of Physiology* **64,** 355–405 (2002).

# List of scientific communications

## Peer-reviewed publications

- Waitzmann, F., Wu, Y. K. & Gjorgjieva, J. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity. *bioRxiv* (2023)
- Wu, Y. K. & Gjorgjieva, J. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity. *Physical Review Research* **5,** 033023 (2023)
- Wu, Y. K., Miehl, C. & Gjorgjieva, J. Regulation of circuit organization and function through inhibitory synaptic plasticity. *Trends in Neurosciences* **45,** 884–898 (2022)
- Wu, Y. K. & Zenke, F. Nonlinear transient amplification in recurrent neural networks with short-term plasticity. *eLife* **10,** e71263 (2021)
- Wu, Y. K., Hengen, K. B., Turrigiano, G. G. & Gjorgjieva, J. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 24514–24525 (2020)
- Pacheco, A. T., Tilden, E. I., Grutzner, S. M., Lane, B. J., Wu, Y., Hengen, K. B., Gjorgjieva, J. & Turrigiano, G. G. Rapid and active stabilization of visual cortical firing rates across light-dark transitions. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 18068–18077 (2019)

## Peer-reviewed poster presentations

- Wu, Y. K., Gjorgjieva, J. & Ahmadian, Y. Impact of second-order connectivity motifs on neural dynamics in low-rank stabilized supralinear networks. *Poster presentation at Bernstein Conference, Berlin, Germany* (2023)
- Kepçe, A., Wu, Y. K. & Gjorgjieva, J. Cell-type-specific synaptic scaling mechanisms in associative learning. *Poster presentation at Bernstein Conference, Berlin, Germany* (2023)

- Wu, Y. K., Waitzmann, F. & Gjorgjieva, J. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity. *Poster presentation at Bernstein Conference, Berlin, Germany* (2022)
- Wu, Y. K. & Gjorgjieva, J. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity. *Poster presentation at FENS Conference, Paris, France* (2022)
- Wu, Y. K., Waitzmann, F. & Gjorgjieva, J. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity. *Poster presentation at COSYNE Conference, Lisbon, Portugal* (2022)
- Wu, Y. K., Friedrich, R. W. & Zenke, F. Auto-associative memory without persistent activity in networks with co-tuned excitation and inhibition. *Poster presentation at COSYNE Conference, Online* (2021)
- Wu, Y. K., Hengen, K. B., Turrigiano, G. G. & Gjorgjieva, J. Circuit Dynamics and Plasticity during Activity Deprivation in Visual Cortex of Freely Behaving Rodents. *Poster presentation at COSYNE Conference, Lisbon, Portugal* (2019)

## Oral presentations

- Wu, Y. K. Top-Down Modulation in Canonical Cortical Circuits with Inhibitory Short-Term Plasticity. *Online talk at Imperial College London in the lab of Dr. Sadra Sadeh* (2023)

# Acknowledgments

I would like to express my greatest gratitude to my supervisor Julijana Gjorgjieva, for teaching me how to approach research questions, conduct rigorous research, give great presentations, and write clear, precise, and easy-to-understand manuscripts. She is a great role model and cares about details while keeping big pictures in mind. Over the years, she has influenced me greatly on how to do science and beyond. I thank her for always being so supportive.

I thank Gina Turrigiano with whom I worked on the homeostatic regulation project. It was my first project in computational neuroscience. I was incredibly lucky to be able to work with her and work on this collaborative project. I have learned a lot from the project. I thank her for supporting me along the way.

I thank Yashar Ahmadian for hosting me in Cambridge. It was a unique experience for me. I very much enjoyed working with him. I thank Julia Veit for giving me the great opportunity to visit her lab and exchange ideas between theorists and experimentalists. I thank Friedemann Zenke for working on the NTA project together and teaching me how to write good manuscripts. I thank the members of my thesis advisory committee, Srdjan Ostojic and Tilo Schwalger for their helpful comments and suggestions. Special thanks to Srdjan Ostojic for his support along my PhD journey from Switzerland to Germany.

I thank old lab members who accompanied me in Frankfurt, Christoph Miehl, Juan Luis Riquelme, Samuel Eckmann, Sebastian Onasch, and Leonidas Richter. I thank them for the good old days and those memorable conversations on science, life, and philosophy. I thank all current lab members with whom I spent time together in Munich. I feel privileged to be surrounded by a group of nice and supportive colleagues. I thank Felix Waitzmann with whom I work closely in the past two years. It had been a great pleasure working with him. I thank Elizabeth Herbert for providing useful feedback on my thesis and previous manuscripts. I thank Tianlin Liu and Adam Carte for the great time in Switzerland. I thank Giulio Bondanelli, Natalie Schieferstein, Roxana Zeraati, Matyas Varadi, and Yuxiu Shao for the scientific exchange, support and encouragement.

Sometimes, life is not as easy as we want it to be. I also would like to thank the

people who brought me not only joy but also pain. I have learned to let it be, then let it go.

Finally, I would like to thank my parents for their endless love, support and encouragement.

This thesis is dedicated to my parents.

# Appendix

## I. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics

Wu, Y. K., Hengen, K. B., Turrigiano, G. G. & Gjorgjieva, J. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 24514-24525 (2020). https://doi.org/10.1073/pnas.1918368117

# Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics

Yue Kris Wu[a], Keith B. Hengen[b,c], Gina G. Turrigiano[b,1], and Julijana Gjorgjieva[a,d,1]

[a]Computation in Neural Circuits Group, Max Planck Institute for Brain Research, 60438 Frankfurt, Germany; [b]Department of Biology, Brandeis University, Waltham, MA 02454; [c]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130; and [d]School of Life Sciences, Technical University of Munich, 85354 Freising, Germany

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved August 04, 2020 (received for review October 20, 2019)

**Homeostasis is indispensable to counteract the destabilizing effects of Hebbian plasticity. Although it is commonly assumed that homeostasis modulates synaptic strength, membrane excitability, and firing rates, its role at the neural circuit and network level is unknown. Here, we identify changes in higher-order network properties of freely behaving rodents during prolonged visual deprivation. Strikingly, our data reveal that functional pairwise correlations and their structure are subject to homeostatic regulation. Using a computational model, we demonstrate that the interplay of different plasticity and homeostatic mechanisms can capture the initial drop and delayed recovery of firing rates and correlations observed experimentally. Moreover, our model indicates that synaptic scaling is crucial for the recovery of correlations and network structure, while intrinsic plasticity is essential for the rebound of firing rates, suggesting that synaptic scaling and intrinsic plasticity can serve distinct functions in homeostatically regulating network dynamics.**

homeostasis | cortical circuits | functional correlation | synaptic scaling | intrinsic plasticity

Neural circuits are faced with a fundamental problem: how to allow experience to alter and refine network connectivity during learning and experience-dependent plasticity, while still maintaining stability of function. Generating a neural system that is both stable and flexible is a nontrivial challenge and requires a prolonged period of development when multiple mechanisms at the level of single neurons and networks of neurons interact. Two powerful and fundamentally different forms of plasticity involved in this process are Hebbian mechanisms, which alter synaptic connectivity in a synapse-specific manner, and homeostatic mechanisms that maintain stable function by globally adjusting overall synaptic weights and neuronal excitability.

The development and refinement of visual response properties in the primary visual cortex (V1) involves classic synapse-specific mechanisms implementing the bidirectional form of Hebbian plasticity, such as long-term potentiation (LTP) and long-term depression (LTD), considered to be the cellular substrate for learning and memory (1). Associative Hebbian plasticity, however, drives positive feedback processes that lead to unstable network dynamics, and some form of homeostasis is needed to compensate for this inherent instability (2, 3). A large body of evidence shows that various homeostatic plasticity mechanisms, including synaptic scaling and intrinsic plasticity (4, 5), operate in the brain to maintain stability despite various internal and external perturbations. More specifically, homeostatic plasticity can elevate neural activity in response to sensory deprivation (6, 7) and suppress activity under conditions of overexcitation (8, 9).

Despite great efforts to describe homeostatic mechanisms at the single cell level, how network properties are homeostatically regulated is largely unknown. While Hebbian and homeostatic mechanisms operate at different timescales and can be induced by distinct cues (10–13), how they interact within complex, highly recurrent microcircuits, as those found in the cortex, to refine and maintain circuit function has remained elusive. A critical challenge has been the lack of detailed measurements of individual synaptic strengths and their potential impact on large-scale network dynamics, especially in a highly recurrent network like the cortex.

Here, we investigate two main questions. First, which aspects of network function are under homeostatic control? Second, why are there so many homeostatic mechanisms, and do they serve redundant or unique functions? To address these questions, we combine analysis of in vivo electrophysiological data during sensory deprivation in the rodent visual cortex and computational modeling of cortical synaptic plasticity and network dynamics. First, we analyzed the collective activity of multiple neurons in the monocular region of the primary visual cortex (V1m) during a classic monocular deprivation (MD) paradigm (lid suture) in freely behaving rats over 9 d during the critical period (14). Earlier work demonstrated that MD induces an initial drop in firing followed by the rates' homeostatic recovery despite long-lasting deprivation (14). Here, we reanalyzed these datasets to characterize the temporal evolution of higher-order network properties over the same 9-d period. Individual pairwise correlations, including correlation structure, weakened during brief MD but recovered during prolonged MD. Second, to understand how the cortical network exploits diverse homeostatic mechanisms to return firing rates and correlations to baseline (BL) after prolonged MD, we took advantage of a plastic spiking recurrent network model equipped with known plasticity and homeostatic mechanisms. Our work suggests that synaptic scaling is crucial

## Significance

**Despite decades of intense studies on homeostasis, network properties undergoing homeostatic regulation remain elusive. Furthermore, whether diverse forms of homeostatic plasticity are simply redundant or serve distinct functions is unclear. Here, our data show that functional correlations are subject to homeostatic regulation, both in terms of average amplitude and their structure. A computational model demonstrates that synaptic scaling is essential for the restoration of correlations and network structure, whereas intrinsic plasticity is crucial for the recovery of firing rates after perturbations, suggesting that synaptic scaling and intrinsic plasticity distinctly contribute to homeostatic regulation.**

Author contributions: Y.K.W. and J.G. designed research; Y.K.W. performed research; Y.K.W., K.B.H., and J.G. contributed new reagents/analytic tools; Y.K.W. and J.G. analyzed data; Y.K.W. and J.G. wrote the paper; K.B.H. and G.G.T. provided experimental data; and G.G.T. provided input during writing of the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.



[1]To whom correspondence may be addressed. Email: turrigiano@brandeis.edu or gjorgjieva@brain.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1918368117/-/DCSupplemental.

First published September 11, 2020.

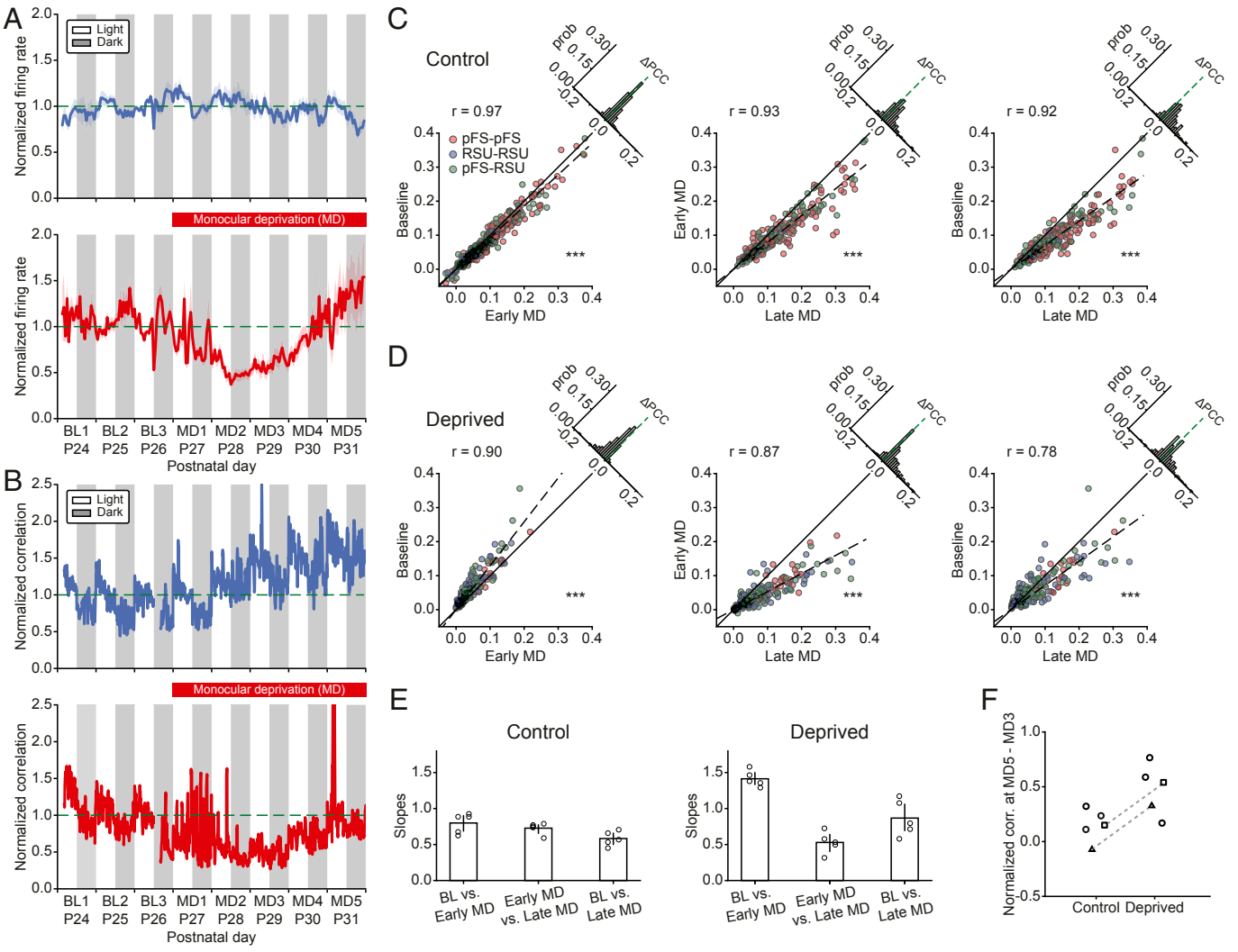


**Fig. 1.** MD induces an initial drop in correlations followed by their homeostatic recovery. (*A*) The average firing rates of 80 neurons from 5 control hemispheres (*Top*) and 104 neurons from 5 deprived hemispheres (*Bottom*) normalized to the firing rates at P26 in the light (horizontal dashed line). (*B*) The average pairwise correlations of 970 pairs from 5 control hemispheres (*Top*) and 2,455 pairs from 5 deprived hemispheres (*Bottom*) normalized to the correlations at P26 in the light (horizontal dashed line). (*C*) Correlation comparisons between BL and early MD (*Left*), between early MD and late MD (*Middle*), and between BL and late MD (*Right*) at the single cell-pair level of one control hemisphere. Different colors represent the correlations between different neuron types. Dashed lines are fitted regression lines crossing the origin. Upper left histograms indicate the distributions of correlation differences. $***P < 0.001$ (Wilcoxon signed-rank test). (*D*) Same as *C* but for one deprived hemisphere. Here, for two hemispheres, we used MD3, and for the other three hemispheres, we used MD2, as early MD because different animals showed the biggest drop in correlations at different times. $***P < 0.001$ (Wilcoxon signed-rank test). (*E*) Slopes of fitted regression lines for the correlation comparisons as in *C* and *D* for five control and five deprived hemispheres. (*F*) Change in the average normalized correlations between MD3 and MD5. Each data point denotes one hemisphere. Hemispheres from the same animals are marked with square or triangle symbols and connected by a dashed line. Data are shown as means ± SEM.

for the recovery of correlations and network structure, whereas intrinsic plasticity is essential for the rebound of firing rates. These results indicate that different homeostatic mechanisms act in the brain to independently regulate distinct network features.

## Results

**Pairwise Correlations during the Critical Period and in Response to MD.** We first confirmed previous analysis of individual neurons recorded in vivo in the primary visual cortex during the critical period of plasticity (postnatal day [P]24 to P32). In these experiments, MD was performed after 3 d of BL activity and continued for the rest of the recordings. While firing rates of individual neurons remained relatively stable during normal development (Fig. 1*A*, *Top*), brief 2-d MD caused the firing rates to decrease to 40% of their BL values (Fig. 1*A*, *Bottom*) (6, 14). However, despite prolonged MD, over the next 3 to 4 d, firing rates gradually recovered to BL after an initial overshoot (Fig. 1*A*, *Bottom*) (6, 14). These effects were not only observed at the population level but also at the level of individual neurons (14). Here, we investigated higher-order network properties during normal development and following prolonged MD by calculating the next statistical moment beyond the firing rates, namely the pairwise spiking correlations between different neuron types (*Methods*). Specifically, we quantified the temporal evolution of the correlation coefficient of individual neuron pairs and of the average correlations across all pairs both during normal development and after perturbing visual input through MD. In control hemispheres, correlations, unlike firing rates, increased slightly as a function of age ($n = 5$ animals; Fig. 1*B*, *Top*). By contrast, in deprived hemispheres, correlations initially dropped over the first 2 d and then gradually rebounded to predeprivation levels ($n = 5$ animals; Fig. 1*B*, *Bottom*), displaying a similar pattern as the firing rates (Fig. 1*A*). As previously reported, we observed light–dark oscillations in the correlation

amplitudes, with higher correlations in the light and lower correlations in the dark (15).

To assess the degree to which correlations of individual neuron pairs changed beyond the population level, we evaluated single cell-pair correlations on different days. As in the earlier analysis, neurons were separated into putative parvalbumin-positive ($PV^+$) fast-spiking units (pFS) or regular-spiking units (RSUs) based on waveform and spiking characteristics (6, 14). Specifically, we focused on three different 12-h periods recorded in the light: 1) BL corresponding to P26; 2) a period that we called "early MD" when the largest drop of firing rates and correlations occurred, typically 2 or 3 d after BL (i.e., P28 or P29); and 3) a period that we called "late MD" corresponding to the time when the firing rates and correlations nearly recovered to BL (i.e., P31). As observed already for the average correlations, when combining all neuron pairs and animals, single cell-pair correlations increased during normal development covering the 6-d period during which recordings were performed. The increase between BL and early MD was small ($n = 435$ pairs; Fig. 1*C*, *Left*) ($r = 0.97$; $P < 10^{-21}$ [Wilcoxon signed-rank test]). Correlations at late MD were significantly greater than at early MD ($n = 253$ pairs; Fig. 1*C*, *Middle*) ($r = 0.93$; $P < 10^{-28}$ [Wilcoxon signed-rank test]). The developmental increase in correlations during the critical period became most obvious when we compared BL versus late MD ($n = 253$ pairs; Fig. 1*C*, *Right*) ($r = 0.92$; $P < 10^{-37}$ [Wilcoxon signed-rank test]). We did not observe any obvious differences in correlations among different cell types in that they all showed similar patterns of temporal evolution. Moreover, almost all neuronal pairs in a control hemisphere demonstrated an increase in correlation (Fig. 1*C*, *Right*).

Conversely, in deprived hemispheres, correlations of the majority of individual cell pairs, independent of their type, underwent a significant drop during early MD ($n = 190$ pairs; Fig. 1*D*, *Left*) ($r = 0.90$; $P < 10^{-24}$ [Wilcoxon signed-rank test]), followed by an increase during late MD ($n = 231$ pairs; Fig. 1*D*, *Middle*) ($r = 0.87$; $P < 10^{-27}$ [Wilcoxon signed-rank test]). The correlations during late MD recovered to a higher level than BL ($n = 190$ pairs; Fig. 1*D*, *Right*) ($r = 0.78$; $P < 10^{-4}$ [Wilcoxon signed-rank test]). We summarized the gradual increase of correlations in control hemispheres and the drop followed by recovery in deprived hemispheres by the slopes of the fitted regression lines of the individual pair data for each animal (Fig. 1*E*). Remarkably, despite a degree of variability across animals, the drop and recovery of correlations induced by MD were ubiquitous (*SI Appendix*, Fig. S1).

There are several possible mechanisms for the recovery of correlations during late MD in deprived hemispheres. First, it is possible that the correlations in the cortex simply follow the firing rates (16), which are homeostatically regulated. However, this scenario assumes a feedforward framework of signal transmission in which input correlations are fixed. In our experiments, input correlations under conditions of normal vision consist of a combination of signal and noise correlations. Closure of the eye during MD destroys signal correlations, thus decreasing overall input correlations, even though thalamic firing rates do not change during MD (17). Therefore, the only source of input correlations during prolonged MD is noise correlations. Under normal vision, cortical correlations in the dark (driven by noise input correlations) are approximately two-thirds of the correlations in the light (driven by intact signal input correlations) (15) (Fig. 1*B*). Combining these two results supports the conclusion that the homeostatic recovery of firing rates cannot explain the full recovery of cortical correlations and that network mechanisms are likely involved.

A second possibility suggests that the increase of correlations in deprived hemispheres could arise from the same underlying, possibly developmental, mechanism as in control hemispheres. To investigate this, we compared the increase in the average correlations between early MD, when the largest drop in the correlations occurs, and late MD, when the correlations have mostly recovered. We found that the increase of correlations in deprived hemispheres was consistently higher than in control hemispheres (Fig. 1*F*). This suggests that the increase of correlations in deprived hemispheres does not only have a developmental, age-dependent component but also a homeostatic recovery component in response to prolonged MD. We further found that the correlation changes between two adjacent 12-h light periods, as quantified by the slopes of the fitted regression lines, were different in the deprived from the control hemisphere in the same animals (*SI Appendix*, Fig. S2). This indicates that the increase in correlations between P29 and P31, the period corresponding to late MD, follows different temporal dynamics in control and deprived hemispheres.

Taken together, our results demonstrate an increase of correlations in deprived hemispheres during prolonged MD that is larger than the developmental increase of correlations in control hemispheres during normal development. Excluding other mechanisms such as coregulation with firing rates and age dependence, we propose that the recovery of correlations in deprived hemispheres during prolonged MD is due to homeostatic mechanisms, which are well known to operate in response to such perturbations (4, 5).

**Network Structure after MD.** While correlations at the single cell-pair level recovered during late MD, the difference between correlations at late MD and BL (Fig. 1*D*, *Right*) raised the possibility that the recovered network might have a different structure after recovery. To examine the evolution of network structure during normal development over the critical period and during prolonged MD, we examined the correlation matrices on different days. An example experiment shows that in the control hemisphere, the structure of the correlation matrix remained consistent over time ($n = 11$ neurons; Fig. 2*A*), whereas in the deprived hemisphere, the correlation structure initially weakened and recovered to a similar structure as BL ($n = 14$ neurons; Fig. 2*B*). MD induced heterogeneous changes in correlation structure across animals, despite an overall initial decrease and subsequent recovery (*SI Appendix*, Fig. S3). To quantify the similarity between the structure of correlation matrices at distinct time points, we calculated the $L1$ distance between correlations (*Methods*), which measures the absolute difference between them. Combining multiple animals revealed that in both control and deprived hemispheres, the correlation matrix at BL is more similar to the correlation matrix at early MD relative to randomly shuffling the latter for control ($n = 609$ pairs; Fig. 2*C*, *Left*) ($P < 10^{-58}$ [Wilcoxon signed-rank test]) and deprived hemispheres ($n = 505$ pairs; Fig. 2*D*, *Left*) ($P < 10^{-33}$ [Wilcoxon signed-rank test]). Additionally, the correlation structures at BL and late MD are more similar than chance level for control ($n = 609$ pairs; Fig. 2*C*, *Right*) ($P < 10^{-28}$ [Wilcoxon signed-rank test]) and deprived hemispheres ($n = 505$ pairs; Fig. 2*D*, *Right*) ($P < 10^{-30}$ [Wilcoxon signed-rank test]). These results suggest that despite a decrease in the correlation amplitude during early MD, the correlation structure is maintained throughout MD; hence, the network does not reorganize as correlations recover during late MD. In line with this finding, we found that the distance between BL and early MD in deprived hemispheres was not significantly different from that between BL and late MD ($n = 505$ pairs; Fig. 2*D*) ($P = 0.771$ [Wilcoxon signed-rank test]). However, for control hemispheres, the distance between BL and late MD was significantly higher than that between BL and early MD ($n = 609$ pairs; Fig. 2*C*) ($P < 10^{-46}$ [Wilcoxon signed-rank test]), due to the large increase in correlation amplitude during development (Fig. 1*C*). Interestingly, the correlation matrices

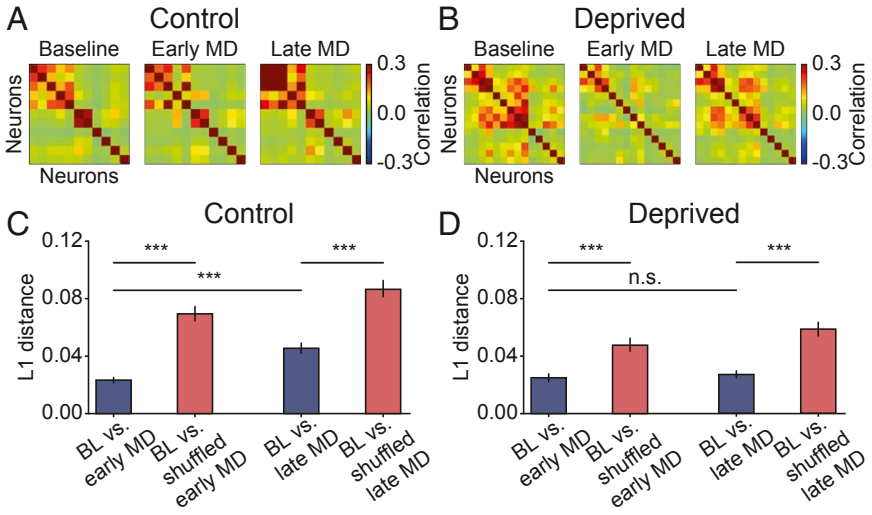


**Fig. 2.** Structure of correlation matrices is maintained after recovery. (*A*) Example correlation matrix of 11 neurons from 1 control hemisphere at 3 different time points. (*B*) Same as *A* but of 14 neurons from 1 deprived hemisphere. (*C*) *L*1 distance between correlations at BL and early MD, and at BL and late MD, vs. shuffled data for control hemispheres. (*D*) Same as *C* for deprived hemispheres. Data are shown as means ± SEM. ***$P < 0.001$; n.s., not significant ($P > 0.05$) (Wilcoxon signed-rank test).

are composed of several assemblies—groups of neurons exhibiting strong pairwise correlations (Fig. 2 *A* and *B*)—reminiscent of the clustered network structure reported in previous studies (18–21).

In conclusion, our analysis of V1 cortical activity recorded in vivo demonstrates that the pairwise correlations, in amplitude and structure, of these networks are homeostatically regulated following prolonged perturbation of normal sensory experience.

**Formation of Structured Connectivity Assemblies during Training in a Recurrent Network Model.** We next asked what mechanisms underlie the observed neuronal- and network-level changes during normal development and following a perturbation like MD. To understand how neural circuits exploit various synaptic plasticity and homeostatic mechanisms to first decrease and then recover both firing rates and correlations during MD, we built a plastic recurrent network model consisting of randomly connected excitatory and inhibitory spiking neurons (*Methods*). Model neurons received thalamic inputs, with thalamocortical synaptic efficacy onto inhibitory neurons set higher than onto excitatory neurons, consistent with previous experimental studies (22–24). Neuronal and network parameters were chosen to generate in vivo-like firing rates, with excitatory neurons firing at 5 Hz and inhibitory neurons firing at 13 Hz (6).

To generate the experimentally observed clustered correlation structure (Fig. 2 *A* and *B*), we included several experimentally characterized plasticity mechanisms (25) (*Methods*). We first tasked the network with the imprinting of connectivity assemblies starting from an initially random connectivity (Fig. 3*A*, *Left*). In contrast to previous models that used random, uncorrelated Poisson inputs (25) and in line with our observation that the networks show stronger pairwise correlations in the light than in the dark (15), we postulated that input correlations—as would be generated during natural vision—matter for the generation of clustered connections. Therefore, we trained the recurrent network by stimulating excitatory neurons with thalamocortical Poisson spiking inputs that had identical firing rates but differed in their correlation structures. For the training, excitatory neurons were randomly grouped into four identical assemblies, thereby simplifying network structure despite known heterogeneities in the data (*SI Appendix*, Fig. S3).

Before training with correlated inputs, the initial synaptic connections in the entire network were weak and identical between any pair of neurons of the same type (Fig. 3*A*, *Left*), resulting in asynchronous irregular network activity (Fig. 3*A*, *Middle*) and low correlations without clustered structure (Fig. 3*A*, *Right*). During training, excitatory neurons within a targeted assembly

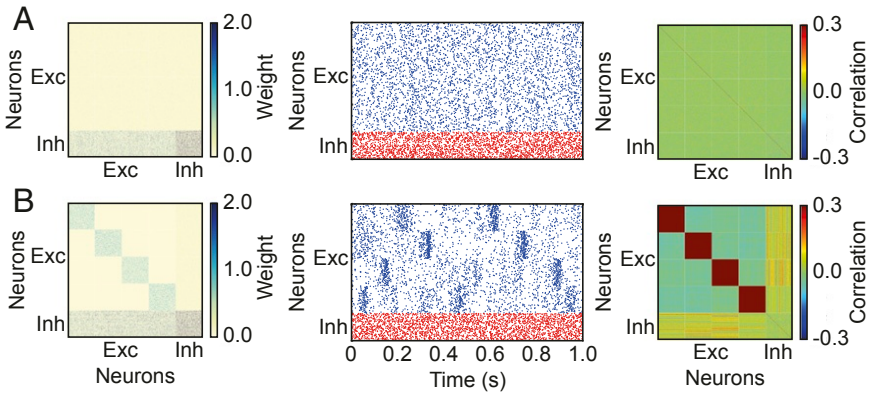


**Fig. 3.** Imprinting-connectivity assemblies with correlated inputs. (*A*) Before training: connectivity matrix (*Left*), spontaneous activity of excitatory (blue) and inhibitory (red) neurons (*Middle*), and correlation matrix (*Right*). (*B*) Same as *A* but after training.

received correlated inputs (*Methods*), which strengthened connectivity between them through Hebbian plasticity. After training, the excitatory subnetwork became structured with stronger synaptic connections between excitatory neurons within assemblies, while inhibitory neurons remained unstructured whereby inhibition is global and nonspecifically connects to all excitatory neurons (Fig. 3*B*, *Left*). As a result of this structure, the network no longer exhibited asynchronous irregular activity but rather blocks of activity in the excitatory neurons defined as occasional periods of high firing rate (Fig. 3*B*, *Middle*). The structured connectivity and block activity selectively generated high correlations between excitatory neurons within assemblies (Fig. 3*B*, *Right*).

**A Model with Persistent Hebbian LTD and Homeostatic Plasticity Cannot Recover Correlations after MD.** Using the structured model network as a baseline following normal cortical development after eye opening, we next wanted to investigate how this network responds to a sensory perturbation resembling MD. To achieve this, we needed to know how the inputs to the network are modified during MD. Previous experimental studies have reported that MD induces no change in the average firing rates of LGN, the visual area of the thalamus (17). Therefore, to simulate MD in our model network, we kept the firing rates of LGN inputs identical to that at BL but assumed that eye closure during MD considerably diminished input correlations. In the model, the excitatory neurons received uncorrelated Poisson inputs to denote the start of MD (Fig. 4*A*).

In addition to these changes in input correlations, recent experiments have revealed that brief MD (2 d) induces LTD at thalamocortical synapses onto excitatory and inhibitory neurons, with thalamocortical synapses onto inhibitory neurons depressing more than synapses onto excitatory neurons (24). The process of LTD is not instantaneous, so we assumed that synaptic connections from the thalamus to excitatory and inhibitory neurons undergo a linear decrease during the first 2 d of MD. To match experimental findings, the decrease in thalamocortical connections onto inhibitory neurons was larger (Fig. 4*A* and *Methods*). It is currently unknown when during MD this thalamocortical depression saturates, but since deprived-eye responsiveness reaches its minimum 2 to 3 d after the onset of MD (26), we assumed that the feedforward connections did not further decrease after this point, while keeping the inputs uncorrelated for the entire MD (Fig. 4*A*).

How does the recurrent network respond to these changes in input correlation structure and depression of feedforward connectivity strength that occur following MD? Although there are potentially multiple ways to achieve network stability and regulate network function, there are two fundamentally different mechanisms that have been well characterized experimentally: homeostatic adjustment of synaptic strengths and of intrinsic excitability (3, 27, 28). Excitatory neurons can regulate their activity by scaling incoming synaptic strengths in response to perturbations—a process known as synaptic scaling (4). This scaling is bidirectional in that it can increase and decrease synaptic strengths; it is global and operates in a multiplicative manner. In addition to synaptic scaling, neurons can alter the number of different ion channels to adjust intrinsic excitability, and consequently modify their firing thresholds, in response to perturbations (5, 13, 29).

Based on these experimental findings, in addition to Hebbian plasticity during training, we modeled these two distinct homeostatic mechanisms following MD: 1) synaptic scaling, which acts only on excitatory synapses (4, 6); and 2) intrinsic plasticity, which modifies the intrinsic excitability of both excitatory and inhibitory neurons (29, 30) (*Methods*). In the presence of persistent thalamocortical LTD, as during training, and both homeostatic mechanisms, the average firing rates of excitatory

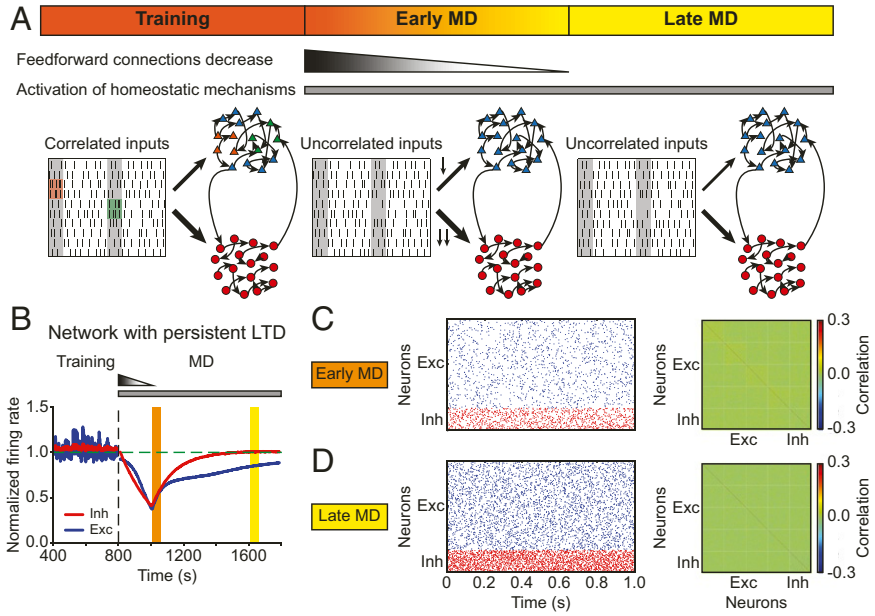


**Fig. 4.** The model with persistent Hebbian LTD and homeostatic plasticity fails to recover correlations. (*A*) Schematic description of the modeling framework during MD. The network consists of 80% excitatory neurons (triangles) and 20% inhibitory neurons (circles). Both of them receive thalamic inputs, and thalamocortical connections onto inhibitory neurons are stronger than onto excitatory neurons. Neurons receive correlated input during training at BL (*Left*) but uncorrelated input during early MD (*Middle*) and late MD (*Right*). During early MD, thalamocortical connections onto both excitatory and inhibitory neurons are depressed following a linear function in time, with thalamocortical connections onto inhibitory neurons depressed more strongly. The connections remain fixed during late MD. (*B*) The average normalized firing rates of excitatory (blue) and inhibitory (red) neurons. The vertical dashed line indicates the onset of MD. The horizontal dashed line indicates a normalized firing rate of 1.0. (*C*, *Left*) Spontaneous activity of excitatory (blue) and inhibitory (red) neurons during early MD. (*C*, *Right*) Correlation matrix during early MD indicated by the orange region in *B*. (*D*) Same as *C* but during late MD indicated by the yellow region in *B*.

and inhibitory neurons in the model network first decreased to 40% of BL, because slow homeostatic mechanisms could not overcome the feedforward synaptic depression and input decorrelation to recover firing rates. At the time that feedforward LTD saturated, firing rates started to increase due to homeostatic plasticity, resembling the recovery to BL observed experimentally during late MD (Fig. 4*B*; compare with Fig. 1*A*, *Bottom*).

Next, we investigated the evolution of higher-order aspects of network dynamics. Similar to the analysis of our data, we focused on two key time points after MD onset in the model: early MD, corresponding to the largest drop of firing rates (Fig. 4*B*, orange); and late MD, corresponding to the time when the firing rates recovered close to BL (Fig. 4*B*, yellow). The network showed irregular spiking dynamics with different firing rates during these two periods (Fig. 4 *C* and *D*, *Left*). The correlations between excitatory neurons first decreased during the period modeling early MD, as observed experimentally (Fig. 4*C*, *Right*; compare with Fig. 2*B*, *Middle*), but did not recover during the period corresponding to late MD (Fig. 4*D*, *Right*; compare with Fig. 2*B*, *Right*). We speculated that this failure to recover the correlations in the model network, despite the recovery of firing rates, could be the result of perturbing the structured connectivity between excitatory neurons within assemblies generated through training (Fig. 3*B*). Indeed, the average weights between excitatory neurons within an assembly depressed during the period corresponding to late MD (*SI Appendix*, Fig. S4).

To reveal the origin of this depression in the model network, we investigated the specific contribution of Hebbian plasticity and synaptic scaling to the average excitatory weight change within assemblies. Despite the overall potentiation of excitatory weights within assemblies induced by synaptic scaling during the period corresponding to early MD, continued LTD from Hebbian plasticity dominated over homeostatic plasticity, depressing all excitatory weights within assemblies and preventing the recovery of excitatory-to-excitatory correlations (*SI Appendix*, Fig. S5). In conclusion, this dominance of depression after MD prevents the recovery of structured connectivity, and consequently correlations, between excitatory neurons in a model with persistent Hebbian LTD despite homeostatic plasticity. This suggests that the relative timing and resulting competition between the two homeostatic mechanisms and ongoing Hebbian plasticity could be important for recovering different aspects of network dynamics.

**The Attenuation of Hebbian LTD Together with Homeostatic Mechanisms Restores Firing Rates and Correlations during Prolonged MD.** Previous work involving ocular dominance plasticity has shown that blocking Hebbian plasticity under normal rearing or after 6 d of MD does not cause any significant change in the response strength in the binocular region of V1, suggesting that the effects of Hebbian and homeostatic plasticity are negligible at each of the two steady states. These experiments also argued that the total effect of Hebbian plasticity in the deprived eye during the recovery phase is dominated by LTD but gradually approaches zero when homeostatic plasticity reaches its steady state (31). Motivated by these findings, we asked whether the recovery of excitatory correlations during the period corresponding to late MD in the model can be rescued by reducing the effect of Hebbian LTD. We proposed that the attenuation of Hebbian plasticity might occur through a metaplastic process where the amplitude of LTD dynamically adapts to the history of neuronal activity (*Methods*) (32, 33). Implementing metaplastic LTD preserved the recovery of average firing rates of both excitatory and inhibitory neurons (Fig. 5*A*). Similarly, the spiking rasters during the period corresponding to early MD showed asynchronous irregular activity (Fig. 5*B*, *Left*). In contrast to the model with persistent LTD, however, the metaplastic reduction in LTD enabled the return of structured excitatory activity during late MD (Fig. 5*C*, *Left*). Importantly, the excitatory correlation structure in the model during late MD homeostatically recovered after its initial dilution during early MD (Fig. 5*B* and Fig. 5*C*, *Right*; compare with Fig. 2*B*, *Middle* and *Right*). The decrease and recovery of correlations was the same across all neuron pairs within assemblies in our model because the trained assemblies were identical, unlike the heterogeneity in the data where the correlations of different neuron pairs

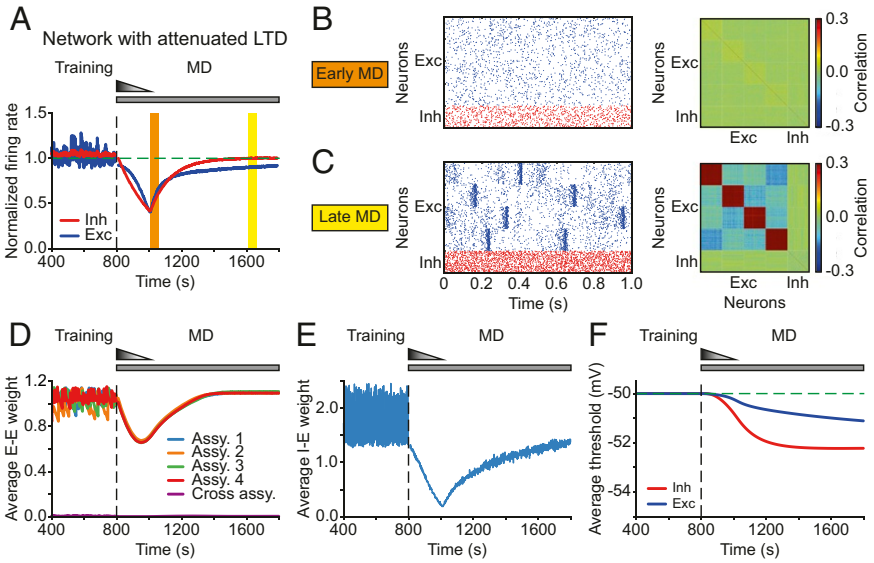


**Fig. 5.** The model with attenuated LTD recovers excitatory and inhibitory firing rates and excitatory correlations during MD. (*A*) The average normalized firing rates of excitatory (blue) and inhibitory (red) neurons. The vertical dashed line indicates the onset of MD. The horizontal dashed line indicates a normalized firing rate of 1.0. (*B*, *Left*) Spontaneous activity of excitatory (blue) and inhibitory (red) neurons during early MD. (*B*, *Right*) Correlation matrix during early MD indicated by the orange region in *A*. (*C*) Same as *B* but during late MD indicated by the yellow region in *A*. (*D*) Average excitatory-to-excitatory weights for each assembly and across assemblies. (*E*) Average inhibitory-to-excitatory weights that target all excitatory neurons independent of assembly membership. (*F*) Average firing thresholds of excitatory (blue) and inhibitory (red) neurons. The horizontal dashed line indicates the initial firing threshold.

underwent a varying degree of decrease and recovery (*SI Appendix*, Fig. S3). Adding heterogeneity to the model assemblies—for instance, by diversifying synaptic strengths, connectivity probabilities, or sizes—might be necessary to capture the diverse changes in correlation structures in the data. The metaplastic down-regulation of LTD in our model shifted the network from an LTD-dominant regime during early MD to an LTP/LTD-balanced regime during the recovery phase. Although this regime differs from previous studies in which the network remains in an LTD-dominant regime during most of the recovery phase (31), firing rates and correlations will recover provided that homeostatic plasticity greatly dominates over Hebbian LTD during the recovery phase.

We further investigated what other properties of the network changed as we modeled MD. Along with firing rates and excitatory correlations, the average excitatory weights within assemblies manifested the same pattern of drop and rebound (Fig. 5*D*), in contrast to the corresponding weights in the initial model with persistent LTD (*SI Appendix*, Fig. S4). Average inhibitory onto excitatory weights also decreased during early MD in the model (Fig. 5*E*), suggesting that the network reduced the amount of inhibition to elevate the decreased firing rates of excitatory neurons. During the period corresponding to late MD, overall inhibition increased to balance the gradually recovered excitation, keeping excitatory–inhibitory balance and avoiding winner-take-all dynamics where a single strongly connected assembly dominates the entire network (25). Furthermore, the average firing thresholds of excitatory and inhibitory neurons in the model network decreased as we modeled prolonged MD and reached a steady state as the firing rates approached their BL values (Fig. 5*F*).

Our experimental analysis revealed that, despite a decrease in the correlation amplitude during early MD, correlation structure is maintained throughout MD (Fig. 2*D*). Consistent with this, if network structure is completely erased during early MD in our model (as in the scenario without metaplastic LTD; Fig. 4), then homeostatic synaptic scaling during late MD cannot recover excitatory correlations because the backbone of recurrent circuitry from which to rebuild them has been lost. Otherwise, synaptic scaling can still rescue correlations even when the intensity or duration of LTD at thalamocortical synapses increases (*SI Appendix*, Fig. S6). In particular, we found that the intensity and duration of feedforward LTD have a different impact on the excitatory synaptic weights within assemblies, which shape excitatory correlation structure in the model. More intense and prolonged LTD causes a larger decrease in the firing rates, enabling the fast upscaling of excitatory synaptic weights within assemblies that recover correlations well before firing rates (*SI Appendix*, Fig. S6). Only prolonging feedforward LTD without affecting its intensity does not decrease firing rates as much (*SI Appendix*, Fig. S6), due to the lower firing thresholds of the neurons (*SI Appendix*, Fig. S7). The smaller drop in firing rates constrains the amount of synaptic upscaling, resulting in weaker excitatory correlation structure during the recovery phase. Consequently, network connectivity and correlations recover later than firing rates (*SI Appendix*, Fig. S6). These results suggest that correlation changes do not necessarily follow firing rate changes but are the product of interacting homeostatic mechanisms at the network level.

In summary, metaplastic regulation of LTD, together with synaptic scaling and intrinsic plasticity, is sufficient to capture both the recovery of excitatory and inhibitory firing rates and excitatory correlations during MD. Maintaining network structure during early MD is necessary for synaptic scaling to recover correlation structure during late MD. Hence, homeostatic modifications of overall synaptic weights and intrinsic excitability cooperate with Hebbian LTD to recover several aspects of network function following input perturbations.

**Individual Homeostatic Mechanisms Have Different Functionality during MD.** To determine the distinct contributions of the different homeostatic mechanisms for the recovery of firing rates and correlations during prolonged MD, we selectively eliminated each mechanism. When deactivating synaptic scaling during the entire period of MD in the model, we found that excitatory and inhibitory firing rates still recovered (Fig. 6*A*), whereas the excitatory correlations did not (Fig. 6*C*). Since synaptic scaling affects excitatory synaptic strengths, we hypothesized that the correlations failed to recover due to the inability of the network to recover its structured excitatory connectivity. Indeed, the average weights between excitatory neurons within assemblies remained low in the absence of synaptic scaling (*SI Appendix*, Fig. S8), eliminating structured block activity (Fig. 6*B*) and preventing the recovery of excitatory correlation structure during late MD (Fig. 6*C*). This suggests that synaptic scaling on excitatory synapses is indispensable for the recovery of excitatory correlations.

Similarly, without intrinsic plasticity during the entire MD period, neither excitatory nor inhibitory firing rates in the model recovered (Fig. 6*D*). This result was independent of the recovery of correlations. When the overall excitatory drive received by a single neuron within the same assembly was weak, low firing rates were accompanied by a poor degree of synchrony within assemblies (Fig. 6*E*), resulting in weak correlations (Fig. 6*F*). Increasing the overall excitation to a neuron, for instance, by increasing the connectivity probability within assemblies, could still generate structured block activity resulting in high correlations within assemblies but without recovering firing rates, especially for inhibitory neurons.

In conclusion, we demonstrated that two important forms of homeostatic plasticity, synaptic scaling and intrinsic homeostatic plasticity, are able to regulate distinct aspects of network activity.

**Recovery of Inhibitory Correlations Requires Cotuning of Excitation and Inhibition.** So far, we have focused on the recovery of excitatory and inhibitory firing rates through intrinsic plasticity and correlations between excitatory neurons through synaptic scaling on excitatory synapses. However, our results in Fig. 1 indicate that the other types of correlations that involve the inhibitory neurons undergo the same temporal profile during prolonged MD, with a drop during early MD and a recovery during late MD. Here, we investigated if the same homeostatic mechanisms identified above have different functionality during MD by minimally modifying our network architecture. In our network, the action of inhibition is global, where inhibitory neurons nonspecifically connect to all excitatory neurons. Hence, inhibitory neurons activate together with any excitatory assembly, resulting in weak excitatory–inhibitory and inhibitory–inhibitory correlations. Rather than considering global inhibition, we next modeled inhibition as cotuned with excitation where individual assemblies of inhibitory neurons connect exclusively to individual assemblies of excitatory neurons (*Methods*), inspired by recent experiments in the visual cortex (34). Due to this cotuning of inhibition with excitation, correlated structure emerged across all types of neurons. The excitatory–inhibitory and inhibitory–inhibitory correlations were high as the block activity generated by a given cortical excitatory assembly provides a major drive to inhibitory neurons (*SI Appendix*, Fig. S9).

Implementing the same protocol for inducing MD in our model, with depression of the feedforward weights and decorrelation of thalamocortical input, generated the same drop and recovery of firing rates and correlations, now involving both excitatory and inhibitory neurons (*SI Appendix*, Fig. S9). Notably, the same two forms of homeostatic plasticity, synaptic scaling and intrinsic homeostatic plasticity, successfully regulate distinct aspects of network activity also in these cotuned networks (*SI Appendix*, Fig. S10). Taken together, our result that homeostatic

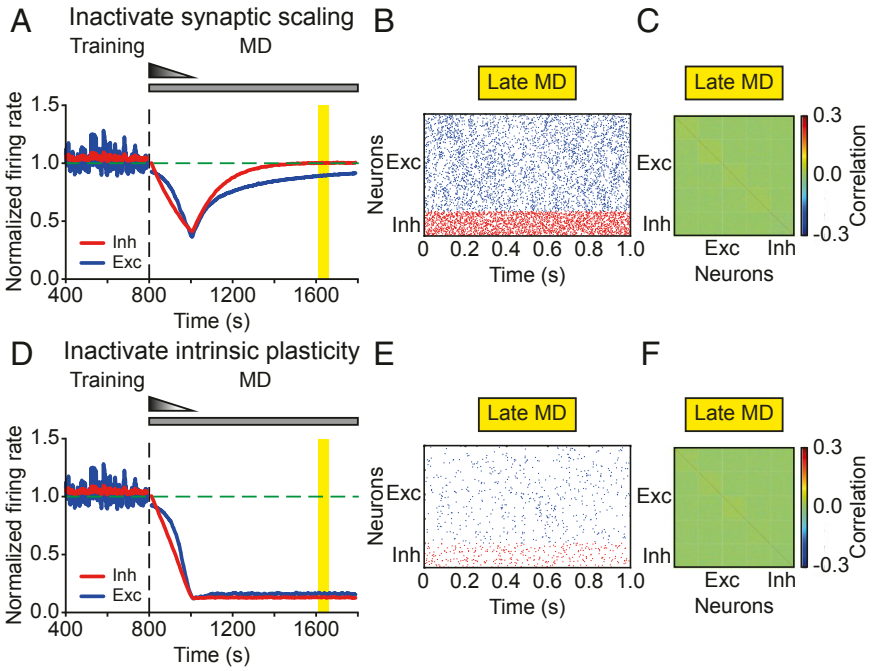


**Fig. 6.** Individual homeostatic mechanisms have different functionality during MD. (*A* and *D*) The average normalized firing rates of excitatory (blue) and inhibitory (red) neurons without synaptic scaling (*A*) or without intrinsic plasticity (*D*). The vertical dashed line indicates the onset of MD. The horizontal dashed line indicates a normalized firing rate of 1.0. (*B* and *E*) Spontaneous activity of excitatory (blue) and inhibitory (red) neurons during late MD without synaptic scaling (*B*) or without intrinsic plasticity (*E*). (*C* and *F*) Correlation matrix during late MD indicated by the yellow region in *A* and *D*, without synaptic scaling (*C*) or without intrinsic plasticity (*F*).



mechanisms regulate distinct aspects of cortical circuit dynamics applies also to different network architectures, suggesting that synaptic scaling and intrinsic plasticity quite generally influence different aspects of network function.

## Discussion

A key question in the field of homeostatic plasticity is which aspects of neuronal activity are under homeostatic control. Recent studies have shown that, despite a high degree of synaptic plasticity during the critical period (35), firing rates of individual neurons remain remarkably constant during normal development (6) and when perturbed by sensory deprivation, rebound back to an individual set point despite continued deprivation (14). Here, we used in vivo data in rodent visual cortex to investigate whether higher-order cortical network properties are under homeostatic control. We found that—distinct from firing rates—correlations in control hemispheres increased slightly during early development. In contrast, correlations in deprived hemispheres initially decreased over the first 2–3 d and then gradually recovered to predeprivation levels, including in their structure. This recovery of correlations was independent of the recovery of firing rates and had a homeostatic component beyond the developmental increase of correlations. Modeling of this process revealed that this restoration of correlation structure could be accomplished through synaptic scaling, while firing rate homeostasis was dependent on intrinsic homeostatic plasticity. Together, these findings provide evidence that functional correlation structures are subject to homeostatic regulation.

Recovery of stimulus preference at the single cell level, as well as network correlation structure, has also been reported during repeated episodes of MD in the binocular region of visual cortex, each followed by eye reopening (36). However, in these ocular dominance plasticity studies, recovery occurring following eye reopening is TrkB-dependent and mediated by Hebbian LTP (37). This is mechanistically distinct from our work where recovery is governed by homeostatic mechanisms and where there is no competition between the closed and open eye.

Our modeling results suggest that the difference in the visual input from the thalamus at MD compared to BL does not seem to be important for cortical correlations. A proper experimental verification of this result would require the measurement of correlations in the thalamus during BL and during MD. Although these data are currently unavailable, there are data to indirectly verify this. Our analysis revealed that cortical correlations in deprived hemispheres recover to their BL level after 5 to 6 d of MD (Fig. 1*B*), regardless of possible correlation changes in the thalamus. In addition, we have previously shown that correlations in the dark are approximately two-thirds of the correlations in the light when the animals are in the awake behavioral state (15). These results suggest that correlated visual inputs only have a modest impact on the amplitude of cortical correlations, while recurrent connections might be the dominant contributor. Hence, following the elimination of visual input during MD, homeostatic mechanisms such as synaptic scaling can recover cortical correlations.

What might be the purpose of the recovered network correlations? Following lid suture to induce MD, the transmitted light through the closed eye lids is relatively weaker compared to the predeprivation condition. Therefore, we propose that the network's homeostatic recovery of correlations might be a way to amplify weak signals, promoting successful signal propagation to other cortical regions (38), which is essential for the animals' perception of the sensory environment (39). We predict that the recovery of correlation structure also has important functional implications for information transmission across cortical hierarchies. For instance, neurons in layer 2/3 process inputs from neurons in layer 4 and are highly influenced by its connectivity. If the recovered network in one layer undergoes a profound remodeling and ends up having a completely different correlation structure, adjustments in successive layers would be needed to keep the cortical network functional.

We cannot conclude from our data whether neurons with higher correlations are more strongly connected. However, as previously shown, functionally correlated neurons are more likely to be connected and more strongly if so (18, 19). We therefore assume that correlation strength is indicative of connection strength. In that sense, the identified clusters with strong correlations come from strongly connected assemblies consistent with previous experimental work (18, 19, 40). However, this is only the case for excitatory neurons (identified RSUs); since the number of sorted pFS cells was significantly lower than RSUs, we could not investigate their correlation structure.

To dissect the role of various homeostatic mechanisms to restore firing rates and correlations to BL despite prolonged MD, we built and analyzed a computational model with spiking neurons and biologically realistic plasticity rules. Upon training with correlated input patterns (41), imitating the BL condition in which animals receive normal visual inputs, the network exhibited structured spontaneous activity and developed stronger correlations within assemblies. Our model showed that decreasing thalamocortical connection strength (24) and decorrelating input patterns during MD degraded synaptic weights and decreased firing rates and correlations. This was accompanied by a depression in excitatory synaptic weights within assemblies and overall inhibitory synaptic weights in the model. Although experiments have not found significant changes in the strength of recurrent excitation within layer 4 (42), in layer 2/3, there is a general depression of excitatory input (28); a more systematic analysis that includes measurements within and across assemblies would be necessary to reveal selective depression of some connections.

Other modeling studies have investigated the interaction between Hebbian and homeostatic plasticity for the stable formation and maintenance of Hebbian assemblies in the context of memory storage and recall (43–45), which are different from the sensory deprivation paradigm studied here. Interestingly, a recent modeling framework for the homeostatic recovery from visual deprivation proposed that the disinhibitory effect of inhibitory plasticity, rather than synaptic scaling, can drive the recovery of firing rates and correlations in specific subnetworks of excitatory neurons (46), based on experimental results (40). We did not observe such specificity in our data, and inhibitory plasticity in our model was insufficient to recover either firing rates or correlations, necessitating instead intrinsic plasticity and synaptic scaling.

Our modeling results indicated that attenuating the depression effect of Hebbian plasticity was required to maintain clustered network structure during the process of recovery. This suggests that the effect of Hebbian plasticity becomes attenuated during prolonged MD, which then allows homeostatic plasticity to "catch up" and restore network properties. This is consistent with several experimental findings. For example, brief MD leads to occlusion of LTD in layer 4 in the primary visual cortex (24, 47), while homeostatic strengthening of CA1 synapses in the hippocampus is accompanied by a reduced ability of synapses to exhibit LTP (48). Furthermore, during MD, the effects of Hebbian plasticity, which is originally LTD-dominant, become negligible as homeostatic plasticity reaches its steady state (31).

Importantly, in the face of ongoing plasticity, we found that two different forms of homeostatic plasticity can serve distinct functions in recovering network function. First, intrinsic plasticity as a mechanism that affects individual neuron properties, such as the firing threshold, is essential for the rebound of firing rates. Since it does not act directly on the synaptic weights, it has no significant impact on the recovery of correlations. We implemented intrinsic plasticity by adjusting the firing threshold, which effectively shifts the neuronal input–output function to keep the model sufficiently general. Biophysically, intrinsic plasticity can be implemented by changes in the density and function of voltage-gated channels (5, 49, 50).

Unlike intrinsic plasticity, synaptic scaling regulates synaptic strengths directly and is crucial for the recovery of correlation and network structure in the model. Mechanistically, this regulation is fundamentally distinct from Hebbian plasticity. The regulating process involves an enhanced accumulation of $\alpha$-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor (AMPAR) in the postsynaptic membrane, which can be mediated by the proinflammatory cytokine tumor-necrosis factor-$\alpha$ (TNF-$\alpha$) produced by glia (10), the immediate-early gene Arc (11), $\beta 3$ integrins (51), and other molecules. Crucially, the scaling is bidirectional, global, and operates in a multiplicative manner (4), although there is some evidence for dendritic branch-specific scaling in some neocortical cell types (52). During recovery, multiplicative scaling potentiates synaptic weights within assemblies more than across assemblies in our model, preserving the relative strength of synaptic inputs and enabling the recovery of correlation structure.

The distinct functional roles fulfilled by synaptic scaling and intrinsic plasticity apply in the context of the present constellation of plasticity rules. We found that synaptic scaling alone is insufficient to recover the firing rates in our model, especially inhibitory firing rates. The critical model assumption that derives this conclusion is that excitatory and inhibitory connections onto inhibitory neurons do not change during MD (*SI Appendix*, *Supplementary Text*). However, increasing synaptic strengths also boosts neuronal responses, which raises the possibility that synaptic scaling alone might be able to recover firing rates with a different combination of plasticity rules. One straightforward possibility to recover the firing rates of inhibitory neurons is either to increase the total excitation to inhibitory neurons, for example, by upscaling the excitatory-to-inhibitory connections, or to decrease the total inhibition to inhibitory neurons, for example, by downscaling the inhibitory-to-inhibitory connections. Interestingly, synaptic scaling onto inhibitory neurons was recently found to organize model recurrent networks around criticality, independently of firing rates (53). This suggests that homeostatic plasticity in excitatory elements might be important for the recovery of firing rates and correlations, while plasticity in inhibitory elements for the recovery of criticality. It still remains to be tested whether and how excitatory and inhibitory connections onto inhibitory neurons change in the context of homeostatic regulation in vivo. We highlight that including spiking neurons in our model and training the BL network with correlated inputs enabled us to study the emergence, dilution, and recovery of correlation structure during prolonged MD, which is not possible in the unstructured randomly connected networks studied in other models (53), even if firing rates recover. Furthermore, our implementation of Hebbian and homeostatic plasticity with appropriate biologically motivated timescales suggests a nontrivial cooperation between Hebbian and homeostatic plasticity, with the first being attenuated while the latter is in full operation.

In conclusion, our analysis reveals an important, previously unidentified network feature that is homeostatically regulated during perturbation of normal circuit dynamics in the visual cortex. The finding that not only the average correlations but also the correlation structure recover has interesting implications for the recovery of computations in these circuits that might be encoded in nonrandom connectivity patterns. Moreover, our network model with spiking neurons and experimentally characterized homeostatic mechanisms allowed us to dissect the role of each on different aspects of network dynamics, suggesting that different homeostatic mechanisms serve unique, rather than redundant, functions.

**Table 1. Neuron model parameters**

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $U^{\text{rest}}$ | −70 | mV | Resting membrane potential |
| $U^{\text{exc}}$ | 0 | mV | Excitatory reversal potential |
| $U^{\text{inh}}$ | −80 | mV | Inhibitory reversal potential |
| $\tau^{\text{ref}}$ | 5 | ms | Duration of refractory period |
| $\tau^{m}_{\text{exc}}$ | 20 | ms | Membrane time constant of excitatory neurons |
| $\tau^{m}_{\text{inh}}$ | 10 | ms | Membrane time constant of inhibitory neurons |
| $\tau^{\text{ampa}}$ | 5 | ms | AMPA decay time constant |
| $\tau^{\text{gaba}}$ | 10 | ms | GABA decay time constant |
| $\tau^{\text{nmda}}$ | 100 | ms | NMDA decay time constant |
| $\alpha$ | 0.5 | - | Receptor weighting factor |

-, no units.

## Methods

**Firing Rates.** To obtain the normalized firing rate evolution for different animals, the firing rates of each animal were normalized to the average firing rate at P26 during the light period. Note that, here, the analysis of firing rates was restricted to MD5 because for the higher-order network feature analysis (the pairwise correlations), the number of available, continuously recorded cells beyond this period was insufficient. Therefore, although the firing rates still seem to be above BL at MD5—a trend identical to that reported in the previous study (14)—they eventually return to BL by MD6 (14).

**Pairwise Correlations.** Each spike train was binned into spike counts of bin size 100 ms, generating a vector of spike counts for each cell. The spike-count correlation coefficient $\rho$ for a pair of neurons was computed in 30-min episodes using a sliding window of 5 min. We averaged these values for each pair every single half day (12 h), thus computing the correlation coefficient for light and dark conditions separately:

$$\rho_{X,Y} = \frac{\mathbf{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where $X$ and $Y$ represent the spike-count vectors of two cells, respectively; $\mu_X$ and $\mu_Y$ are the means of $X$ and $Y$; $\sigma_X$ and $\sigma_Y$ denote the standard deviations of $X$ and $Y$; **E** is the expectation. This produced the matrices of pairwise spike-count correlations on different half days. Just like the firing rates, to generate the normalized correlation curve across animals, the correlations of each animal were normalized to the average correlations at P26 during the light period.

The correlation matrices in Fig. 2 *A* and *B* were clustered using hierarchical clustering during BL, and the same neuron order was preserved at later time points.

**Quantification of Changes in Correlation Structure.** We first generated a shuffled matrix $\boldsymbol{A}'$ by redistributing the off-diagonal entries of the original matrix $\boldsymbol{A}$ while keeping the matrix $\boldsymbol{A}'$ symmetric. Then, we computed the absolute difference between the shuffled matrix $\boldsymbol{A}'$ and the correlation matrix at BL $\boldsymbol{B}$:

$$\boldsymbol{M} = |\boldsymbol{A}' - \boldsymbol{B}|.$$

The elements of the upper triangular part of $\boldsymbol{M}$ were used to form a vector of the absolute difference, known as the $L1$ distance, between correlations. Vectors from different animals were then concatenated into a single vector. During shuffling, only the elements corresponding to a given animal were shuffled, i.e., animal identity was preserved.

**Neuron and Network Model.** Single neurons were modeled as leaky integrate-and-fire with membrane potential of neuron $i$, $U_i$, given by (54):

$$\tau^{m}\frac{dU_i}{dt} = (U^{\text{rest}} - U_i) + g_i^{\text{ext}}(t)(U^{\text{exc}} - U_i) + g_i^{\text{inh}}(t)(U^{\text{inh}} - U_i),$$

where $\tau^m$ is the membrane time constant, and $U^{\text{rest}}$ is the resting potential. The neuron elicited a spike when its membrane potential reached the spiking threshold $U^{\text{thr}}$. After a spike, the membrane potential was reset to $U^{\text{rest}}$. The neuron also had a refractory period $\tau^{\text{ref}}$ after a spike. Inhibitory neurons also followed the same integrate-and-fire formalism but with a shorter membrane time constant. The values of all neuron model parameters are listed in Table 1.

The network model consisted of 800 excitatory and 200 inhibitory leaky integrate-and-fire neurons, which were randomly connected with a probability of 20%. Excitatory neurons were randomly grouped into four nonoverlapping groups. Each excitatory and inhibitory neuron received external excitatory input from 1,000 neurons firing with Poisson statistics at an average firing rate of 5 Hz, with synaptic strength $J^{\text{ext}\to E}$ and $J^{\text{ext}\to I}$, respectively.

Excitatory synapses have a fast AMPA component and a slow *N*-methyl-D-aspartic acid (NMDA) component. Dynamics of excitatory conductances are given by:

$$\tau^{\text{ampa}}\frac{dg_i^{\text{ampa}}}{dt} = -g_i^{\text{ampa}} + \sum_{j\in\text{exc}} J_{ij}S_j(t),$$

$$\tau^{\text{nmda}}\frac{dg_i^{\text{nmda}}}{dt} = -g_i^{\text{nmda}} + g_i^{\text{ampa}},$$

$$g_i^{\text{exc}}(t) = \alpha g_i^{\text{ampa}}(t) + (1-\alpha)g_i^{\text{nmda}}(t).$$

Here, $J_{ij}$ is the synaptic strength from neuron $j$ to neuron $i$. If the connection does not exist, $J_{ij}$ was set to 0. $S_j(t)$ is the spike train of neuron $j$, which is defined as $S_j(t) = \sum_k \delta(t - t_j^k)$, where $\delta$ is the Dirac delta function and $t_j^k$, the spikes times $k$ of neuron $j$. $\alpha$ is a weighting parameter. Dynamics of inhibitory conductances are given by:

$$\tau^{\text{gaba}}\frac{dg_i^{\text{inh}}}{dt} = -g_i^{\text{inh}} + \sum_{j\in\text{inh}} J_{ij}S_j(t).$$

The values of all network parameters are listed in Table 2.

**Training Procedure.** We implemented the network in three stages: initialization stage, a training stage, and an MD stage. All plasticity except for excitatory-to-excitatory plasticity was present in the first 100 s of the simulation to initialize the network and obtain network activity before training.

Subsequently, the training process started. During training, correlated stimuli were presented sequentially to each assembly for 1 s, with 3-s gaps in between stimulus activations. While correlated stimuli were presented to 1 assembly, the remaining neurons received inputs from 1,000 independent neurons firing with Poisson statistics at an average firing rate of 5 Hz. The firing rate of the correlated inputs was also 5 Hz. Correlated inputs for the training were generated following previous studies (33, 55). Specifically, we used a copying probability of 0.4 from individual uncorrelated Poisson source trains and a copying probability of 0.6 from a common Poisson source, all with the same firing rates.

The weight matrix obtained after training was used to induce MD in the simulations. MD simulations started with 3 s without plasticity when inhibitory spike timing-dependent plasticity (iSTDP) was activated, while other plasticity and homeostatic mechanisms were activated at 10 s. At the same time, the feedforward connections onto excitatory and inhibitory neurons linearly decreased by 8 and 15% from 10 to 210 s and, afterward, were kept fixed.

**Table 2. Network model parameters**

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $N_E$ | 800 | - | Number of excitatory neurons |
| $N_I$ | 200 | - | Number of inhibitory neurons |
| $p$ | 0.2 | - | Connectivity probability |
| $J^{EE}$ | 0.2 | - | Initial E-to-E connection weight |
| $J^{EI}$ | 2.0 | - | Initial I-to-E connection weight |
| $J^{IE}$ | 0.2 | - | E-to-I connection weight |
| $J^{II}$ | 2.0 | - | I-to-I connection weight |
| $J^{EE}_{\text{min}}$ | 0.0 | - | Minimal E-to-E connection weight |
| $J^{EE}_{\text{max}}$ | 1.2 | - | Maximal E-to-E connection weight |
| $J^{EI}_{\text{min}}$ | 0.0 | - | Minimal I-to-E connection weight |
| $J^{EI}_{\text{max}}$ | 6.0 | - | Maximal I-to-E connection weight |
| $J^{\text{ext}\to E}$ | 0.78 | - | Initial external-to-E connection weight |
| $J^{\text{ext}\to I}$ | 0.85 | - | Initial external-to-I connection weight |

E, excitatory; I, inhibitory. -, no units.

**Table 3. Plasticity model parameters**

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $r_0^E$ | 5 | Hz | Target firing rate of excitatory neurons |
| $r_0^I$ | 13 | Hz | Target firing rate of inhibitory neurons |
| $\tau^+$ | 16.8 | ms | Time constant of presynaptic detector |
| $\tau^-$ | 33.7 | ms | Time constant of faster postsynaptic detector |
| $\tau^{\text{slow}}$ | 114 | ms | Time constant of slower postsynaptic detector |
| $A^-$ | 0.0071 | - | Amplitude of LTD |
| $A^+$ | 0.0065 | - | Amplitude of LTP |
| $\tau^{\text{iSTDP}}$ | 0.02 | s | Time constant of synaptic trace for iSTDP |
| $\eta^{\text{iSTDP}}$ | 1 | - | Learning rate of iSTDP |
| $\tau^{\text{est}}$ | 20 | s | Time constant of firing rate estimator |
| $\tau^{ss}$ | 200 | s | Time constant of synaptic scaling |
| $\eta^{\text{ip}}$ | 0.00125 | mV/s | Learning rate of intrinsic plasticity |

-, no units.

**Plasticity.** To form the clustered correlation structure observed experimentally, we followed previous modeling studies (25) and modeled the plasticity of excitatory-to-excitatory synapses using triplet STDP (32) of inhibitory-to-excitatory synapses using iSTDP (56, 57) and also included heterosynaptic plasticity operating on excitatory-to-excitatory synapses.

The triplet STDP rule describes synaptic plasticity based on triplets of spikes and captures experiments where the rate of pre- and postsynaptic neurons varies (58). The triplet STDP rule enables the formation of bidirectional connections, a necessity for the formation of clustered architectures (41, 59). According to this rule, the synaptic strength from excitatory neuron $j$ to excitatory neuron $i$ follows:

$$\frac{dJ_{ij}^{EE}}{dt} = -z_i^-(t)A^- S_j(t) + z_j^+(t)A^+ z_i^{\text{slow}}(t-\epsilon)S_i(t).$$

Here, $A^-$ and $A^+$ are the amplitude of the weight change induced by a post–pre pair or a post–pre–post triplet of spikes. $\epsilon$ is a small positive constant. The synaptic traces for neuron $i$ (and similarly for neuron $j$) $z_i^+(t)$, $z_i^-(t)$, and $z_i^{\text{slow}}(t)$ evolve according to $\frac{dz_i^n}{dt} = -\frac{z_i^n}{\tau^n} + S_i(t)$ with different time constants $\tau^n$, where $n = \{+, -, \text{slow}\}$.

According to iSTDP, the synaptic strength from inhibitory neuron $j$ to excitatory neuron $i$ follows:

$$\frac{dJ_{ij}^{EI}}{dt} = \eta^{\text{iSTDP}}(x_i - 2r_i^0\tau^{\text{iSTDP}})S_j(t) + \eta^{\text{iSTDP}}x_j S_i(t),$$

where $x_i$ and $x_j$ are the synaptic traces of the postsynaptic excitatory neuron $i$ and presynaptic inhibitory neuron $j$, which are described by $\frac{dx_i}{dt} = -\frac{x_i}{\tau^{\text{iSTDP}}} + S_i(t)$, with $r_i^0$, $\tau^{\text{iSTDP}}$, and $\eta$ denoting the target firing rate of neuron $i$ (and similarly for neuron $j$), the time constant of the synaptic trace and the learning rate of iSTDP, respectively.

Excitatory-to-inhibitory connections and inhibitory-to-inhibitory connections were nonplastic since their plasticity has been much less investigated experimentally and computationally. All plastic weights were subject to upper bounds.

***Heterosynaptic plasticity.*** We also modeled normalization in the form of heterosynaptic plasticity, which ensures that the sum of all incoming excitatory synaptic weights at each postsynaptic excitatory neuron is kept below a target (60). This form of normalization has been found to be essential in maintaining clustered structures upon their formation (25). Hence, the synaptic strength from excitatory neuron $j$ to excitatory neuron $i$ was modified according to heterosynaptic plasticity as follows:

$$J_{ij}^{EE}(t) \leftarrow J_{ij}^{EE}(t) - \frac{1}{N_i^E}\left(\sum_j J_{ij}^{EE}(t) - \beta\sum_j J_{ij}^{EE}(0)\right),$$

where $N_i^E$ is the number of nonzero elements. As heterosynaptic plasticity also imposed a constraint on the excitatory-to-excitatory synaptic weight, $\beta$ was set to 1.08 so that $J_{ij}^{EE}$ becomes approximately $J_{max}^{EE}$. Heterosynaptic plasticity was implemented every 1 s and only acting when the $\sum_j J_{ij}^{EE}(t)$ was larger than $\beta\sum_j J_{ij}^{EE}(0)$.

***Metaplasticity.*** The amplitude of LTD for neuron $i$, $A_i^-$, follows:

$$A_i^- \leftarrow A_i^- \frac{x_i^{\text{est}}}{\tau^{\text{est}} r_i^0}.$$

Here, $x_i^{\text{est}}$ denotes the firing-rate estimator defined as $\frac{dx_i^{\text{est}}}{dt} = -\frac{x_i^{\text{est}}}{\tau^{\text{est}}} + S_i(t)$, with $\tau^{\text{est}}$ being the integration time constant of $x_i^{\text{est}}$. If the firing rate of a neuron was close to its target, $r_i^0$, then $\frac{x_i^{\text{est}}}{\tau^{\text{est}} r_i^0} \approx 1$. Metaplasticity was implemented every 30 s. Furthermore, $A_i^-$ was bounded below by 15% of its initial value to ensure that the effect of Hebbian plasticity eventually becomes negligible, as shown previously (31).

***Homeostatic mechanisms: synaptic scaling and intrinsic plasticity.*** The evolution of synaptic strength from excitatory neuron $j$ to excitatory neuron $i$ via synaptic scaling is given by:

$$\tau^{ss}\frac{dJ_{ij}^{EE}}{dt} = J_{ij}^{EE}\left(1 - \frac{x_i^{\text{est}}}{\tau^{\text{est}} r_i^0}\right),$$

where $\tau^{ss}$ represents the time constant of synaptic scaling.

The firing threshold of neuron $i$ regulated by intrinsic plasticity is given by:

$$\frac{dU_i^{\text{thr}}}{dt} = \eta^{ip}\left(\frac{x_i^{\text{est}}}{\tau^{\text{est}}} - r_i^0\right),$$

where $\eta^{ip}$ is the learning rate of intrinsic plasticity. Initial firing threshold was set to $-50$ mV.

The values of all plasticity parameters are listed in Table 3.

**Cotuned Network.** The cotuned network model consisted of 800 excitatory and 200 inhibitory neurons. Excitatory and inhibitory neurons were divided into four nonoverlapping groups. The connectivity probability within the same groups is 20%. Inhibitory neurons exclusively connected with excitatory neurons in the same group. The simulations started with 3 s without plasticity when iSTDP was activated, while other plasticity and homeostatic mechanisms were inactivated for the first 210 s. After that, other plasticity and homeostatic mechanisms were activated. The feedforward connections onto excitatory and inhibitory neurons linearly decreased by 4 and 8% from 210 to 410 s. From 410 s onward, feedforward connections were kept fixed. For the sake of simplicity, we implemented metaplasticity differently from the original model. Instead of dynamically modifying the LTD amplitude, here, we disabled Hebbian plasticity at 410 s. Parameters used in cotuned network models, which are different from the original model, are listed in *SI Appendix*, Table S1.

**Simulations.** Data analysis and numerical simulations were performed in Python and Julia. All differential equations were implemented by Euler integration with a time step of 0.1 ms.

**Data Availability.** The code used for data analysis and model simulations is available at GitHub (https://github.com/comp-neural-circuits/homeostasis). The data is available at Figshare (https://figshare.com/projects/Homeostasis/80936).

**ACKNOWLEDGMENTS.** We thank all members of J.G.'s group for comments and discussions. This work was supported by the Max Planck Society (Y.K.W. and J.G.), NIH Grants R01 EY025613 and R35 NS111562 (to G.G.T.), and NIH Grant R00 NS089800 (to K.B.H.). This project has received funding from the European Research Council under European Union's Horizon 2020 Research and Innovation Program Grant 804824 (to J.G.).

1. G. B. Smith, A. J. Heynen, M. F. Bear, Bidirectional synaptic mechanisms of ocular dominance plasticity in visual cortex. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 357–367 (2008).
2. L. F. Abbott, S. B. Nelson, Synaptic plasticity: Taming the beast. *Nat. Neurosci.* **3** 1178–1183 (2000).
3. G. G. Turrigiano, S. B. Nelson, Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.* **5**, 97–107 (2004).
4. G. G. Turrigiano, K. R. Leslie, N. S. Desai, L. C. Rutherford, S. B. Nelson, Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* **391**, 892–896 (1998).
5. N. S. Desai, L. C. Rutherford, G. G. Turrigiano, Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.* **2**, 515–520 (1999).
6. K. B. Hengen, M. E. Lambo, SD. Van Hooser, D. B. Katz, G. G. Turrigiano, Firing rate homeostasis in visual cortex of freely behaving rodents. *Neuron* **80**, 335–342 (2013).

7. T. Keck *et al.*, Synaptic scaling and homeostatic plasticity in the mouse visual cortex in vivo. *Neuron* **80**, 327–334 (2013).
8. D. P. Seeburg, M. Feliu-Mojer, J. Gaiottino, D. T. Pak, M. Sheng, Critical role of cdk5 and polo-like kinase 2 in homeostatic synaptic plasticity during elevated activity. *Neuron* **58**, 571–583 (2008).
9. D. M. Evers *et al.*, Plk2 attachment to NSF induces homeostatic removal of GluA2 during chronic overexcitation. *Nat. Neurosci.* **13**, 1199–1207 (2010).
10. D. Stellwagen, R. C. Malenka, Synaptic scaling mediated by glial TNF-$\alpha$. *Nature* **440**, 1054–1059 (2006).
11. J. D. Shepherd *et al.*, Arc/arg3.1 mediates homeostatic synaptic scaling of AMPA receptors. *Neuron* **52**, 475–484 (2006).
12. A. Joseph, G. G. Turrigiano, All for one but not one for all: Excitatory synaptic scaling and intrinsic excitability are coregulated by CaMKIV, whereas inhibitory synaptic scaling is under independent control. *J. Neurosci.* **37**, 6778–6785 (2017).
13. G. Daoudal, D. Debanne, Long-term plasticity of intrinsic excitability: Learning rules and mechanisms. *Learn. Mem.* **10**, 456–465 (2003).
14. K. B. Hengen, A. T. Pacheco, J. N. McGregor, S. D. Van Hooser, G. G. Turrigiano, Neuronal firing rate homeostasis is inhibited by sleep and promoted by wake. *Cell* **165**, 180–191 (2016).
15. A. T. Pacheco *et al.*, Rapid and active stabilization of visual cortical firing rates across light–dark transitions. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 18068–18077 (2019).
16. J. De La Rocha, B. Doiron, E. Shea-Brown, K. Josić, A. Reyes, Correlation between neural spike trains increases with firing rate. *Nature* **448**, 802–806 (2007).
17. M. L. Linden, A. J. Heynen, R. H. Haslinger, M. F. Bear, Thalamic activity that drives visual cortical plasticity. *Nat. Neurosci.* **12**, 390–392 (2009).
18. H. Ko *et al.*, Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**, 87–91 (2011).
19. L. Cossell *et al.*, Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518**, 399–403 (2015).
20. R. Perin, T. K. Berger, H. Markram, A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5419–5424 (2011).
21. Y. Yoshimura, J. L. Dantzker, E. M. Callaway, Excitatory cortical neurons form fine-scale functional networks. *Nature* **433**, 868–873 (2005).
22. S. J. Cruikshank, T. J. Lewis, B. W. Connors, Synaptic basis for intense thalamocortical activation of feedforward inhibitory cells in neocortex. *Nat. Neurosci.* **10**, 462–468 (2007).
23. X.-Y. Ji *et al.*, Thalamocortical innervation pattern in mouse auditory and visual cortex: Laminar and cell-type specificity. *Cereb. Cortex* **26**, 2612–2625 (2015).
24. N. J. Miska, L. M. Richter, B. A. Cary, J. Gjorgjieva, G. G. Turrigiano, Sensory experience inversely regulates feedforward and feedback excitation-inhibition ratio in rodent visual cortex. *eLife* **7**, e38846 (2018).
25. A. Litwin-Kumar, B. Doiron, Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nat. Commun.* **5**, 5319 (2014).
26. M. Y. Frenkel, M. F. Bear, How monocular deprivation shifts ocular dominance in visual cortex of young mice. *Neuron* **44**, 917–923 (2004).
27. G. Turrigiano, Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annu. Rev. Neurosci.* **34**, 89–103 (2011).
28. M. E. Lambo, G. G. Turrigiano, Synaptic and intrinsic homeostatic mechanisms cooperate to increase L2/3 pyramidal neuron excitability during a late phase of critical period plasticity. *J. Neurosci.* **33**, 8810–8819 (2013).
29. M. S. Grubb, J. Burrone, Activity-dependent relocation of the axon initial segment fine-tunes neuronal excitability. *Nature* **465**, 1070–1074 (2010).
30. E. Campanac *et al.*, Enhanced intrinsic excitability in basket cells maintains excitatory-inhibitory balance in hippocampal circuits. *Neuron* **77**, 712–722 (2013).
31. T. Toyoizumi, M. Kaneko, M. P. Stryker, K. D. Miller, Modeling the dynamic interaction of Hebbian and homeostatic plasticity. *Neuron* **84**, 497–510 (2014).
32. J. P. Pfister, W. Gerstner, Triplets of spikes in a model of spike timing-dependent plasticity. *J. Neurosci.* **26**, 9673–9682 (2006).
33. J. Gjorgjieva, C. Clopath, J. Audet, J. P. Pfister, A triplet spike-timing-dependent plasticity model generalizes the Bienenstock-Cooper-Munro rule to higher-order spatiotemporal correlations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19383–19388 (2011).
34. P. Znamenskiy *et al.*, Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. bioRxiv DOI: http://dx.doi.org/10.1101/294835 (4 April 2018).
35. C. N. Levelt, M. Hübener, Critical-period plasticity in the visual cortex. *Annu. Rev. Neurosci.* **35**, 309–330 (2012).
36. T. Rose, J. Jaepel, M. Hübener, T. Bonhoeffer, Cell-specific restoration of stimulus preference after monocular deprivation in the visual cortex. *Science* **352**, 1319–1322 (2016).
37. M. Kaneko, J. Hanover, P. England, M. Stryker, TrkB kinase is required for recovery, but not loss, of cortical responses following monocular deprivation. *Nat. Neurosci.* **11**, 497–504 (2008).
38. T. P. Vogels, L. F. Abbott, Signal propagation and logic gating in networks of integrate-and-fire neurons. *J. Neurosci.* **25**, 10786–10795 (2005).
39. B. Van Vugt *et al.*, The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360**, 537–542 (2018).
40. S. J. Barnes *et al.*, Subnetwork-specific homeostatic plasticity in mouse visual cortex in vivo. *Neuron* **86**, 1290–1303 (2015).
41. G. K. Ocker, B. Doiron, Training and spontaneous reinforcement of neuronal assemblies by spike timing plasticity. *Cerebr. Cortex* **29**, 937–951 (2018).
42. A. Maffei, K. Nataraj, S. B. Nelson, G. G. Turrigiano, Potentiation of cortical inhibition by visual deprivation. *Nature* **7**, 81–84 (2006).
43. J. M. Auth, T. Nachstedt, C. Tetzlaff, The interplay of synaptic plasticity and scaling enables self-organized formation and allocation of multiple memory representations. bioRxiv DOI: http://dx.doi.org/10.1101/260950 (14 October 2018).
44. J. Humble, K. Hiratsuka, H. Kasai, T. Toyoizumi, Intrinsic spine dynamics are critical for recurrent network learning in models with and without autism spectrum disorder. *Front. Comput. Neurosci.* **13**, 38 (2019).
45. F. Zenke, E. J. Agnes, W. Gerstner, Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nat. Commun.* **6**, 6922 (2015).
46. Y. Sweeney, S. J. Barnes, C. Clopath, Diverse homeostatic responses to visual deprivation by uncovering recurrent subnetworks. bioRxiv DOI: http://doi.org/10.1101/312926 (2 May 2018).
47. R. A. Crozier, Y. Wang, C. H. Liu, M. F. Bear, Deprivation-induced synaptic depression by distinct mechanisms in different layers of mouse visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1383–1388 (2007).
48. C. Soares, K. F. Lee, J. C. Béïque, Metaplasticity at CA1 synapses by homeostatic control of presynaptic release dynamics. *Cell Rep.* **21**, 1293–1303 (2017).
49. G. LeMasson, E. Marder, L. Abbott, Activity-dependent regulation of conductances in model neurons. *Science* **259**, 1915–1917 (1993).
50. G. Turrigiano, G. LeMasson, E. Marder, Selective regulation of current densities underlies spontaneous changes in the activity of cultured neurons. *J. Neurosci.* **15**, 3640–3652 (1995).
51. L. A. Cingolani,*et al.*, Activity-dependent regulation of synaptic AMPA receptor composition and abundance by $\beta$3 integrins. *Neuron* **58**, 749–762 (2008).
52. S. J. Barnes *et al.*, Deprivation-induced homeostatic spine scaling in vivo is localized to dendritic branches that have undergone recent spine loss. *Neuron* **96**, 871–882 (2017).
53. Z. Ma, G. G. Turrigiano, R. Wessel, K. B. Hengen, Cortical circuit dynamics are homeostatically tuned to criticality in vivo. *Neuron* **104**, 655–664 (2019).
54. F. Zenke, G. Hennequin, W. Gerstner, Synaptic plasticity in neural networks needs homeostasis with a fast rate detector. *PLoS Comput. Biol.* **9**, e1003330 (2013).
55. R. Brette, Generation of correlated spike trains. *Neural Comput.* **21**, 188–215 (2009).
56. T. P. Vogels, H. Sprekeler, F. Zenke, C. Clopath, W. Gerstner, Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569–1573 (2011).
57. J. A. D'amour, R. C. Froemke, Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* **86**, 514–528 (2015).
58. P. J. Sjöström, G. G. Turrigiano, S. B. Nelson, Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* **32**, 1149–1164 (2001).
59. L. Montangie, C. Miehl, J. Gjorgjieva, Autonomous emergence of connectivity assemblies via spike triplet interactions. *PLoS Comput. Biol.* **16**, e1007835 (2020).
60. I. R. Fiete, W. Senn, C. Z. Wang, R. H. Hahnloser, Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron* **65**, 563–576 (2010).

# II. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity

Wu, Y. K. & Gjorgjieva, J. Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity. *Physical Review Research* **5**, 033023 (2023). https://doi.org/10.1103/PhysRevResearch.5.033023

# Inhibition stabilization and paradoxical effects in recurrent neural networks with short-term plasticity

Yue Kris Wu * and Julijana Gjorgjieva
*School of Life Sciences, Technical University of Munich, 85354 Freising, Germany*
*and Max Planck Institute for Brain Research, 60438 Frankfurt, Germany*

(Received 22 December 2022; accepted 5 June 2023; published 12 July 2023)

*This article is part of the Physical Review Research collection titled Physics of Neuroscience.*

Inhibition stabilization is considered a ubiquitous property of cortical networks, whereby inhibition controls network activity in the presence of strong recurrent excitation. In networks with fixed connectivity, an identifying characteristic of inhibition stabilization is that increasing (decreasing) excitatory input to the inhibitory population leads to a decrease (increase) in inhibitory firing, known as the paradoxical effect. However, population responses to stimulation are highly nonlinear, and drastic changes in synaptic strengths induced by short-term plasticity (STP) can occur on the timescale of perception. How neuronal nonlinearities and STP affect inhibition stabilization and the paradoxical effect is unclear. Using analytical calculations, we demonstrate that in networks with STP the paradoxical effect implies inhibition stabilization, but inhibition stabilization does not imply the paradoxical effect. Interestingly, networks with neuronal nonlinearities and STP can transition nonmonotonically between inhibition-stabilization and noninhibition-stabilization, and between paradoxically- and nonparadoxically-responding regimes with increasing excitatory activity. Furthermore, we generalize our results to more complex scenarios including networks with multiple interneuron subtypes and any monotonically increasing neuronal nonlinearities. In summary, our work reveals the relationship between inhibition stabilization and the paradoxical effect in the presence of neuronal nonlinearity and STP, yielding several testable predictions.

DOI: 10.1103/PhysRevResearch.5.033023

## I. INTRODUCTION

Cortical networks are typically characterized by inhibition stabilization, where inhibition is needed to keep network activity levels in biologically realistic ranges despite the presence of strong recurrent excitation [1]. Networks operating in the inhibition-stabilized regime are capable of performing various computations, including input amplification, response normalization, and network multistability [2–6]. In networks with fixed connectivity, a hallmark of inhibition stabilization is the paradoxical effect: An increase or a decrease of excitatory input to the inhibitory population respectively decreases or increases the inhibitory firing [7]. Over the past decade, much effort has been made to identify the operating regime of cortical networks based on the paradoxical effect [1,8,9].

Yet, various aspects ranging from the network to the synaptic level can considerably affect network dynamics and the operating regime. First, if individual neurons in the network receive large excitatory and inhibitory currents which precisely cancel each other, the network operates in a balanced state characterized by a linear population response [10–12]. Recent work has argued that neuronal input-output functions are better characterized by supralinear functions, and networks with this type of nonlinearity can exhibit various nonlinear phenomena as observed in biology [13–15]. Second, synapses in the brain are highly dynamic as a result of different STP mechanisms, operating on a timescale of milliseconds to seconds [16,17]. Upon presynaptic stimulation, postsynaptic responses can either get depressed subject to short-term depression (STD) or facilitated subject to short-term facilitation (STF). While short-term synaptic dynamics are widely observed in biological circuits, it is unclear how they interact with the neuronal nonlinearity to jointly determine the network operating regime. Here we ask how the neuronal nonlinearity and STP affect inhibition stabilization and the paradoxical effect.

To address this question, we determine the conditions for inhibition stabilization and the paradoxical effect in networks of excitatory and inhibitory neurons in the presence of STP with linear and supralinear population response functions. We find that, irrespective of the neuronal nonlinearity, in networks with excitatory-to-excitatory (E-to-E) STD, inhibition stabilization does not necessarily imply the paradoxical effect, but the paradoxical effect implies inhibition stabilization. In contrast, in networks with static connectivity or networks with other STP mechanisms, inhibition stabilization and the paradoxical effect imply each other. Interestingly, neuronal nonlinearities and STP endow the network with unconventional behaviors. More specifically, in the presence

*kris.wu@tum.de

of a neuronal nonlinearity and E-to-E STD, monotonically increasing excitatory activity can lead to nonmonotonic transitions between noninhibition-stabilization and inhibition-stabilization, as well as between nonparadoxically-responding and paradoxically-responding regimes. Furthermore, we generalize our results to more complex scenarios including networks with multiple interneuron subtypes and any monotonically increasing neuronal nonlinearities. In conclusion, our work reveals the impact of neuronal nonlinearities and STP on inhibition stabilization and the paradoxical effect, and makes several predictions for future experiments.

## II. RESULTS

To understand the relationship between inhibition stabilization and the paradoxical effect in recurrent neural networks with STP, we studied rate-based population models consisting of an excitatory (E) and an inhibitory (I) population. The dynamics of the network are given by

$$\tau_E \frac{dr_E}{dt} = -r_E + [p_{EE}J_{EE}r_E - p_{EI}J_{EI}r_I + g_E]_+^{\alpha_E}, \quad (1)$$

$$\tau_I \frac{dr_I}{dt} = -r_I + [p_{IE}J_{IE}r_E - p_{II}J_{II}r_I + g_I]_+^{\alpha_I}, \quad (2)$$

where $r_E$ and $r_I$ denote the firing rates of the excitatory and inhibitory population; $\tau_E$ and $\tau_I$ are the corresponding time constants; $J_{AB}$ represents the synaptic strength from population $B$ to population $A$, where $A, B \in \{E, I\}$; $g_E$ and $g_I$ are the external inputs to the respective populations; and $\alpha_E$ and $\alpha_I$ are the exponents of the respective input-output functions. Finally, $p_{AB}$ represents the short-term plasticity variable from population $B$ to population $A$. We implemented short-term plasticity mechanisms based on the Tsodyks and Markram model [16]. For STD, we replaced $p_{AB}$ with $x_{AB}$ and described the STD dynamics as follows:

$$\frac{dx_{AB}}{dt} = \frac{1 - x_{AB}}{\tau_x} - U_d x_{AB} r_B, \quad (3)$$

where $x_{AB}$ is a short-term depression variable that is limited to the interval (0,1] for the synaptic connection from population $B$ to population $A$. Biophysically, the short-term depression variable $x$ represents the fraction of vesicles available for release, $\tau_x$ is the time constant of STD, and $U_d$ is the depression factor controlling the degree of depression induced by the presynaptic activity.

For STF, we replaced $p_{AB}$ by $u_{AB}$ and expressed the STF dynamics as follows:

$$\frac{du_{AB}}{dt} = \frac{1 - u_{AB}}{\tau_u} + U_f(U_{\max} - u_{AB}) r_B, \quad (4)$$

where $u_{AB}$ is a short-term facilitation variable that is constrained to the interval $[1, U_{\max})$ for the synaptic connection from population $B$ to population $A$. Biophysically, the short-term facilitation variable $u$ represents the ability of releasing neurotransmitter, $\tau_u$ 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.

To investigate the impact of neuronal nonlinearities on inhibition stabilization and the paradoxical effect, we considered both threshold-linear networks ($\alpha_E = \alpha_I = 1$) as well as supralinear networks ($\alpha_E = \alpha_I > 1$). In the regime of positive $r_E$ and $r_I$, threshold-linear networks behave as linear networks. In the following, we thus call them linear networks. Furthermore, while we keep our analysis for supralinear networks in a general form, we use $\alpha_E = \alpha_I = 2$ for the numerical simulations. Note that the neuronal nonlinearity in our study refers to the nonlinearity of population-averaged responses to input when no STP mechanisms are taken into account, which is fully determined by $\alpha_E$ and $\alpha_I$.

In addition, for the sake of analytical tractability, we included one STP mechanism at a time. To investigate how inhibition stabilization is affected by the neuronal nonlinearity and STP, we computed the real part of the leading eigenvalue of the Jacobian matrix of the excitatory-to-excitatory subnetwork incorporating STP, and refer to it as the "Inhibition Stabilization index" (IS index) (see Supplemental Material [18]). A positive (negative) IS index implies that the network is in the IS (non-IS) regime. To reveal how inhibition stabilization changes with network activity and network connectivity, we investigated how the IS index changes with the excitatory activity $r_E$ and the excitatory to excitatory connection strength $J_{EE}$. These two quantities, $r_E$ and $J_{EE}$, are directly involved in the definition of the IS index (see Supplemental Material [18]).

### A. Inhibition stabilization in recurrent neural networks with short-term depression at E-to-E synapses

We first examined inhibition stabilization for networks with E-to-E STD, evaluated at the fixed point of the system [Fig. 1(a)]. The distinction between non-IS and IS is reflected in network responses to perturbations induced by injecting additional excitatory currents into excitatory neurons while inhibition is fixed. Networks initially in the non-IS regime return back to their initial activity level after a small transient perturbation to the excitatory activity when inhibition is fixed, whereas networks initially in the IS regime deviate from their initial activity (Fig. S1). For linear networks with E-to-E STD, if $J_{EE}$ is less than one, the network is always in the non-IS regime regardless of $r_E$ [Fig. 1(b)]. If $J_{EE}$ is greater than one, the network transitions from IS to non-IS with increasing $r_E$ [Fig. 1(b)]. In contrast, supralinear networks with E-to-E STD manifest different behaviors. When $J_{EE}$ is large, the network first transitions from non-IS to IS, and then back to non-IS with increasing $r_E$ [Figs. 1(c) and S1]. When $J_{EE}$ is small, the supralinear network stays in the non-IS regime for all values of $r_E$ [Fig. 1(c)].

To better understand the transition between non-IS and IS in the presence of neuronal nonlinearities and E-to-E STD, we investigated how the boundary between non-IS and IS, defined as "IS boundary," changes with $r_E$ (see Fig. S2; Supplemental Material [18]). Mathematically, the IS boundary is determined by the recurrent excitatory-to-excitatory connection strength for different $r_E$ at which the IS index is zero, denoted by $J_{EE}^{IS}$. By computing the derivative of $J_{EE}^{IS}$ with respect to $r_E$ (see Supplemental Material [18]), we found that the derivative is always positive for linear networks with E-to-E STD, suggesting that the IS boundary increases with increasing $r_E$ [Fig. 1(d)]. Therefore, for networks with large

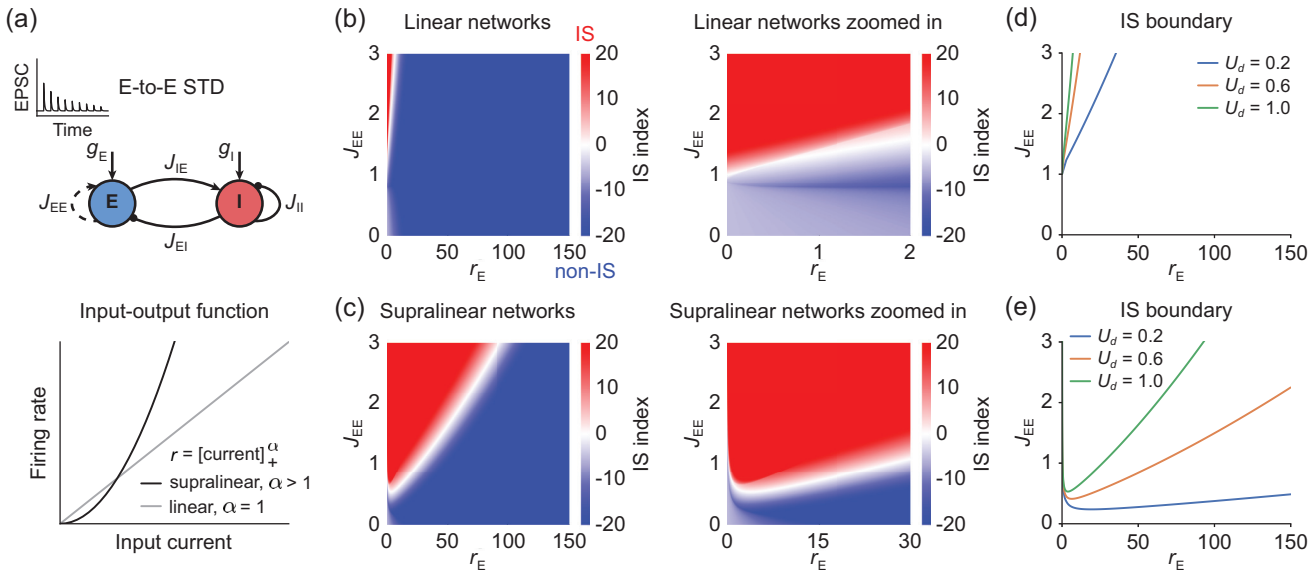


FIG. 1. Inhibition stabilization in recurrent neural networks with E-to-E short-term depression. (a) Top: Schematic of the recurrent network model consisting of an excitatory (blue) and an inhibitory population (red) with E-to-E STD. Bottom: Different nonlinearities of population-averaged responses to input. Linear (gray) and supralinear (black) input-output function given by a rectified power law with exponent $\alpha = 1$ and $\alpha = 2$ [cf. Eqs. (1)and (2)], respectively. (b) Left: IS index as a function of excitatory firing rate $r_E$ and excitatory-to-excitatory connection strength $J_{EE}$ for linear networks with E-to-E STD. The non-IS and IS regime are marked in blue and red, respectively. Right: Zoomed-in version of (b) left. Here, the depression factor $U_d$ is 1.0, the firing rate $r_E$ is plotted in a biologically realistic range from 0 to 150 Hz. (c) Same as (b) but for supralinear networks with E-to-E STD. Here, the depression factor $U_d$ is 1.0. (d) IS boundary for linear networks with E-to-E STD, defined as the corresponding excitatory-to-excitatory connection strength $J_{EE}^{IS}$ for different $r_E$ at which the IS index is zero. Different colors represent the IS boundary for different values of depression factor $U_d$. (e) Same as (d) but for supralinear networks with E-to-E STD. Here, $\tau_x$ is 200 ms in (b)–(e).

$J_{EE}$, as $r_E$ increases, only the transition from IS to non-IS is possible [Figs. 1(b) and 1(d)]. In contrast, for supralinear networks with E-to-E STD, the derivative changes from negative to positive with increasing $r_E$ (see Supplemental Material [18]), implying that the IS boundary first decreases and then increases as $r_E$ increases [Fig. 1(e)]. As a result, networks can undergo nonmonotonic transitions between non-IS and IS with increasing $r_E$. More specifically, networks can switch from non-IS to IS, and then back to non-IS with increasing $r_E$ [Figs. 1(c) and 1(e); Fig. S1]. The nonmonotonic transitions arise from the competition between the increasing neuronal gain due to the supralinear neuronal input-output function and the decreasing synaptic strength due to STD. Despite high firing rates in the large $r_E$ limit, E-to-E synaptic strengths are significantly depressed and STD effectively decouples excitatory neurons, rendering the network non-inhibition stabilized. Furthermore, the turning point of the IS boundary appears when $U_d x_{EE} r_E$ is of order one (see Supplemental Material [18]). Increasing the depression factor $U_d$ or the STD time constant $\tau_x$ shifts the turning point to the upper left in the $(r_E, J_{EE})$ coordinate system (see Fig. S3; Supplemental Material [18]). We also found that these nonmonotonic transitions cannot be observed in networks with static connectivity (see Fig. S4; Supplemental Material [18]). Taken together, our results suggest that E-to-E STD can nontrivially affect the inhibition stabilization property, especially in the presence of neuronal nonlinearities.

### B. Inhibition stabilization in recurrent neural networks with short-term facilitation at E-to-E synapses

To determine if the observed effects are specific to the type of STP at E-to-E synapses, we next examined networks with E-to-E STF [Fig. 2(a)]. Unlike the scenario with STD, for both linear networks or supralinear networks, only a monotonic transition from non-IS to IS is possible with increasing $r_E$ in the presence of E-to-E STF [Figs. 2(b) and 2(c)]. In contrast to supralinear networks, linear networks with $J_{EE}$ larger than one are always in the IS regime regardless of $r_E$. In both cases, the parameter regime of $J_{EE}$ and $r_E$ which supports IS is much larger than in the corresponding network with STD (Fig. 2). Furthermore, independent of the neuronal nonlinearity, the derivative of $J_{EE}^{IS}$ with respect to $r_E$ is always negative (see Supplemental Material [18]), indicating that the IS boundary decreases as $r_E$ increases [Figs. 2(d) and 2(e)]. These results indicate that E-to-E STF exerts a more intuitive influence on the inhibition stabilization property than E-to-E STD even in the presence of neuronal nonlinearities.

### C. Inhibition stabilization in recurrent neural networks with short-term plasticity at other synapses

Finally, we performed the same analyses for networks with different types of STP at all synapses other than E-to-E, including E-to-I STD/STF, I-to-E STD/STF, and I-to-I STD/STF, respectively [see Fig. 3(a); Supplemental

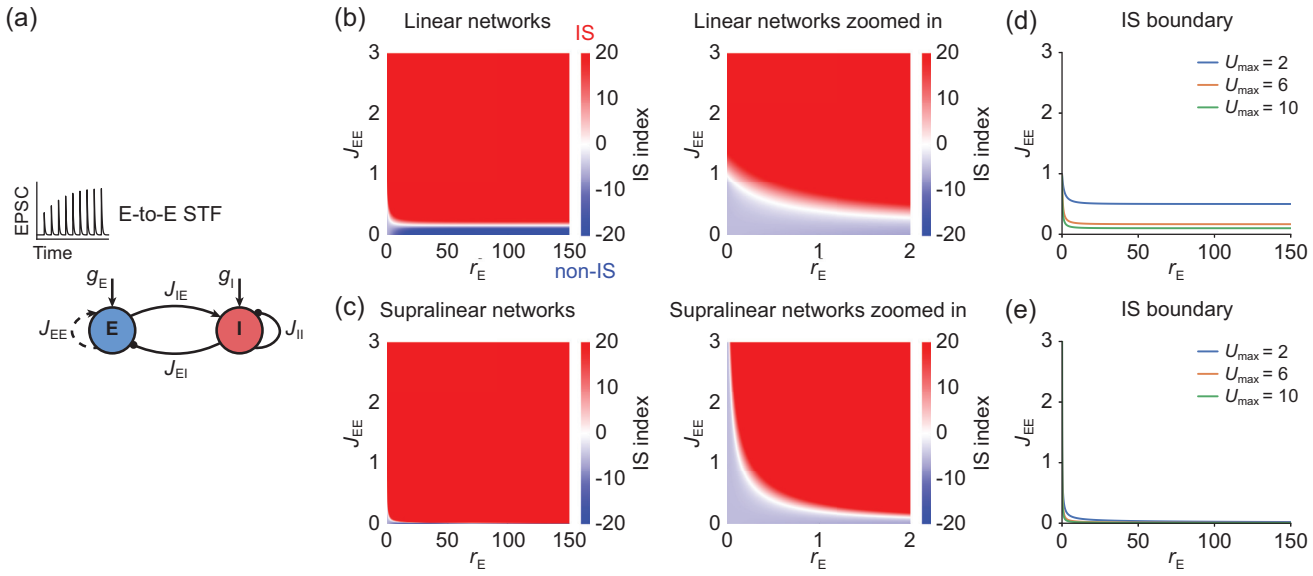


FIG. 2. Inhibition stabilization in recurrent neural networks with E-to-E short-term facilitation. (a) Schematic of the recurrent network model consisting of an excitatory (blue) and an inhibitory population (red) with E-to-E STF. (b) Left: IS index as a function of excitatory firing rate $r_E$ and excitatory-to-excitatory connection strength $J_{EE}$ for linear networks with E-to-E STF. The non-IS and IS regime are marked in blue and red, respectively. Right: Zoomed-in version of (b) left. Here, the facilitation factor $U_f$ is 1.0, and the maximal facilitation value $U_{\max}$ is 6.0. (c) Same as (b) but for supralinear networks with E-to-E STF. Here, the facilitation factor $U_f$ is 1.0, and $U_{\max}$ is 6.0. (d) IS boundary for linear networks with E-to-E STF, defined as the corresponding excitatory-to-excitatory connection strength $J_{EE}^{IS}$ for different $r_E$ at which the IS index is zero. Different colors represent the IS boundary for different values of $U_{\max}$. (e) Same as (d) but for supralinear networks with E-to-E STF. Here, $\tau_u$ is 200 ms in (b)–(e).

Material [18]]. Including these STP mechanisms does not change the IS condition relative to networks with fixed connectivity (see Supplemental Material [18]). For networks with a linear input-output function, the IS boundary does not change with $r_E$ [Figs. 3(b) and 3(d); Supplemental Material [18]], and $J_{EE}$ completely determines whether the network is non-IS or IS. In contrast, for networks with a supralinear input-output function, the derivative of $J_{EE}^{IS}$ with respect to $r_E$ is always negative, suggesting that the IS boundary decreases with increasing $r_E$ [see Figs. 3(c) and 3(e); Supplemental Material [18]]. Therefore, the transition from non-IS to IS with increasing $r_E$ in static supralinear networks or supralinear networks with STP at all synapses other than E-to-E can only happen for large $J_{EE}$ [Figs. 3(c) and 3(e); Fig. S4]. No transition between non-IS and IS can occur in the biological realistic firing regime from 0 to 150 Hz for small $J_{EE}$ [Figs. 3(c) and 3(e)].

In summary, by considering all possible STP mechanisms, our results demonstrate a nontrivial influence of the neuronal nonlinearity and STP on inhibition stabilization. Specifically, we revealed how inhibition stabilization changes with excitatory activity and network connectivity when considering neuronal nonlinearities and STP.

### D. Paradoxical effects in recurrent neural networks with short-term plasticity

Previous theoretical studies have suggested that in excitatory and inhibitory networks with static connectivity, one identifying characteristic of inhibition stabilization is that injecting excitatory (inhibitory) current into inhibitory neurons decreases (increases) inhibitory firing rates, known as the paradoxical effect [4,7]. Here, we sought to identify the conditions under which a paradoxical effect can arise in recurrent neural networks with STP. We assumed that the system is stable locally around the fixed point, in other words, a small transient perturbation to the system will not lead to deviation from the initial fixed point over time. Furthermore, the perturbation used to probe the paradoxical effect (e.g., the excitatory current injected to the inhibitory population) is small enough that it will not lead to a transition between non-IS and IS. To determine the conditions for the presence of the paradoxical effect under these assumptions, we considered the phase plane of the excitatory (abscissa) and inhibitory (ordinate) firing rate dynamics. The first condition involves a larger slope of the inhibitory nullcline than of the excitatory nullcline locally around the fixed point in the phase plane, while the second condition involves a positive slope of the excitatory nullcline around the fixed point (see Ref. [7]; Fig. S5; Supplemental Material [18]). We compared the above two conditions for the presence of the paradoxical effect with the conditions to be in the IS regime. We found that irrespective of the shape of the neuronal nonlinearity, in networks with E-to-E STD, the paradoxical effect implies inhibition stabilization, whereas inhibition stabilization does not imply the paradoxical effect [see Fig. 4(a); Supplemental Material [18]]. In contrast, for networks with E-to-E STF, E-to-I STD/STF, I-to-I STD/STF, and I-to-I STD/STF, inhibition stabilization and the paradoxical effect imply each other [see Fig. 4(a); Supplemental Material [18]]. To highlight the difference between inhibition stabilization and the paradoxical effect in networks with E-to-E STD [Fig. 4(b)], we plotted the paradoxical effect

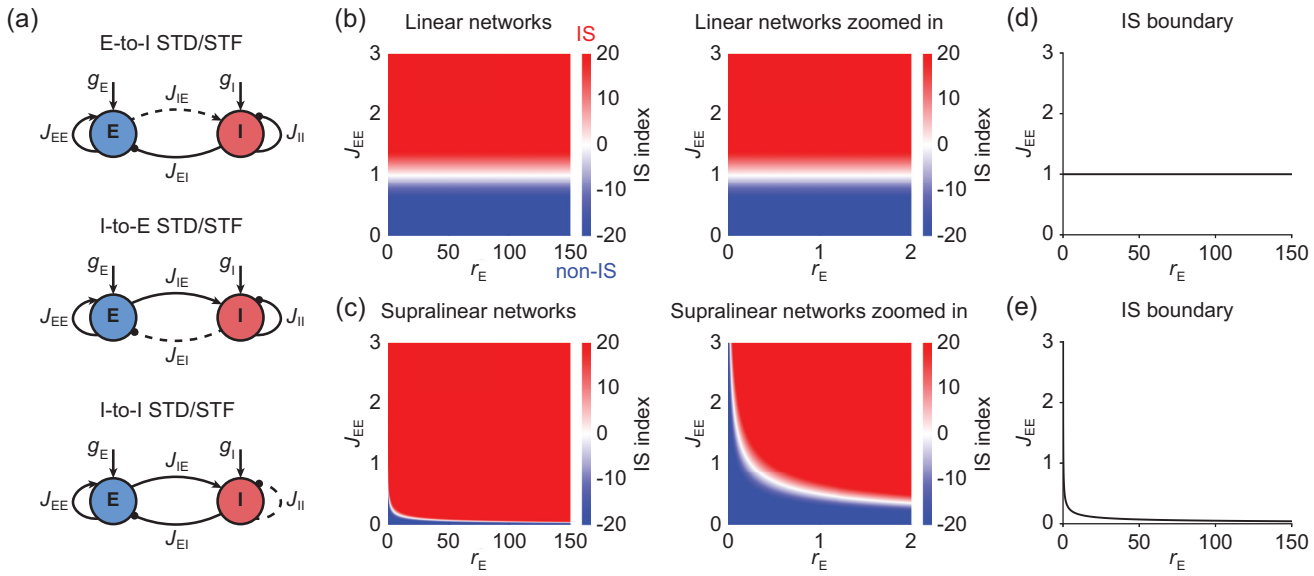


FIG. 3. Inhibition stabilization in recurrent neural networks with E-to-I short-term depression/facilitation, I-to-E short-term depression/facilitation, or I-to-I short-term depression/facilitation. (a) Schematic of the recurrent network model consisting of an excitatory (blue) and an inhibitory population (red) with E-to-I STD/STF (top), with I-to-E STD/STF (middle), or with I-to-I STD/STF (bottom). (b) Left: IS index as a function of excitatory firing rate $r_E$ and excitatory-to-excitatory connection strength $J_{EE}$ for linear networks with E-to-I STD/STF, with I-to-E STD/STF, or with I-to-I STD/STF. The non-IS and IS regime are marked in blue and red, respectively. Right: Zoomed-in version of (b) left. (c) Same as (b) but for supralinear networks with E-to-I STD/STF, with I-to-E STD/STF, or with I-to-I STD/STF. (d) IS boundary for linear networks with E-to-I STD/STF, with I-to-E STD/STF, or with I-to-I STD/STF, defined as the corresponding excitatory-to-excitatory connection strength $J_{EE}^{IS}$ for different $r_E$ at which the IS index is zero. (e) Same as (d) but for supralinear networks with E-to-I STD/STF, with I-to-E STD/STF, or with I-to-I STD/STF.

boundary that separates the paradoxically-responding and the nonparadoxically-responding regime together with the IS boundary for both linear networks and supralinear networks with E-to-E STD [Figs. 4(c)–4(f)]. In the two dimensional $r_E$-$J_{EE}$ plane, the parameter regime for the paradoxical effect is much narrower than the IS regime, suggesting that there is a large parameter space, in which inhibition-stabilized networks do not exhibit the paradoxical effect [Figs. 4(c)–4(f)]. It is noteworthy that in the presence of E-to-E STD, networks in which dynamic STD regulation is required to ensure stability, as studied in [19,20], are a subset of networks which do not exhibit paradoxical effects (see Supplemental Material [18]). For inhibition stabilized networks which do not exhibit paradoxical effects, the corresponding excitatory subnetwork with dynamical short-term plasticity variables is unstable.

Furthermore, by analyzing how the paradoxical effect boundary changes with $r_E$, we found that it exhibits a similar trend to the IS boundary (see Supplemental Material [18]). In particular, the paradoxical effect boundary of supralinear networks with E-to-E STD is also a nonmonotonic function of $r_E$. Therefore, in this case, networks can undergo nonmonotonic transitions between the paradoxically-responding regime and nonparadoxically-responding regime with monotonically changing excitatory activity $r_E$.

### E. Generalization to networks with a mixture of STP mechanisms and multiple interneuron subtypes

So far, we only considered individual STP mechanisms one at a time. Here, we generalized our findings to four more complex and biologically realistic scenarios. First, our results can be generalized to networks with a mixture of STP mechanisms at other types of synapses except E-to-E synapses where we assume only STD or STF (see Supplemental Material [18]). More specifically, the conditions and results derived for networks with either E-to-E STD or STF alone are the same for networks with either E-to-E STD or STF and a mixture of STP mechanisms at other types of synapses.

Second, by incorporating both STD and STF at E-to-E synapses, we found that the paradoxical effect implies inhibition stabilization, whereas inhibition stabilization does not necessarily imply the paradoxical effect (see Supplemental Material [18]).

Third, inhibitory neurons in the cortex are highly diverse in terms of anatomy, electrophysiology, and function [21–23]. Recent studies have investigated the relationship between inhibition stabilization and the paradoxical effect in networks with multiple interneuron subtypes in the absence of STP [24–28]. Yet, synapses between different interneuron subtypes exhibit considerable short-term dynamics [29]. We then extended the analysis to networks with multiple interneuron subtypes. Theoretical studies have shown that in networks with static connectivity, if the excitatory subnetwork is non-IS (IS), then the sign of the change in the firing rate of the excitatory population and of the change in the total inhibitory current to the excitatory population are opposite (the same) [24]. We found that in the presence of E-to-E STD, if the network is IS, the sign of the change in the firing rate of the excitatory population and of the change in the total inhibitory current to the excitatory population are the same (see Supplemental Material [18]). However, the same sign

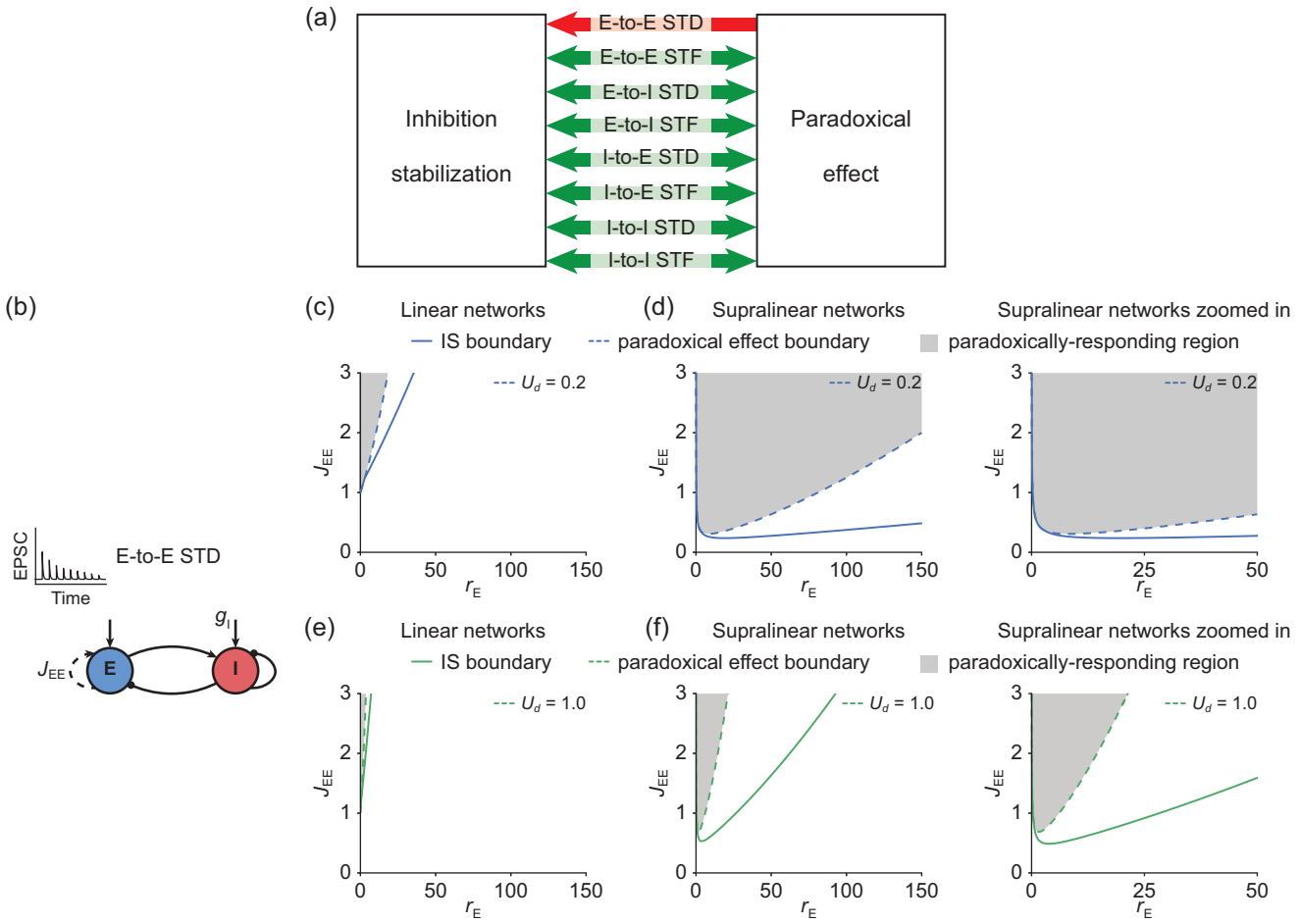


FIG. 4. Relationship between inhibition stabilization and the paradoxical effect in recurrent neural networks with short-term plasticity. (a) In networks with E-to-E STD, the paradoxical effect implies inhibition stabilization, whereas inhibition stabilization does not imply the paradoxical effect indicated by the unidirectional red arrow. In contrast, for networks with E-to-E STF, E-to-I STD/STF, I-to-I STD/STF, and I-to-I STD/STF, inhibition stabilization and the paradoxical effect imply each other indicated by the bidirectional green arrows. (b) Schematic of the recurrent network model consisting of an excitatory (blue) and an inhibitory population (red) with E-to-E STD. (c) An example of the paradoxical effect boundary (dashed line) and the inhibition stabilization boundary (solid line) as a function of excitatory firing rate $r_E$ for linear networks with E-to-E STD. The paradoxically-responding region is marked in gray. Here, the depression factor $U_d$ is 0.2 (Supplemental Methods). (d) Same as (c) but for supralinear networks with E-to-E STD. Right: Zoomed-in version of (d) left. (e), (f) Same as (c), (d) but the depression factor $U_d$ 1.0. Here, $\tau_x$ is 200 ms in (c)–(f).

of the change in the firing rate of the excitatory population and of the change in the total inhibitory current to the excitatory population does not imply that the network is IS (see Supplemental Material [18]). In the presence of E-to-E STF and STP at other types of synapses, inhibition stabilization and the same sign change in the firing rate of the excitatory population and of the change in the total inhibitory current to the excitatory population imply each other (see Supplemental Material [18]).

Finally, the derived relationship between inhibition stabilization and the paradoxical effect is independent from the power of the power-law input-output function. Therefore, the derived relationship holds for any monotonically increasing neuronal nonlinearities including sublinear (see Fig. S6; Supplemental Material [18]), despite the fact that sublinear networks have different IS transitions from supralinear networks (see Fig. S7; Supplemental Material [18]). Taken together, our results indicate that the relationship between inhibition stabilization and paradoxical effect in networks with STP becomes nontrivial in the presence of short-term plasticity.

## III. DISCUSSION

In this study, we combined analytical and numerical methods to reveal the effects of neuronal nonlinearities and STP on inhibition stabilization, the paradoxical effect, and the relationship between them. Including STD at E-to-E synapses, in contrast to other types of STP and other synapse types, generates the most surprising results. Under these conditions, the paradoxical effect implies inhibition stabilization, whereas inhibition stabilization does not imply the paradoxical effect. For networks with other STP mechanisms and networks with static connectivity, inhibition stabilization and the paradoxical effect imply each other. Additionally, in the presence of a neu-

ronal nonlinearity and E-to-E STD, a nonmonotonic transition between different regimes can occur when excitatory activity changes monotonically.

Network models applied to neural circuit development have previously investigated inhibition stabilization in the presence of STP [30]. Recent studies have examined inhibition stabilization and the paradoxical effect in threshold-linear networks with E-to-E STP to demonstrate that inhibition stabilization can be probed by the paradoxical effect [1]. Recent work has conducted similar analysis for supralinear networks with short-term plasticity on specific synapses [31]. Here, we systematically analyzed networks with all forms of STP mechanisms, for both linear networks and supralinear networks. By mathematically defining the IS boundary and the paradoxical effect boundary, we further demonstrated how network activity and connectivity affect the inhibition stabilization property and the paradoxical effect. Importantly, we generalized our results to several more complex scenarios including networks with a mixture of STP mechanisms, networks with both E-to-E STD and STF, networks with multiple interneuron subtypes, and any monotonically increasing neuronal nonlinearities.

In this work, we assumed that the network activity has reached a fixed point, and we did not consider scenarios like multistability or oscillations that could arise from neuronal nonlinearities or STP [15,32,33]. While multistability and oscillations have been observed in the brain [34,35], the single stable fixed point assumed in our study is believed to be a reasonable approximation of awake sensory cortex [36].

Our model makes several predictions that can be tested in optogenetic experiments. Across cortical layers and across brain regions, synaptic strengths can differ by an order of magnitude [37]. Furthermore, the degrees of balance between excitation and inhibition may also vary [38,39], resulting in different neuronal nonlinearities [11,13,39]. Therefore, different behaviors predicted by our analysis may be observable in different neural circuits. For example, in the presence of E-to-E STD, our model shows that networks with weak excitatory to excitatory connection strength $J_{EE}$ are always non-IS in the biologically realistic activity regime and therefore exhibit no paradoxical effects. In contrast, with E-to-E STD and strong $J_{EE}$, network models with a linear population-averaged response function can undergo the transition from IS to non-IS with increasing excitatory activity $r_E$. Different from linear networks, our model predicts that nonmonotonic transitions between non-IS and IS can be found in supralinear networks. More specifically, supralinear networks can switch from non-IS to IS, and then from IS to non-IS with increasing $r_E$. Although inhibition stabilization does not imply the paradoxical effect in the presence of E-to-E STD, the transition between paradoxically-responding and nonparadoxically-responding regime is also nonmonotonic with increasing $r_E$ in supralinear networks, whereas in linear networks, only transitions from the paradoxically-responding to the nonparadoxically-responding regime with increasing $r_E$ is possible. Therefore, depending on the excitation-inhibition balance, and the specific short term plasticity mechanisms operating in different brain regions, our work proposes that the circuits will exhibit different properties when interrogated with common experimental techniques.

Second, in the presence of STF on E-to-E synapses or STP on other synapses, our results demonstrate that inhibition stabilization and the paradoxical effect imply each other. Linear network models with $J_{EE}$ larger than one that have E-to-E STF are always IS and thus exhibit the paradoxical effect. In linear network models with STP on other synapses, activity does not affect inhibition stabilization and the paradoxical effect. In contrast, regardless of the strength of $J_{EE}$, supralinear networks with E-to-E STF or STP on other synapses can switch from non-IS to IS with increasing $r_E$. This transition from non-IS and IS can be directly tested experimentally by probing the paradoxical effect, because of the equivalence of inhibition stabilization and the paradoxical effect found in network models with E-to-E STF or STP on other synapses.

Last, our analysis shows that in most cases substantially altering either $J_{EE}$ or $r_E$ can switch the network operating regime. Multiple factors can modify $J_{EE}$ and $r_E$ experimentally. On a short timescale, the strength of sensory stimulation, and behavioral states like arousal [40], attention [41], and locomotion [42] can dramatically change activity levels $r_E$. Regime switching may therefore be experimentally observable across different stimulation conditions and different behavioral states. On a long timescale, $J_{EE}$ or $r_E$ can be modified by long term plasticity mechanisms [43,44]. In this case, regime switching could be experimentally detectable across different developmental stages.

Taken together, our theoretical framework provides a systematic analysis of how short-term synaptic plasticity and response nonlinearities interact to determine the network operating regime, revealing unexpected relationships and their signatures as a guide for future experimental studies.

The code used for model simulations is available at GitHub [45]. All simulation parameters are listed in Supplemental Material.

## ACKNOWLEDGMENTS

We thank Elizabeth Herbert, Felix Waitzmann, Leonidas M. A. Richter, and Dylan Festa for comments on the manuscript. This work was supported by the Max Planck Society and has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 804824 to J.G.) and from the Deutsche Forschungsgemeinschaft in the Collaborative Research Centre 1080 (project C7 to J.G.). Y.K.W. is supported by the Add-on Fellowship of the Joachim Herz Foundation.

[1] A. Sanzeni, B. Akitake, H. C. Goldbach, C. E. Leedy, N. Brunel, and M. H. Histed, Inhibition stabilization is a widespread property of cortical networks, eLife **9**, 54875 (2020).

[2] D. J. Amit and N. Brunel, Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex, Cereb. Cortex **7**, 237 (1997).

[3] B. K. Murphy and K. D. Miller, Balanced amplification: A new mechanism of selective amplification of neural activity patterns, Neuron **61**, 635 (2009).

[4] H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster, Inhibitory stabilization of the cortical network underlies visual surround suppression, Neuron **62**, 578 (2009).

[5] D. B. Rubin, S. D. VanHooser, and K. D. Miller, The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex, Neuron **85**, 402 (2015).

[6] S. Sadeh and C. Clopath, Inhibitory stabilization and cortical computation, Nat. Rev. Neurosci. **22**, 21 (2021).

[7] M. V. Tsodyks, W. E. Skaggs, T. J. Sejnowski, and B. L. McNaughton, Paradoxical effects of external modulation of inhibitory interneurons, J. Neurosci. **17**, 4382 (1997).

[8] S. Sadeh, R. A. Silver, T. D. Mrsic-Flogel, and D. R. Muir, Assessing the role of inhibition in stabilizing neocortical networks requires large-scale perturbation of the inhibitory population, J. Neurosci. **37**, 12050 (2017).

[9] N. Li, S. Chen, Z. V. Guo, H. Chen, Y. Huo, H. K. Inagaki, G. Chen, C. Davis, D. Hansel, C. Guo, and K. Svoboda, Spatiotemporal constraints on optogenetic inactivation in cortical circuits, eLife **8**, 48622 (2019).

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

[11] C. van Vreeswijk and H. Sompolinsky, Chaotic balanced state in a model of cortical circuits, Neural Comput. **10**, 1321 (1998).

[12] A. Renart, J. D. Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris, The asynchronous state in cortical circuits, Science **327**, 587 (2010).

[13] Y. Ahmadian, D. B. Rubin, and K. D. Miller, Analysis of the stabilized supralinear network, Neural Comput. **25**, 1994 (2013).

[14] G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, and K. D. Miller, The dynamical regime of sensory cortex: Stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability, Neuron **98**, 846 (2018).

[15] N. Kraynyukova and T. Tchumatchenko, Stabilized supralinear network can give rise to bistable, oscillatory, and persistent activity, Proc. Natl. Acad. Sci. USA **115**, 3464 (2018).

[16] M. V. Tsodyks and H. Markram, The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability, Proc. Natl. Acad. Sci. USA **94**, 719 (1997).

[17] H. Markram, E. Muller, S. Ramaswamy, M. W. Reimann, M. Abdellah, C. A. Sanchez, A. Ailamaki, L. Alonso-Nanclares, N. Antille, S. Arsever *et al.*, Reconstruction and simulation of neocortical microcircuitry, Cell **163**, 456 (2015).

[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.5.033023 for detailed mathematical derivations.

[19] A. Loebel and M. Tsodyks, Computation by ensemble synchronization in recurrent networks with synaptic depression, J. Comput. Neurosci **13**, 111 (2002).

[20] A. Loebel, I. Nelken, and M. Tsodyks, Processing of sounds by population spikes in a model of primary auditory cortex, Front. Neurosci. **1**, 197 (2007).

[21] C. K. Pfeffer, M. Xue, M. He, Z. J. Huang, and M. Scanziani, Inhibition of inhibition in visual cortex: The logic of connections between molecularly distinct interneurons, Nat. Neurosci. **16**, 1068 (2013).

[22] X. Jiang, S. Shen, C. R. Cadwell, P. Berens, F. Sinz, A. S. Ecker, S. Patel, and A. S. Tolias, Principles of connectivity among morphologically defined cell types in adult neocortex, Science **350**, aac9462 (2015).

[23] R. Tremblay, S. Lee, and B. Rudy, GABAergic interneurons in the neocortex: From cellular properties to circuits, Neuron **91**, 260 (2016).

[24] A. Litwin-Kumar, R. Rosenbaum, and B. Doiron, Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes, J. Neurophysiol. **115**, 1399 (2016).

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

[26] K. D. Miller and A. Palmigiano, Generalized paradoxical effects in excitatory/inhibitory networks, bioRxiv, https://www.biorxiv.org/content/10.1101/2020.10.13.336727v1.

[27] A. Palmigiano, F. Fumarola, D. P. Mossing, N. Kraynyukova, H. Adesnik, and K. D. Miller, Common rules underlying optogenetic and behavioral modulation of responses in multi-cell-type V1 circuits, bioRxiv, https://www.biorxiv.org/content/10.1101/2020.11.11.378729v3.

[28] L. M. A. Richter and J. Gjorgjieva, A circuit mechanism for independent modulation of excitatory and inhibitory firing rates after sensory deprivation, Proc. Nat. Acad. Sci. USA **119**, e2116895119 (2022).

[29] L. Campagnola, S. C. Seeman, T. Chartrand, L. Kim, A. Hoggarth, C. Gamlin, S. Ito, J. Trinh, P. Davoudian, C. Radaelli *et al.*, Local connectivity and synaptic dynamics in mouse and human neocortex, Science **375**, eabj5861 (2022).

[30] V. Rahmati, K. Kirmse, K. Holthoff, L. Schwabe, and S. J. Kiebel, Developmental emergence of sparse coding: A dynamic systems approach, Sci. Rep. **7**, 13015 (2017).

[31] Y. K. Wu and F. Zenke, Nonlinear transient amplification in recurrent neural networks with short-term plasticity, eLife **10**, 71263 (2021).

[32] G. Mongillo, O. Barak, and M. Tsodyks, Synaptic theory of working memory, Science **319**, 1543 (2008).

[33] G. Mongillo, D. Hansel, and C. van Vreeswijk, Bistability and Spatiotemporal Irregularity in Neuronal Networks with Nonlinear Synaptic Transmission, Phys. Rev. Lett. **108**, 158101 (2012).

[34] X.-J. Wang, Synaptic reverberation underlying mnemonic persistent activity, Trends Neurosci. **24**, 455 (2001).

[35] G. Buzsáki and A. Draguhn, Neuronal oscillations in cortical networks, Science **304**, 1926 (2004).

[36] K. D. Miller, Canonical computations of cerebral cortex, Current Opinion in Neurobiology **37**, 75 (2016).

[37] Allen Institute for Brain Science, Synaptic physiology coarse matrix dataset, 2019.

[38] J. Barral and A. D'Reyes, Synaptic scaling rule preserves excitatory-inhibitory balance and salient neuronal network dynamics, Nat. Neurosci. **19**, 1690 (2016).

[39] Y. Ahmadian and K. D. Miller, What is the dynamical regime of cerebral cortex? Neuron **109**, 3373 (2021).

[40] M. Vinck, R. Batista-Brito, U. Knoblich, and J. A. Cardin, Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding, Neuron **86**, 740 (2015).

[41] J. H. Reynolds and L. Chelazzi, Attentional modulation of visual processing, Annu. Rev. Neurosci. **27**, 611 (2004).

[42] Y. Fu, J. M. Tucciarone, J. S. Espinosa, N. Sheng, D. P. Darcy, R. A. Nicoll, Z. J. Huang, and M. P. Stryker, A cortical circuit for gain control by behavioral state, Cell **156**, 1139 (2014).

[43] N. L. Rochefort, O. Garaschuk, R. I. Milos, M. Narushima, N. Marandi, B. Pichler, Y. Kovalchuk, and A. Konnerth, Sparsification of neuronal activity in the visual cortex at eye-opening, Proc. Natl. Acad. Sci. USA **106**, 15049 (2009).

[44] H. Ko, L. Cossell, C. Baragli, J. Antolik, C. Clopath, S. B. Hofer, and T. D. Mrsic-Flogel, The emergence of functional microcircuits in visual cortex, Nature (London) **496**, 96 (2013).

[45] Code repository, https://github.com/comp-neural-circuits/inhibition-stabilization-paradoxical-effect-STP.

# III. Regulation of circuit organization and function through inhibitory synaptic plasticity

Wu, Y. K.*, Miehl, C.* & Gjorgjieva, J. Regulation of circuit organization and function through inhibitory synaptic plasticity. *Trends in Neurosciences* **45**, 884–898 (2022). https://doi.org/10.1016/j.tins.2022.10.006

CellPress
OPEN ACCESS

Trends in
Neurosciences

Review

# Regulation of circuit organization and function through inhibitory synaptic plasticity

Yue Kris Wu [1,2,3], Christoph Miehl [1,2,3], and Julijana Gjorgjieva [1,2,*]

**Diverse inhibitory neurons in the mammalian brain shape circuit connectivity and dynamics through mechanisms of synaptic plasticity. Inhibitory plasticity can establish excitation/inhibition (E/I) balance, control neuronal firing, and affect local calcium concentration, hence regulating neuronal activity at the network, single neuron, and dendritic level. Computational models can synthesize multiple experimental results and provide insight into how inhibitory plasticity controls circuit dynamics and sculpts connectivity by identifying phenomenological learning rules amenable to mathematical analysis. We highlight recent studies on the role of inhibitory plasticity in modulating excitatory plasticity, forming structured networks underlying memory formation and recall, and implementing adaptive phenomena and novelty detection. We conclude with experimental and modeling progress on the role of interneuron-specific plasticity in circuit computation and context-dependent learning.**

## Highlights

Inhibitory synapses are continuously modified by experience through synaptic plasticity. Different learning rules have been proposed to describe the dependence of plasticity on firing rates, spike timing, calcium levels, and membrane potential.

Inhibitory plasticity affects dendritic, cellular, and network dynamics and influences excitatory plasticity at all levels.

Inhibitory plasticity shapes the formation of feedforward receptive fields and structured connectivity in recurrent circuits, supporting the formation and recall of memories and the generation of adaptive and novelty responses. of memories and the generation of adaptive and novelty responses.

Multiple inhibitory neuron subtypes and interneuron-specific plasticity support various computations, including context-dependent processing and pathway-specific selection, and play unique roles in supporting the stability and competition of neural assemblies.

## Inhibition throughout development and adulthood

Long-term synaptic plasticity is widely considered to underlie circuit assembly and connectivity refinement during early postnatal development, as well as learning and memory in adulthood [1]. Over the past few decades, extensive studies have characterized the plasticity of synapses between excitatory neurons [2–5]. Consistent with Hebbian principles, coincident pre- and postsynaptic activity potentiates synaptic strength, which enhances the correlation between pre- and postsynaptic activity and further potentiates synaptic strength, potentially leading to runaway synaptic growth and abnormal seizure-like activity [6]. To prevent excessive excitation and maintain stable activity levels, neural circuits employ various mechanisms to dynamically coordinate changes in excitation and inhibition [7,8]. The modulation of inhibitory synapses onto excitatory neurons, called **inhibitory plasticity** (see Glossary), is one such mechanism encountered in different regions of the mammalian brain [9–14] (Box 1). Yet, understanding inhibitory plasticity and its functional implications in shaping network connectivity and dynamics remains challenging because of the different roles inhibitory plasticity might play, depending on the varying demands across an animal's lifetime, as well as the considerable anatomical, electrophysiological, and functional diversity of interneurons, which can undergo different forms of plasticity [15–17].

During early development, it has long been thought that the main inhibitory neurotransmitter in the adult, **gamma-aminobutyric acid (GABA)**, is depolarizing [18,19]. The early excitatory action of GABA has been implicated in the activity-dependent growth and differentiation of neurons and the establishment of neural circuits [20,21]. However, while GABA depolarizes immature cortical neurons *in vivo*, its action at the network level (at least in the neocortex) appears to be inhibitory [22–24]. The maturation of GABAergic synaptic transmission triggers the onset of a critical period in which sensory circuits are highly plastic and sensitive to perturbations [25]. During development and early life, the plasticity of inhibitory GABAergic synapses interacts with **excitatory plasticity** [10]. Multiple computational studies have demonstrated that this interaction shapes

[1]School of Life Sciences, Technical University of Munich, Freising, Germany
[2]Max Planck Institute for Brain Research, Frankfurt, Germany
[3]These authors contributed equally to this work.

*Correspondence:
gjorgjieva@tum.de (J. Gjorgjieva).

## Box 1. Inhibitory plasticity in experiments and models

Inhibitory plasticity has been observed in different regions of the mammalian brain [9–12,35]. Experimentally, inhibitory plasticity can be induced by concurrent presynaptic hyperpolarization and postsynaptic depolarization [16,36–39], for instance, via high-frequency stimulation of input pathways [40,41] or pairing of pre- and postsynaptic spikes [16,36,42–44] (see [13] for an extensive summary of experimental studies on inhibitory plasticity).

In computational models, inhibitory plasticity is implemented by phenomenological learning rules, which simplify the underlying complex molecular and biochemical processes [13,14]. In these models, inhibitory synaptic change can depend on firing rates, precise spike times, or membrane potential based on the induction protocol used experimentally [45–49]. A commonly used inhibitory learning rule, which depends on spikes [also called **inhibitory spike-timing-dependent plasticity** (iSTDP)], is the **symmetric Hebbian learning rule** (see Figure I in Box 1). It has a symmetric window as a function of the time difference between pre- and postsynaptic spikes. Spikes near each other in time, independent of their order, lead to inhibitory **long-term potentiation (LTP)** , whereas pre- and postsynaptic spikes far from each other lead to inhibitory **long-term depression (LTD)** [45]. A similar symmetric iSTDP window has been found experimentally in the auditory cortex [44], in the orbitofrontal cortex [50], and in the hippocampus [16]. To account for the diversity of experimentally observed iSTDP windows, computational models have also investigated other learning window shapes, including **asymmetric Hebbian**, where pre-post spike pairs lead to LTP and post-pre spike pairs lead to LTD [51,52], as observed in entorhinal cortex [43]; **asymmetric anti-Hebbian**, where pre-post spike pairs lead to LTD and post-pre spike pairs lead to LTP [52]; and **symmetric anti-Hebbian** window, where spikes near each other in time lead to LTD, while spikes far from each other lead to LTP [53], as observed in hippocampus [36] (see Figure I in Box 1).

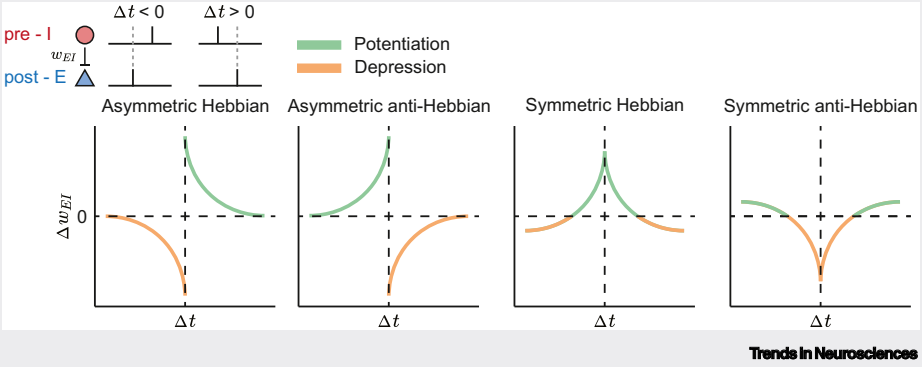


Figure I. Different learning windows of inhibitory spike-timing-dependent plasticity. Inhibitory plasticity can be parameterized into different idealized learning windows as a function of the timing difference between pre- and postsynaptic spikes $\Delta t$, leading to either inhibitory long-term potentiation ($\Delta w_{EI} > 0$, green) or inhibitory long-term depression ($\Delta w_{EI} < 0$, orange): asymmetric Hebbian [51,52], asymmetric anti-Hebbian [52], symmetric Hebbian [45], and symmetric anti-Hebbian [53].

network structures and establishes the appropriate network connectivity driven by developmental patterns of spontaneous activity and sensory experience [26–28]. Following sensory deprivation, especially during the critical period, inhibitory plasticity can regulate the balance of excitation and inhibition (E/I balance) and contribute to firing rate homeostasis [29,30]. To adapt to more complex environments, inhibitory plasticity continues to shape learning and network dynamics throughout adulthood. For example, different interneuron subtypes and interneuron-specific plasticity support diverse computations from context-dependent information processing to predictive coding [16,31–34]. Therefore, through plasticity, inhibition can adjust to the needs of the organism at various stages from development to adulthood.

Here, we present recent experimental and theoretical advances on inhibitory plasticity and the control it exerts on circuit connectivity and dynamics. We outline how inhibitory plasticity controls network firing rates and correlations, as well as the plasticity of excitatory connections. We discuss how the interaction of excitatory and inhibitory plasticity can influence the formation of

## Glossary

**Anti-Hebbian learning rule:** a learning rule in which long-term depression is induced by presynaptic followed by postsynaptic spikes, the opposite of Hebb's principle.
**Asymmetric Hebbian learning rule:** a learning rule that is an asymmetric function of the difference in spike times of pre- and postsynaptic neurons. For asymmetric learning rules, pre-post spike pairs have the opposite impact on the weight change to that of post-pre spike pairs.
**Disinhibition:** loss or reduction of inhibition. Disinhibition can be induced in multiple ways, for example, via neuromodulators that reduce GABA release from inhibitory neurons onto excitatory neurons, or via increasing inhibition onto inhibitory neurons that target excitatory neurons.
**Excitatory plasticity:** the plasticity of synapses from an excitatory to another excitatory neuron.
**Gamma-aminobutyric acid (GABA):** a major inhibitory neurotransmitter in the adult brain.
**Hebbian learning rule:** a learning rule in which long-term potentiation is induced by presynaptic followed by postsynaptic spikes, in agreement with Hebb's principle.
**Inhibition-stabilized network (ISN):** a network consisting of excitatory and inhibitory neurons with strong recurrent excitation, which is stabilized by strong feedback inhibition generated in the circuit.
**Inhibitory plasticity:** the plasticity of synapses from an inhibitory to an excitatory neuron.
**Inhibitory spike-timing-dependent plasticity (iSTDP):** a process that adjusts the (inhibitory) synaptic strength based on the timing of presynaptic and postsynaptic spikes.
**Long-term depression (LTD):** a process involving the weakening of synapses between neurons.
**Long-term potentiation (LTP):** a process involving the strengthening of synapses between neurons.
**Symmetric Hebbian learning rule:** a learning rule that is a symmetric function of the difference in spike times of pre- and postsynaptic neurons. For symmetric learning rules, pre-post spike pairs have the same impact on the weight change to that of post-pre spike pairs.

different network connectivity structures, including, but not limited to, receptive fields and assemblies, modulate these structures during learning and memory formation, and generate adapted and novelty responses. Based on experimental evidence of different interneuron subtypes and their connectivity profiles, we also present modeling studies that explore differences in the plasticity at these synapses. Throughout, a picture emerges that highlights inhibition and inhibitory plasticity as key factors that control circuit dynamics, ensure appropriate circuit function, and provide a substrate for flexible and complex computations driving behavior throughout the entire life of an organism.

## Inhibitory plasticity controls excitation at different spatiotemporal scales

To maintain stable activity levels, inhibitory plasticity can dynamically adjust the amount of inhibition at different spatial and temporal scales during both normal circuit operation and perturbation (Figure 1). At the network level, inhibition is thought to maintain healthy firing rates to prevent runaway dynamics leading to epileptic activity or decreases leading to complete silence (Figure 1A). However, in heavily interconnected neural circuits, the relationship between inhibition and network dynamics is more complicated. In such recurrently dominated networks, strong feedback inhibition generated by the circuit is needed to balance strong recurrent excitation. Both theoretical and experimental studies have put forward such inhibition stabilization as an essential property of cortical networks [54,55]. **Inhibition-stabilized networks (ISNs)** can perform various computations, including input amplification, response normalization, and network multistability [56–58]. A signature of inhibition stabilization is widely considered to be the paradoxical effect, whereby injecting excitatory currents into inhibitory neurons (e.g., via optogenetic stimulation of inhibitory neurons) decreases inhibitory firing [59]. Several circuit aspects, including recurrent excitatory-to-excitatory connection strengths and network activity, can dynamically shape inhibition stabilization [57,60]. For example, in networks where neuronal dynamics are nonlinear, changing the connection from inhibitory to excitatory neurons affects network activity and puts the network in different inhibition-stabilized regimes, as evaluated by the presence of the paradoxical effect (Figure 1B, [57,58,60]). Yet, detecting ISNs via the paradoxical effect is experimentally challenging due to the sensitivity of optogenetic stimulation strength [61] and the complexity introduced by multiple interneuron subtypes [62]. While inhibition stabilization is necessary for various computations, it is still unclear how it can be maintained in the presence of synaptic plasticity, for example, during learning, though recent work addresses this question in the context of balanced excitatory and inhibitory receptive field formation [63].

More broadly, inhibitory plasticity can operate as a homeostatic process and control network activity following perturbation [64,65]. A classical paradigm to explore this process experimentally is elevating or suppressing the activity of cultured neurons, which triggers the potentiation or depression of spontaneous inhibitory synaptic currents into the perturbed neurons [66,67]. In the living animal, a perturbation may involve sensory deprivation, for example, the removal of whiskers in the somatosensory system or the closure of an eye in the visual system [68,69]. Here, inhibitory plasticity could be involved both during the initial circuit response leading to the decrease in network firing rates, as well as later on during their recovery. Initially, the strong potentiation of recurrent inhibition onto excitatory neurons could contribute to the early decrease of network firing rates [30,70,71]. The subsequent gradual upregulation of firing rates could be triggered by the loss of inhibitory synapses onto excitatory neurons [72,73], or the decreased spontaneous inhibitory current frequency [74,75] and amplitude [64,68]. In sum, inhibitory plasticity could act as a common driver behind the homeostatic regulation of network activity immediately after or during a prolonged period following sensory perturbation across sensory cortices.

How could inhibitory plasticity achieve this homeostatic regulation of excitatory firing rates? One answer lies in the concept of E/I balance, which inhibitory plasticity can establish and maintain at

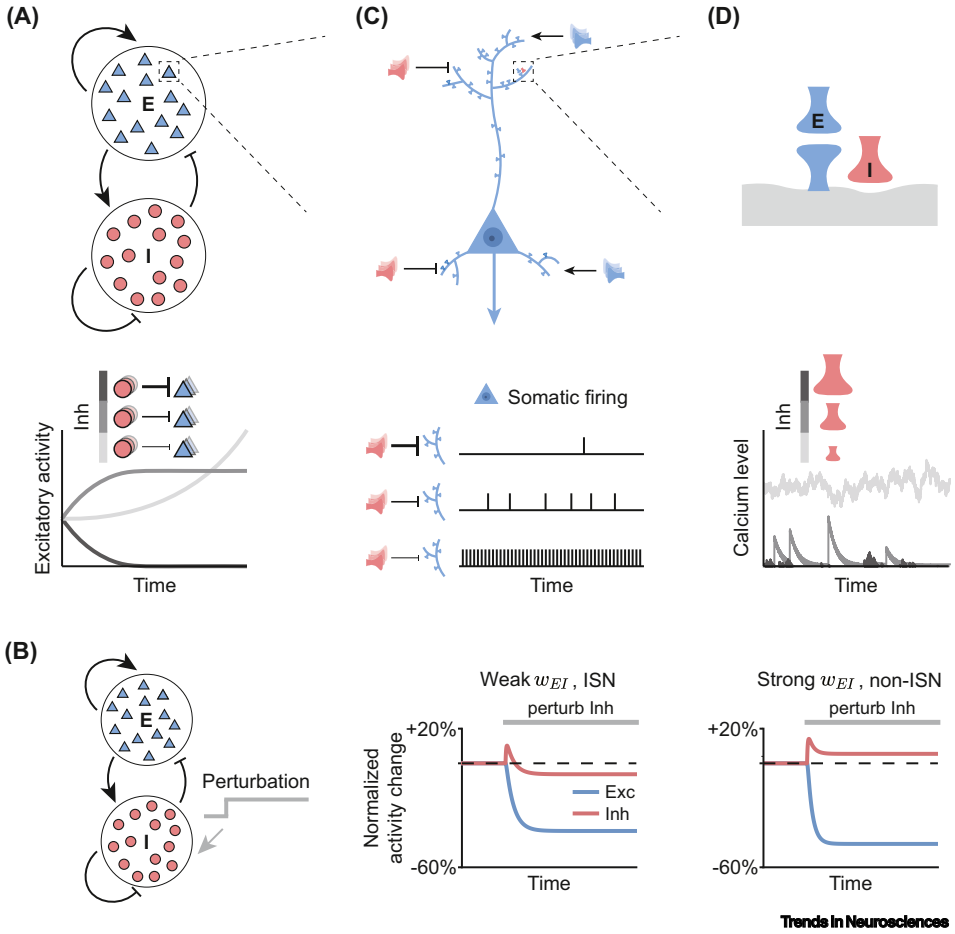


Figure 1. Inhibitory control of excitation at different scales. (A) At the network level (top), inhibition (Inh) affects excitatory population activity (bottom). Excessive inhibition can silence excitatory activity, insufficient inhibition can lead to the explosion of excitatory activity, while the appropriate amount of inhibition stabilizes network dynamics and maintains excitatory activity at a modest level. **(B)** Left: Assessing inhibition stabilization via the paradoxical effect by perturbing the inhibitory population. Middle: For weak inhibitory weights ($w_{EI}$), network activity is high and the network is in the inhibition-stabilized network (ISN) regime. Injecting additional excitatory currents into inhibitory neurons ('perturb Inh') leads to a paradoxical decrease of the inhibitory population response. Right: For strong $w_{EI}$, network activity is low and the network is in the non-ISN regime. Injecting additional excitatory currents into inhibitory neurons ('perturb Inh') does not generate a paradoxical response. (C)At the single neuron level (top), inhibition affects somatic firing (bottom). Excessive inhibition generates very little spiking, insufficient inhibition leads to high levels of spiking, while appropriate amount of inhibition leads to appropriate spiking levels. (D) At the dendritic level (top), inhibition influences the local calcium level (bottom). Excessive inhibition leads to extremely low calcium level locally on the dendrite, insufficient inhibition leads to extraordinarily high local calcium level, while the appropriate amount of inhibition leads to an appropriate local calcium level.

the network, cellular, and subcellular level, with different computational implications for circuit processing (Box 2) [29,74–78]. E/I balance is typically quantified by the E/I ratio, defined as the ratio of excitatory to inhibitory input currents. The E/I ratio can in return also affect the amount of inhibitory plasticity, with high initial E/I ratios resulting in stronger inhibitory potentiation, as shown in the mouse auditory cortex [44,79].

Various inhibitory plasticity rules have been proposed to regulate E/I balance in computational models [45,51,52,80–82]. The best-studied model of inhibitory plasticity, which has a symmetric Hebbian learning window (see Figure I in Box 1), can establish a precise E/I balance at the single-

Box 2. Different types of E/I balance

Neural circuits are known to maintain E/I balance [7,10]. E/I balance generally refers to the coregulation of excitation and inhibition and is typically measured by the ratio of excitatory and inhibitory inputs [10]. When excitation and inhibition are balanced at the population level but not necessarily at the single neuron level, the E/I balance is known as global balance [95,102]. Global balance can be achieved via input-dependent inhibitory plasticity rules [84]. If excitatory and inhibitory input currents onto a single neuron are balanced, or co-tuned, across the stimulus space, this is referred to as detailed balance [76–78,103]. Detailed balance can be established via inhibitory plasticity rules, which maintain a target firing rate at the single neuron level [45]. Additionally, when excitatory and inhibitory inputs are balanced also on a millisecond timescale, as observed experimentally [104,105], the E/I balance is known as tight balance, and loose balance otherwise [106]. The coexistence of tight and detailed balance is referred to as precise E/I balance and has been observed in several circuits, such as the zebrafish homolog of olfactory cortex [107] and mammalian hippocampus [108], where it is involved in efficient memory storage, millisecond-range input gating, and subthreshold gain control.

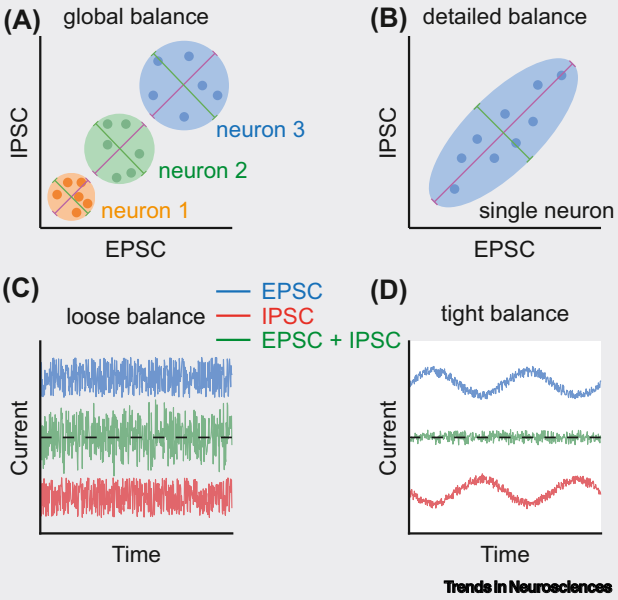


Figure I. Different types of excitation/inhibition (E/I) balance. (A) Global balance is characterized by a high degree of correlation between excitatory postsynaptic currents (EPSCs) and inhibitory postsynaptic currents (IPSCs) at the population level but a low degree of correlation for individual neurons across stimuli. Each dot represents a neuron–stimulus pair. Data for different neurons are marked in different colors. (B) Detailed balance is characterized by a high degree of correlation between EPSCs and IPSCs at the individual neuron level across stimuli. (C) Loose balance is characterized by a low degree of correlation between EPSCs and IPSCs over time. (D) Tight balance is characterized by tightly correlated EPSCs and IPSCs on a millisecond timescale. Panels (A) and (B) are adapted from [107].

neuron level on a millisecond timescale [45,83]. The learning rule achieves the balance by a negative feedback mechanism, which increases inhibitory synaptic strength for high postsynaptic firing rates and decreases inhibitory strength for low firing rates to counteract deviations from a target firing rate (Figure 1C), therefore maintaining a firing rate set-point for each individual neuron. How such a negative feedback mechanism might be implemented biologically remains an open question (see [14] for a discussion of the molecular mechanisms underlying inhibitory plasticity). Due to the resulting robust homeostatic properties, this rule is commonly used in recurrent network models [28,45]. Computational work has proposed several alternatives, including an input-dependent inhibitory plasticity rule [84], or a voltage-dependent plasticity rule [49], both of which can achieve firing rate heterogeneity as observed experimentally [69,85]. One caveat of all these inhibitory plasticity rules is the mismatch between timescales assumed in models and timescales measured in experiments. Most computational models rely on fast inhibitory plasticity

to guarantee homeostasis and establish an E/I balance [48,65]; however, it takes several tens of minutes to reach a stable baseline of inhibitory synaptic strength following plasticity induction in the mouse auditory cortex [44,76,78].

Recent experimental evidence suggests that E/I balance can even extend to local dendritic segments of single neurons [86] (Figure 1D). Inhibitory synapses form and change in strength to complement the dendritic organization of excitatory synaptic inputs, which often form local clusters based on coactivation [87,88], to regulate excitatory synaptic dynamics and plasticity [86,89,90]. For example, in the hippocampus, stimulating clustered excitatory synapses has been shown to trigger the *de novo* formation of inhibitory synapses [91], and a push–pull plasticity mechanism has been found to maintain the balance of local dendritic excitatory and inhibitory strength [92]. Also, inhibitory synapses in the neocortex remain stable if located in the proximity of excitatory synapses during normal visual experience [72]. Thus, while the presence of E/I balance on local stretches of dendrites is supported by experimental data, how it emerges during early postnatal development and how it is maintained during learning and perturbations remains an open question.

Besides regulating E/I balance and firing rates, inhibitory plasticity plays a more nuanced role in controlling the firing patterns of single neurons. By regulating the precise arrival of inhibitory inputs relative to excitatory inputs, experiments in the hippocampus have showed that inhibition can close or open the time window in which a spike is triggered [93]. Inhibitory plasticity can therefore dramatically affect the spike generation properties and spiking statistics of excitatory neurons, including neuronal input–output functions [94], pairwise spike correlations and spiking regularity [95,96], and criticality [97,98]. Both experimental and modeling work have showed that potentiating inhibition can decorrelate network activity [24,99,100] and switch network firing regimes [95] from oscillatory states supporting memory consolidation [101] to asynchronous irregular states supporting high memory capacity, despite the presence of noise [81]. Such switching could occur at different behavioral state transitions (e.g., from sleep to wake). Yet, direct evidence of inhibitory plasticity contributing to a dynamical switching between network firing regimes remains to be examined experimentally.

## Inhibitory control of excitatory plasticity

Experimental evidence has revealed that excitatory plasticity is jointly determined by factors like pre- and postsynaptic firing rates [2,4], spike timing [3,4], and dendritic calcium levels [5]. Since inhibition can influence all of these factors, it naturally also affects excitatory plasticity [12,109–111].

In experiments, the frequency of presynaptic stimulation can determine the sign of excitatory synaptic plasticity, with low-frequency stimulation favoring excitatory LTDand high-frequency stimulation inducing excitatory LTP [2]. Decreasing inhibition decreases the excitatory LTD/LTP threshold, making LTP induction easier, while increasing inhibition increases the LTD/LTP threshold and makes LTP induction more difficult [112] (Figure 2A). Based on these results, computational studies have demonstrated that a change of the inhibitory input (e.g., via inhibitory plasticity) can shift the threshold between LTP and LTD [47,48]. By keeping the firing rates exactly at the LTD/LTP threshold, inhibitory plasticity has been suggested as a mechanism to effectively switch excitatory plasticity off [48] (Figure 2A). Any deviation of the firing rates (e.g., via **disinhibition**) can then turn on excitatory plasticity. Such gating of excitatory plasticity has also been modeled at the level of individual inhibitory inputs on dendritic trees by affecting the amplitude of backpropagating action potentials and calcium spikes [113,114] (Figure 2B). Therefore, changes in inhibition can switch excitatory plasticity on or off, regulate how much plasticity is induced, or even dictate the sign of excitatory plasticity [38,115].

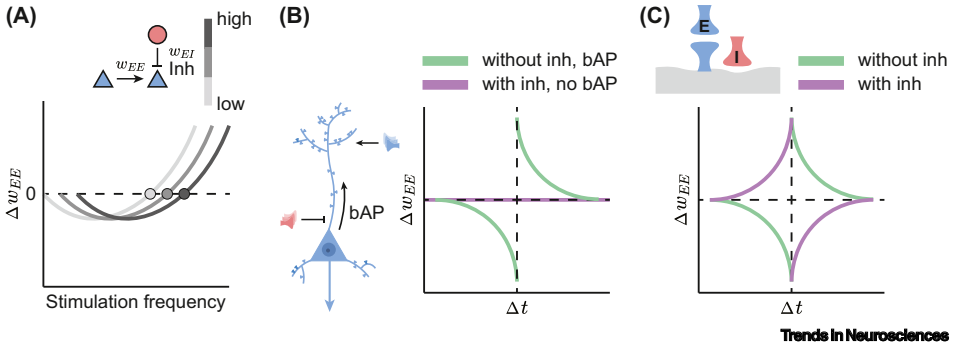


Figure 2. Inhibitory control of excitatory plasticity. (A) The level of inhibition (Inh), modulated by inhibitory weights ($w_{EI}$) or inhibitory firing rates, controls excitatory plasticity ($\Delta w_{EE}$). Higher (lower) level of inhibition leads to higher (lower) long-term depression (LTD)/long-term potentiation (LTP) threshold of excitatory plasticity as a function of the presynaptic stimulation frequency. Different dots represent corresponding LTD/LTP thresholds that separate the depression ($\Delta w_{EE} < 0$) and potentiation ($\Delta w_{EE} > 0$) of excitatory synapses onto excitatory neurons. Different grays represent different levels of inhibition. Panel (A) is adapted from [48,112]. (B) Strong inhibitory input can switch excitatory plasticity on or off via gating of a backpropagating action potential (bAP). In the absence of inhibition, the bAP propagates into the dendrite and spike-timing-dependent plasticity at the excitatory synapse is induced (green). By contrast, in the presence of inhibition, the bAP is suppressed and no synaptic plasticity is induced (purple). Panel (B) is adapted from [113]. **C.** (C) Local inhibitory input can affect calcium concentration in the dendritic spine and flip the excitatory spike-timing-dependent plasticity. Panel (C) is adapted from [115,125].

Multiple experimental studies have suggested disinhibition as a mechanism for the gating of excitatory plasticity [116]. Disinhibition can be induced by neuromodulators, including but not limited to acetylcholine, noradrenalin, and oxytocin [10,76], or by disinhibitory pathways involving multiple interneuron subtypes [117,118] (Box 3). For instance, elevated activity in vasoactive intestinal peptide (VIP)-expressing inhibitory neurons receiving top-down inputs can suppress activity in somatostatin (SST)-expressing inhibitory neurons and, as a result, disinhibit excitatory neurons and control excitatory plasticity [111,117–120].

At the dendritic level, inhibitory input onto the dendrite can affect postsynaptic calcium concentration at nearby excitatory spines [111,121] and, therefore, influence local excitatory plasticity [122,123]. Computational models have proposed that the dynamic local balancing of excitation by inhibition can change the shape of the learning rule for excitatory synapses [124–126]. For example, blocking inhibitory inputs can flip the spike-timing-dependency of excitatory plasticity [125], consistent with previous experimental findings [115] (Figure 2C). Furthermore, local changes in excitatory and inhibitory synapses are coordinated with each other via crosstalk, giving rise to the codependence of excitatory and inhibitory plasticity [7,8]. While these works clearly show that inhibitory synapses can control excitatory plasticity at multiple spatial scales, how this control is used during learning and its impact on behavior remains to be explored.

## Inhibitory plasticity in the formation of structured networks and resulting computation

Non-random structure is a hallmark of biological networks. Multiple computational studies have demonstrated that various network structures can form from the coordinated interaction between excitatory and inhibitory plasticity. This includes the emergence of receptive fields [45,47,48], place fields [27], and grid fields [27] through the refinement of feedforward excitatory and inhibitory connectivity, typically in settings with a single postsynaptic neuron based on input statistics [51–53]. In recurrent circuits, inhibitory plasticity also shapes neuronal assemblies [26,48] and chain-like structure [127,128], as well as ensuing tuning diversity and efficient sensory representation [100].

Box 3. Interneuron diversity

Interneurons exhibit high anatomical, electrophysiological, and functional diversity [157,158]. In the mouse neocortex, three major classes of interneurons expressing parvalbumin (PV), somatostatin (SST), and vasoactive intestinal peptide (VIP) constitute more than 80% of GABAergic interneurons [15]. Distinct interneuron subtypes target different domains of pyramidal cells. More specifically, PV neurons preferentially target perisomatic regions of pyramidal neurons, whereas SST neurons target distal dendritic regions of pyramidal neurons that also receive inhibition from neuron-derived neurotrophic factor (NDNF)-expressing interneurons in layer 1 [15,159].

The multiplicity of interneuron subtypes is implicated in diverse computations and cognitive functions, such as locomotion-induced gain modulation [160], selective attention [127], context-dependent modulation [31,33], predictive processing [32,161], and gating of synaptic plasticity [117,120]. For instance, long-range cortico-cortical projections activating upstream VIP neurons in the primary visual cortex exert spatially specific top-down modulation of visual processing, resembling selective attention [127]. In predictive processing framework, mismatches between sensory inputs and internally generated predictive signals evoke the activity of prediction-error neurons [32]. In the layer 2/3 of the primary visual cortex, prediction-error neurons balance inhibitory visual input mediated by SST against excitatory motor-related predictive input targeting VIP [161].

Strongly interconnected groups of excitatory neurons form assemblies, which have been proposed to be the basis of associative memory [129,130]. Inhibition can influence excitatory assemblies in two distinct ways. First, inhibitory neurons may be nonspecific and nonpreferentially target different excitatory assemblies, known as 'blanket of inhibition' [131] (Figure 3A). Second, inhibition may be stimulus-specific if distinct inhibitory neurons receive stimulus-specific feedforward drive, or if excitatory and inhibitory neurons with a similar stimulus tuning connect more strongly and form E/I assemblies, known as stimulus-specific feedback inhibition [132] (Figure 3B).

While many mechanisms are involved in the formation of excitatory assemblies [133], computational models have proposed an important role of inhibitory plasticity in preventing runaway excitation that results from the assemblies' repeated coactivation and preventing winner-take-all dynamics whereby a single assembly is always active [26,28,48]. Specific to forming E/I assemblies, both inhibitory synapses onto excitatory neurons and excitatory synapses onto inhibitory neurons need to be plastic in the recurrent circuit [63,134]. The resulting co-tuned feedback inhibition in networks with E/I assemblies can support network stability [60,132], changes in neuronal variability [135], and decision making in the presence of noise [136].

Irrespective of whether inhibition is unspecific or specific, modeling studies suggest that the plasticity of lateral inhibitory connections across assemblies can ensure that different memories encoded by different assemblies are easily discriminated [50,137]. Concurrently, multiple experimental studies have found evidence for the role of inhibition in memory recall. For instance, inactive memories can be unmasked by suppressing inhibitory neurons [138]. Using E/I assemblies as a model for associative memories, the inactive memories seem to remain in the quiescent state until being recalled by disinhibition [138,139]. Recent work in the human neocortex has further suggested that specific inhibition can avoid inappropriate interference of overlapping memories and permit continual learning [140,141].

The activation of E/I assemblies shaped by inhibitory plasticity has also been hypothesized to underlie the adaptation of behavioral responses to repeated stimulation (i.e., 'habituation') [139,142]. The ability to adapt to repeated stimuli, detect unexpected stimuli in the environment, and identify their relevance to execute appropriate behavioral reactions is important for survival. Inhibitory plasticity has been suggested to be important in shaping adaptation to repeated responses also at the cellular level in the mouse auditory cortex [143]. A recent computational study has provided a mechanistic insight on how inhibitory plasticity can shape the responses to repeated and novel stimuli [144]. While the repeated presentation of a stimulus evokes initially high activity of the excitatory assembly representing the stimulus, the subsequent increase of

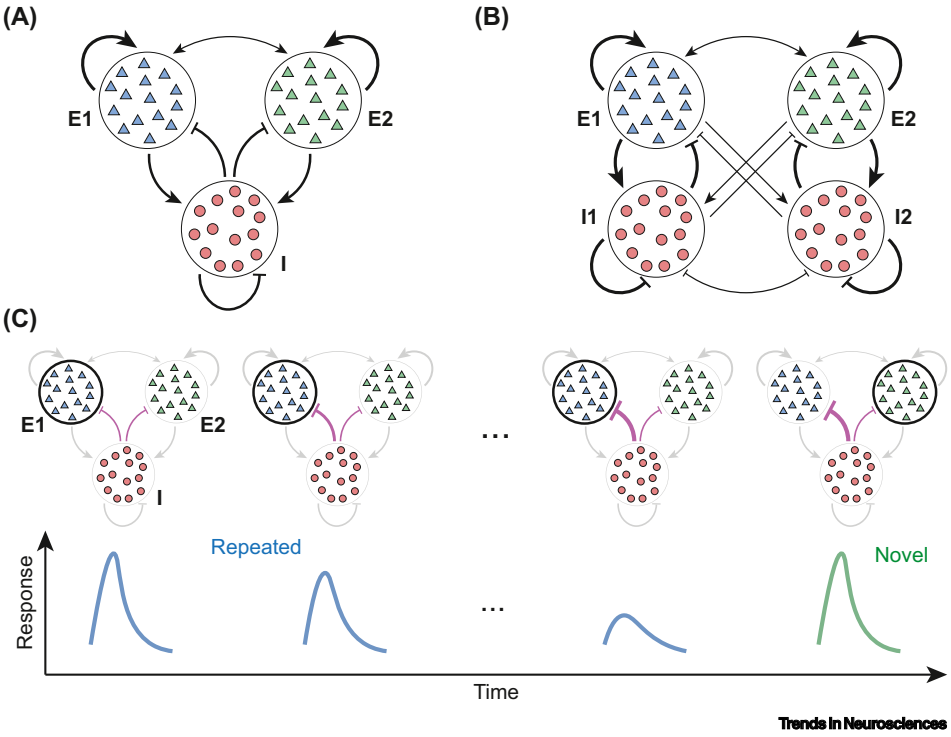


Figure 3. Unspecific versus specific inhibitory connectivity and the generation of adaptive and novelty responses. (A) Network with unspecific inhibition, in which different excitatory assemblies are inhibited by a single inhibitory population. (B) Network with stimulus-specific feedback inhibition, in which distinct excitatory assemblies are inhibited by non-overlapping inhibitory subpopulations. (C)The repeated and novel stimuli activate distinct excitatory assemblies, E1 and E2, respectively (activation marked with bold circles). Repeated presentation of the same stimulus leads to an increase of specific inhibitory synaptic strength onto the E1 assembly and a reduction of the evoked response (blue), while presenting the novel stimulus triggers a high response due to the weak inhibitory synaptic strength onto the E2 assembly (green).

inhibitory synaptic strengths suppresses the ensuing responses upon stimulus repetition. By contrast, a novel stimulus evokes a high response of its corresponding excitatory assembly since the inhibitory synapses onto the assembly do not potentiate (Figure 3C). While both blanket and stimulus-specific inhibition can capture adapted and elevated responses to repeated and novel stimuli, stimulus-specific inhibition is necessary for other adaptive phenomena [144]. This includes stimulus-specific adaptation, whereby excitatory neurons that are equally driven by two stimuli exhibit a higher response to the rarely presented stimulus, but a lower response to the frequently presented stimulus [145].

## Interneuron-specific plasticity and its functional implications

Inhibitory neurons can be divided into multiple distinct subtypes based on their electrophysiological, morphological, and transcriptomic properties (Box 3). Accumulating evidence also suggests that synapses from and to different interneuron subtypes undergo distinct forms of synaptic plasticity [16,17,37,146,147]. Computational models have capitalized on these experimental results of interneuron-specific plasticity and explored its role in different settings. In feedforward networks, modeling work has showed that the receptive field of a neuron may not be solely determined by the feedforward excitatory weight profiles, but is heavily modulated by inhibition from different pathways [53]. By exploring several candidate plasticity rules for the different inhibitory pathways, the authors found that the neuron's receptive field strongly depends on the

modulatory state of inhibition as an example of context-dependence [53].

Recent studies in the rodent hippocampus have identified learning rules describing the LTD of parvalbumin (PV) synapses and the LTP of SST synapses onto excitatory pyramidal neurons in CA1 during physiological activity patterns [16]. As PV and SST mainly target perisomatic regions receiving inputs from CA3 and distal dendritic regions receiving inputs from pyramidal neurons in entorhinal cortex, respectively, both experiments and modeling suggest that interneuron-specific plasticity might prioritize inputs from one pathway over another [16] (Figure 4A). As stronger inhibition resulting from the potentiation of SST synapses onto excitatory neurons can limit excitatory plasticity [120], modeling has suggested that interneuron-specific plasticity can promote the stability of place cells [16]. Recent experiments in CA1 suggest that even synapses from different interneurons targeting the same perisomatic regions of excitatory neurons can undergo opposite changes when animals explore novel environments [17] (Figure 4B). Since these two types of interneurons preferentially receiving different inputs fire at different phases of network theta rhythms associated with memory encoding and retrieval [148], the opposite regulation of interneuron-specific plasticity may impact memory formation and maintenance. Future computational models could help uncover how the opposing plasticity mechanisms support long-term memories.

In addition to hippocampus, interneuron-specific plasticity rules based on spike timing have been reported in layer 2/3 of mouse orbitofrontal cortex and implicated in assembly formation in recurrent network models [50]. More specifically, PV synapses onto excitatory neurons follow a symmetric Hebbian learning rule and appear to be important for network stability; by contrast, SST synapses onto excitatory neurons follow an asymmetric Hebbian learning rule and appear to enhance competition between assemblies [50] (Box 1). Although a learning rule has not yet been characterized for neuron-derived neurotrophic factor (NDNF)-expressing interneurons, experimental studies have revealed that inhibition mediated by NDNF interneurons in layer 1 of the auditory cortex changes after associative auditory fear conditioning , and have suggested that NDNF interneurons and their plasticity are involved in the formation of associative memories [149].

While significantly less studied, recent work has begun to explore synapses between inhibitory neurons, including their impact on E/I balance in recent connectomic studies [150], on generating long neuronal timescales that support working memory, and on memory storage in computational models [151,152]. Yet, little is known about the plasticity of these inhibitory-to-inhibitory connections experimentally. Computational models here play an important role in revealing the functional consequences of this type of plasticity. For instance, a two-stage model showed that an initial stage of SST to PV plasticity guides the subsequent plasticity of excitatory-to-excitatory connections in a recurrent network underlying visual stimulus selectivity [153]. Recent modeling work has also begun investigating recurrent network models where multiple synapse types are simultaneously plastic and found that experimentally observed dynamics and computations can emerge from the complex interplay of many plasticity mechanisms. Given the high-dimensional space of learning rule parameters, when such models succeed in finding stable regimes, they can provide predictions for the learning mechanisms in real biological circuits. Deriving learning rules via optimizing a desired function has provided a new promising approach to study plasticity [154,155]. In an elegant example, recent studies derived plasticity rules from the perspective of optimizing a loss function to achieve firing rate set-points; the emergent networks could then generate self-sustained, inhibition-stabilized dynamics [156] and stimulus-specific feedback inhibition [134]. Even without deriving novel learning rules, combining classical Hebbian plasticity with synapse-type-specific competition for synaptic resources can yield novel dynamics such as the development of stimulus selectivity, E/I balance, decorrelated neural activity, assembly structures, and response normalization [63].

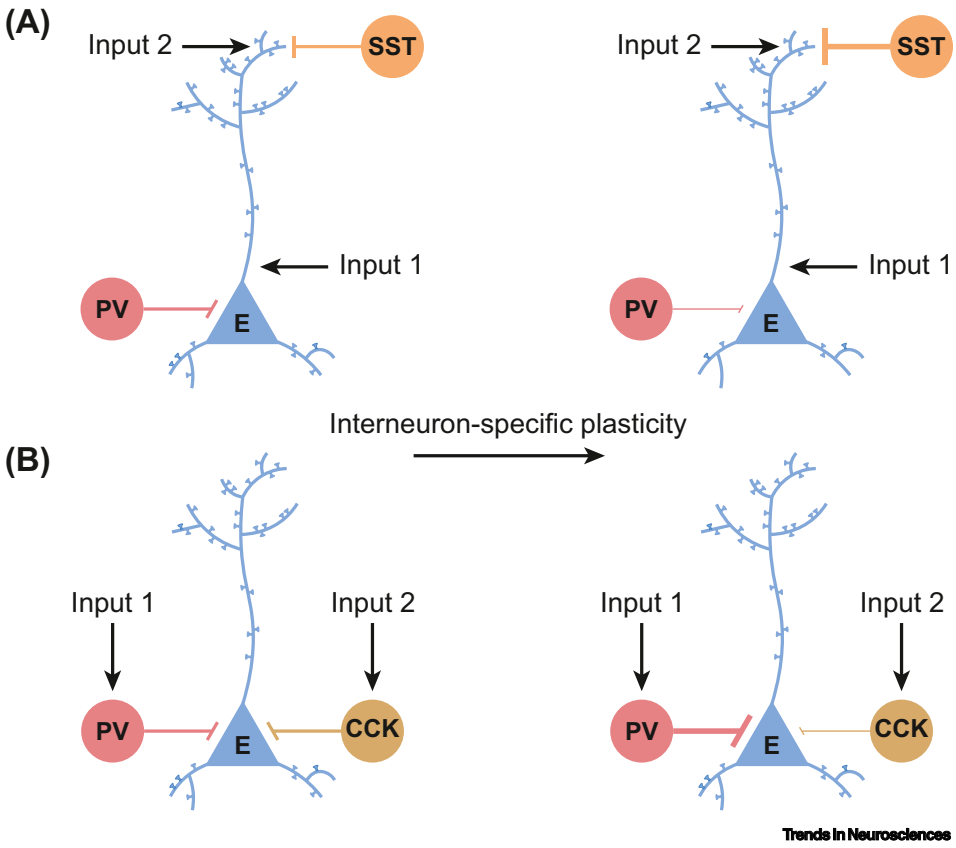


Figure 4. Interneuron-specific plasticity. (A) Inhibitory synapses from parvalbumin (PV)- (red) and somatostatin (SST)- (orange) expressing neurons onto hippocampal CA1 pyramidal neurons (blue) are weakened and enhanced, respectively, during physiological firing patterns [16]. This interneuron-specific plasticity can prioritize proximal input from CA3 over distal input from entorhinal cortex. (B) Perisomatic inhibitory synapses from PV- (red) and cholecystokinin (CCK)-expressing (brown) neurons onto recently activated hippocampal CA1 pyramidal neurons (blue) undergo long-term potentiation and long-term depression, respectively when animals are engaged in novel environments [17].

## Concluding remarks and future perspectives

Over the past two decades, our understanding of the inhibitory control of circuit organization and dynamics, as well as the potential to modulate this control via plastic inhibition, has significantly grown. Inhibitory synapses in the brain are highly dynamic and regulated by various plasticity mechanisms, including short-term plasticity operating at the timescale of milliseconds to seconds [162] as well as long-term plasticity acting at the timescale of minutes to hours [44]. Here, we summarized studies on the long-term plasticity of inhibitory-to-excitatory synapses, referred to as inhibitory plasticity. As discussed in this review, abundant evidence suggests that inhibitory plasticity is important for establishing and maintaining E/I balance, achieving firing rate homeostasis, controlling excitatory plasticity, and shaping network connectivity throughout the entire life of an organism. Nonetheless, it remains unclear if the learning rules that characterize inhibitory plasticity in development are the same as those operating in adulthood (see Outstanding questions). Complementary to the growing number of experimental studies on inhibitory plasticity, theoretical and computational approaches have played an important role in synthesizing the available data to reveal how inhibition regulates various aspects of circuit function. This has generated mechanistic insights into the function of inhibitory plasticity at several spatial scales, from the local dendritic regulation of E/I balance, to the cellular control of spiking properties, and the maintenance of stable

### Outstanding questions

Neuronal activity during development is typically generated spontaneously in the absence of sensory experience. This activity operates on much slower timescales (hundreds of milliseconds) compared with the sensory-driven activity patterns (few to tens of milliseconds) in adulthood. Do the activity-dependent learning rules that characterize inhibitory plasticity integrate activity at different timescales in development and adulthood?

The phenomenological learning rules that determine how inhibitory plasticity depends on rates and spike timing can be modulated by various external factors. How do different neuromodulators, behavioral states, and environmental perturbations affect inhibitory plasticity rules?

How are phenomenological descriptions of inhibitory plasticity implemented with the biological machinery of molecular interactions?

Distinct forms of E/I balance might be beneficial for different demands in development versus adulthood. How are different types of E/I balance dynamically regulated by inhibitory plasticity over multiple timescales to serve specific goals?

E/I balance also exists at different spatial scales. Are there shared principles underlying the establishment of E/I balance across these different scales? What are the functional implications of breaking E/I balance at some spatial scales but not others?

Interneurons come in diverse subtypes, receive inputs from different pathways, and target excitatory neurons in different locations (e.g., cell body versus dendrite). This diversity is also reflected in the types of plasticity rules experienced at the synapses. How can interneuron-specific plasticity rules be described as a function of firing rates, spike timing, and calcium level?

Inhibitory plasticity rules might also differ across brain regions. How do different brain regions coordinate the potentially different forms of inhibitory plasticity they express to maintain biologically reasonable activity levels and process information?

activity patterns and connectivity structures at the network level. At the same time, we have highlighted that inhibitory control also occurs at multiple temporal scales from the regulation of fast spiking to the slower calcium dynamics and even slower timescales at which measurable changes in synaptic strength can be observed.

Despite this progress, many open challenges remain due to the high diversity of inhibitory neurons and the interneuron-specific plasticity at different synapse types. Experimentally, the development of transgenic and recording techniques opens new possibilities to record activity from multiple interneuron subtypes simultaneously and probe the rules that govern synaptic plasticity. Concurrently, computational models and theories are becoming paramount. First, they are essential to understand the complex interactions of different plasticity mechanisms, especially in highly recurrent circuits with non-intuitive dynamics. Second, models can explore candidate plasticity mechanisms and study their functional implications. Last, theoretical work also enables the exploration of more abstract concepts, like inhibition-stabilization, as general frameworks for circuit processing, which can be established and modulated through inhibitory plasticity.

Different computations, such as selective attention, context-dependent modulation, and predictive processing, typically require diverse interneuron subtypes with specific synaptic connections. How do interneuron-specific plasticity mechanisms establish the network connectivity enabling diverse computations?

## Acknowledgments

This work was supported by the Max Planck Society and has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 804824 to J.G.) and from the Deutsche Forschungsgemeinschaft in the Collaborative Research Centre 1080. We thank Katharina A. Wilmes, Everton J. Agnes, Elizabeth Herbert, and Dylan Festa for comments on the manuscript.

## Declaration of interests

The authors declare no conflicts of interest.

## References

1. Magee, J.C. and Grienberger, C. (2020) Synaptic plasticity forms and functions. *Annu. Rev. Neurosci.* 43, 95–117
2. Kirkwood, A. *et al.* (1996) Experience-dependent modification of synaptic plasticity in visual cortex. *Nature* 381, 526–528
3. Bi, G.-g. and Poo, M.-M. (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* 18, 10464–10472
4. Sjöström, P.J. *et al.* (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32, 1149–1164
5. Inglebert, Y. *et al.* (2020) Synaptic plasticity rules with physiological calcium levels. *Proc. Natl. Acad. Sci. U. S. A.* 117, 33639–33648
6. Turrigiano, G.G. and Nelson, S.B. (2004) Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.* 5, 97–107
7. Hennequin, G. *et al.* (2017) Inhibitory plasticity: balance, control, and codependence. *Annu. Rev. Neurosci.* 40, 557–579
8. Herstel, L.J. and Wierenga, C.J. (2021) Network control through coordinated inhibition. *Curr. Opin. Neurobiol.* 67, 34–41
9. Chen, J.L. and Nedivi, E. (2013) Highly specific structural plasticity of inhibitory circuits in the adult neocortex. *Neuroscientist* 19, 384–393
10. Froemke, R.C. (2015) Plasticity of cortical excitatory-inhibitory balance. *Annu. Rev. Neurosci.* 38, 195–219
11. Kripkee, B. and Froemke, R.C. (2017) *Organization and Plasticity of Cortical Inhibition*, Oxford University Press
12. Chiu, C.Q. *et al.* (2019) Preserving the balance diverse forms of long-term GABAergic synaptic plasticity. *Nat. Rev. Neurosci.* 20, 272–281
13. Gandolfi, D. *et al.* (2020) Inhibitory plasticity: from molecules to computation and beyond. *Int. J. Mol. Sci.* 21, 1805
14. Capogna, M. *et al.* (2021) The ins and outs of inhibitory synaptic plasticity: neuron types, molecular mechanisms and functional roles. *Eur. J. Neurosci.* 54, 6882–6901
15. Tremblay, R. *et al.* (2016) GABAergic interneurons in the neocortex: from cellular properties to circuits. *Neuron* 91, 260–292
16. Udakis, M. *et al.* (2020) Interneuron-specific plasticity at parvalbumin and somatostatin in-hibitory synapses onto CA1 pyramidal neurons shapes hippocampal output. *Nat. Commun.* 11, 4395
17. Yap, E.L. *et al.* (2021) Bidirectional perisomatic inhibitory plasticity of a Fos neuronal network. *Nature* 590, 115–121
18. Ganguly, K. *et al.* (2001) GABA itself promotes the developmental switch of neuronal GABAergic responses from excitation to inhibition. *Cell* 105, 521–532
19. Ben-Ari, Y. *et al.* (2007) GABA: a pioneer transmitter that excites immature neurons and generates primitive oscillations. *Physiol. Rev.* 87, 1215–1284
20. Cancedda, L. *et al.* (2007) Excitatory GABA action is essential for morphological maturation of cortical neurons in vivo. *J. Neurosci.* 27, 5224–5235
21. Wang, D.D. and Kriegstein, A.R. (2008) GABA regulates excitatory synapse formation in the neocortex via nmda NMDA receptor activation. *J. Neurosci.* 28, 5547–5558
22. Kirmse, K. *et al.* (2015) GABA depolarizes immature neurons and inhibits network activity in the neonatal neocortex in vivo. *Nat. Commun.* 6, 7750
23. Murata, Y. and Colonnese, M.T. (2020) GABAergic interneurons excite neonatal hippocampus in vivo. *Sci. Adv.* 6, 1–11
24. Chini, M. *et al.* (2022) An increase of inhibition drives the developmental decorrelation of neural activity. *eLife* 11, 1–28
25. Hensch, T.K. (2005) Critical period plasticity in local cortical circuits. *Nat. Rev. Neurosci.* 6, 877–888
26. Litwin-Kumar, A. and Doiron, B. (2014) Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nat. Commun.* 5, 5319
27. Weber, S.N. and Sprekeler, H. (2018) Learning place cells, grid cells and invariances with excitatory and inhibitory plasticity. *eLife* 7, e34560
28. Ocker, G.K. and Doiron, B. (2019) Training and spontaneous reinforcement of neuronal assemblies by spike timing plasticity. *Cereb. Cortex* 29, 937–951

29. House, D.R. *et al.* (2011) Parallel regulation of feedforward inhibition and excitation during whisker map plasticity. *Neuron* 72, 819–831
30. Maffei, A. *et al.* (2006) Potentiation of cortical inhibition by visual deprivation. *Nature* 443, 81–84
31. Kuchibhotla, K.V. *et al.* (2017) Parallel processing by cortical inhibition enables context-dependent behavior. *Nat. Neurosci.* 20, 62–71
32. Keller, G.B. and Mrsic-Flogel, T.D. (2018) Predictive processing: a canonical cortical computation. *Neuron* 100, 424–435
33. Keller, A.J. *et al.* (2020) A disinhibitory circuit for contextual modulation in primary visual cortex. *Neuron* 108, 1181–1193
34. Hertäg, L. and Clopath, C. (2022) Prediction-error neurons in circuits with multiple neuron types: formation, refinement and functional implications. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115699119
35. Vickers, E.D. *et al.* (2018) Parvalbumin-interneuron output synapses show spike-timing-dependent plasticity that contributes to auditory map remodeling. *Neuron* 99, 720–735
36. Woodin, M.A. *et al.* (2003) Coincident pre- and postsynaptic activity modifies GABAergic synapses by postsynaptic changes in Cl- transporter activity. *Neuron* 39, 807–820
37. Chiu, C.Q. *et al.* (2018) Input-specific NMDAR-dependent potentiation of dendritic GABAergic inhibition. *Neuron* 97, 368–377
38. Wang, L. and Maffei, A. (2014) Inhibitory plasticity dictates the sign of plasticity at excitatory synapses. *J. Neurosci.* 34, 1083–1093
39. Mellor, J. (2018) Synaptic plasticity at hippocampal synapses: experimental background. In *Springer Series in Computational Neuroscience* (Cutsuridis, V. *et al*., eds), pp. 201–226, Springer
40. Caillard, O. *et al.* (1999) Long-term potentiation of GABAergic synaptic transmission in neonatal rat hippocampus. *J. Physiol.* 518, 109–119
41. Shew, T. *et al.* (2000) Mechanisms involved in tetanus induced potentiation of fast IPSCs in rat hippocampal CA1 neurons. *J. Neurophysiol.* 83, 3388–3401
42. Holmgren, C.D. and Zilberter, Y. (2001) Coincident spiking activity induces long-term changes in inhibition of neocortical pyramidal cells. *J. Neurosci.* 21, 8270–8277
43. Haas, J.S. *et al.* (2006) Spike-timing-dependent plasticity of inhibitory synapses in the entorhinal cortex. *J. Neurophysiol.* 96, 3305–3313
44. D'amour, J.A. and Froemke, R.C. (2015) Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* 86, 514–528
45. Vogels, T.P. *et al.* (2011) Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* 334, 1569–1573
46. Bourjaily, M.A. and Miller, P. (2011) Synaptic plasticity and connectivity requirements to produce stimulus-pair specific responses in recurrent networks of spiking neurons. *PLoS Comput. Biol.* 7, e1001091
47. Clopath, C. *et al.* (2016) Receptive field formation by interacting excitatory and inhibitory synaptic plasticity. *bioRxiv* Published online July 29, 2016. https://doi.org/10.1101/066589
48. Miehl, C. and Gjorgjieva, J. (2022) Stability and learning in excitatory synapses by non-linear inhibitory plasticity. *bioRxiv* Published online March 29, 2022. https://doi.org/10.1101/2022.03.28.486052
49. Pedrosa, V. and Clopath, C. (2020) Voltage-based inhibitory synaptic plasticity: network regulation, diversity, and flexibility. *bioRxiv* Published online December 9, 2020. https://doi.org/10.1101/2020.12.08.416263
50. Lagzi, F. *et al.* (2021) Assembly formation is stabilized by parvalbumin neurons and accelerated by somatostatin neurons. *bioRxiv* Published online September 7, 2021. https://doi.org/10.1101/2021.09.06.459211
51. Luz, Y. and Shamir, M. (2012) Balancing feed-forward excitation and inhibition via Hebbian inhibitory synaptic plasticity. *PLoS Comput. Biol.* 8, e1002334
52. Kleberg, F.I. *et al.* (2014) Excitatory and inhibitory STDP jointly tune feed-forward neural circuits to selectively propagate correlated spiking activity. *Front. Comput. Neurosci.* 8, 53
53. Agnes, E.J. *et al.* (2020) Complementary inhibitory weight profiles emerge from plasticity and allow flexible switching of receptive fields. *J. Neurosci.* 40, 9634–9649
54. Li, N. *et al.* (2019) Spatiotemporal constraints on optogenetic inactivation in cortical circuits. *eLife* 8, e48622
55. Sanzeni, A. *et al.* (2020) Inhibition stabilization is a widespread property of cortical networks. *eLife* 9, e54875
56. Murphy, B.K. and Miller, K.D. (2009) Article balanced amplification a new mechanism of selective amplification of neural activity patterns. *Neuron* 61, 635–648
57. Rubin, D.B. *et al.* (2015) The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* 85, 402–417
58. Sadeh, S. and Clopath, C. (2021) Inhibitory stabilization and cortical computation. *Nat. Rev. Neurosci.* 22, 21–37
59. Tsodyks, M.V. *et al.* (1997) Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.* 17, 4382–4388
60. Wu, Y.K. and Zenke, F. (2021) Nonlinear transient amplification in recurrent neural networks with short-term plasticity. *eLife* 10, e71263
61. Sadeh, S. *et al.* (2017) Assessing the role of inhibition in stabilizing neocortical networks requires large-scale perturbation of the inhibitory population. *J. Neurosci.* 37, 12050–12067
62. Mahrach, A. *et al.* (2020) Mechanisms underlying the response of mouse cortical networks to optogenetic manipulation. *eLife* 9, 1–37
63. Eckmann, S. and Gjorgjieva, J. (2022) Synapse-type-specific competitive Hebbian learning forms functional recurrent networks. *bioRxiv* Published online March 14, 2022. https://doi.org/10.1101/2022.03.11.483899
64. Gainey, M.A. and Feldman, D.E. (2017) Multiple shared mechanisms for homeostatic plasticity in rodent somatosensory and visual cortex. *Philos. Trans. R. Soc. B. Biol. Sci.* 372, 20160157
65. Sprekeler, H. (2017) Functional consequences of inhibitory plasticity: homeostasis, the excitation-inhibition balance and beyond. *Curr. Opin. Neurobiol.* 43, 198–203
66. Kilman, V. *et al.* (2002) Activity deprivation reduces miniature IPSC amplitude by decreasing the number of postsynaptic GABAA receptors clustered at neocortical synapses. *J. Neurosci.* 22, 1328–1337
67. Hartman, K.N. *et al.* (2006) Activity-dependent regulation of inhibitory synaptic transmission in hippocampal neurons. *Nat. Neurosci.* 9, 642–649
68. Li, L. *et al.* (2014) Rapid homeostasis by disinhibition during whisker map plasticity. *Proc. Natl. Acad. Sci. U. S. A.* 111, 1616–1621
69. Hengen, K.B. *et al.* (2016) Neuronal firing rate homeostasis is inhibited by sleep and promoted by wake. *Cell* 165, 180–191
70. Nahmani, M. and Turrigiano, G.G. (2014) Deprivation-induced strengthening of presynaptic and postsynaptic inhibitory transmission in layer 4 of visual cortex during the critical period. *J. Neurosci.* 34, 2571–2582
71. Miska, N.J. *et al.* (2018) Sensory experience inversely regulates feedforward and feedback excitation-inhibition ratio in rodent visual cortex. *eLife* 7, e38846
72. Chen, J.L. *et al.* (2012) Clustered dynamics of inhibitory synapses and dendritic spines in the adult neocortex. *Neuron* 74, 361–373
73. van Versendaal, D. *et al.* (2012) Elimination of inhibitory synapses is a major component of adult ocular dominance plasticity. *Neuron* 74, 374–383
74. Barnes, S.J. *et al.* (2015) Subnetwork-specific homeostatic plasticity in mouse visual cortex in vivo. *Neuron* 86, 1290–1303
75. Keck, T. *et al.* (2017) Interactions between synaptic homeostatic mechanisms an attempt to reconcile BCM theory, synaptic scaling, and changing excitation/inhibition balance. *Curr. Opin. Neurobiol.* 43, 87–93
76. Froemke, R.C. *et al.* (2007) A synaptic memory trace for cortical receptive field plasticity. *Nature* 450, 425–429
77. Dorm, A.L. *et al.* (2010) Developmental sensory experience balances cortical excitation and inhibition. *Nature* 465, 932–936
78. Field, R.E. *et al.* (2020) Heterosynaptic plasticity determines the set point for cortical excitatory-inhibitory balance. *Neuron* 106, 842–854

79. Aljadeff, J. *et al.* (2019) Cortical credit assignment by Hebbian, neuromodulatory and inhibitory plasticity. *arXiv* Published online November 1, 2019. https://doi.org/10.48550/arxiv.1911.00307
80. Yger, P. *et al.* (2015) Fast learning with weak synaptic plasticity. *J. Neurosci.* 35, 13351–13362
81. Rubin, R. *et al.* (2017) Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *Proc. Natl. Acad. Sci. U. S. A.* 114, E9366–E9375
82. Baker, C. *et al.* (2020) Nonlinear stimulus representations in neural circuits with approximate excitatory-inhibitory balance. *PLoS Comput. Biol.* 16, e1008192
83. Akil, A.E. *et al.* (2021) Balanced networks under spike-time dependent plasticity. *PLoS Comput. Biol.* 17, e1008958
84. Kaleb, K. *et al.* (2021) Network-centered homeostasis through inhibition maintains hippocampal spatial map and cortical circuit function. *Cell Rep.* 36, 109577
85. Buzsáki, G. and Mizuseki, K. (2014) The log-dynamic brain: how skewed distributions affect network operations. *Nat. Rev. Neurosci.* 15, 264–278
86. Iascone, D.M. *et al.* (2020) Whole-neuron synaptic mapping reveals spatially precise excitatory / inhibitory balance limiting dendritic and somatic spiking. *Neuron* 106, 566–578
87. Kleindienst, T. *et al.* (2011) Activity-dependent clustering of functional synaptic inputs on developing hippocampal dendrites. *Neuron* 72, 1012–1024
88. Takahashi, N. *et al.* (2016) Active cortical dendrites modulate perception. *Science* 354, 1159–1165
89. Boivin, J.R. and Nedivi, E. (2018) Functional implications of inhibitory synapse placement on signal processing in pyramidal neuron dendrites. *Curr. Opin. Neurobiol.* 51, 16–22
90. Kirchner, J.H. and Gjorgjieva, J. (2021) Emergence of local and global synaptic organization on cortical dendrites. *Nat. Commun.* 12, 4005
91. Hu, H.Y. *et al.* (2019) Endocannabinoid signaling mediates local dendritic coordination between excitatory and inhibitory synapses. *Cell Rep.* 27, 666–675
92. Liu, G. (2004) Local structural balance and functional interaction of excitatory and inhibitory synapses in hippocampal dendrites. *Nat. Neurosci.* 7, 373–379
93. Pouille, F. and Scanziani, M. (2001) Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. *Science* 293, 1159–1163
94. Carvalho, T.P. and Buonomano, D.V. (2009) Differential effects of excitatory and inhibitory plasticity on synaptically driven neuronal input-output functions. *Neuron* 61, 774–785
95. Brunel, N. (2000) Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.* 8, 183–208
96. Cardin, J.A. (2018) Inhibitory interneurons regulate temporal precision and correlations in cortical circuits. *Trends Neurosci.* 41, 689–700
97. Stepp, N. *et al.* (2015) Synaptic plasticity enables adaptive self-tuning critical networks. *PLoS Comput. Biol.* 11, e1004043
98. Ma, Z. *et al.* (2019) Cortical circuit dynamics are homeostatically tuned to criticality in vivo. *Neuron* 104, 655–664
99. Duarte, R.C.F. and Morrison, A. (2014) Dynamic stability of sequential stimulus representations in adapting neuronal networks. *Front. Comput. Neurosci.* 8, 1066828
100. Larisch, R. *et al.* (2021) Sensory coding and contrast invariance emerge from the control of plastic inhibition over emergent selectivity. *PLoS Comput. Biol.* 17, e1009566
101. Buzsáki, G. and Draguhn, A. (2004) Neuronal oscillations in cortical networks. *Science* 304, 1926–1929
102. Van Vreeswijk, C. and Sompolinsky, H. (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274, 1724–1726
103. Wehr, M. and Zador, A.M. (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* 426, 442–446
104. Wilent, W.B. and Contreras, D. (2005) Dynamics of excitation and inhibition underlying stimulus selectivity in rat somatosensory cortex. *Nat. Neurosci.* 8, 1364–1370
105. Okun, M. and Lampl, I. (2008) Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.* 11, 535–537
106. Denève, S. and Machens, C.K. (2016) Efficient codes and balanced networks. *Nat. Neurosci.* 19, 375–382
107. Rupprecht, P. and Friedrich, R.W. (2018) Precise synaptic balance in the zebrafish homolog of olfactory cortex. *Neuron* 100, 669–683
108. Bhatia, A. *et al.* (2019) Precise excitation-inhibition balance controls gain and timing in the hippocampus. *eLife* 8, e43415
109. Paulsen, O. and Moser, E.I. (1998) A model of hippocampal memory encoding and retrieval: GABAergic control of synaptic plasticity. *Trends Neurosci.* 21, 273–278
110. Vogels, T.P. *et al.* (2013) Inhibitory synaptic plasticity: spike timing-dependence and putative network function. *Front. Neural Circ.* 7, 119
111. Hattori, R. *et al.* (2017) Functions and dysfunctions of neocortical inhibitory neuron subtypes. *Nat. Neurosci.* 20, 1199–1208
112. Steele, P.M. and Mauk, M.D. (1999) Inhibitory control of LTP and LTD: stability of synapse strength. *J. Neurophysiol.* 81, 1559–1566
113. Wilmes, K.A. *et al.* (2016) Inhibition as a binary switch for excitatory plasticity in pyramidal neurons. *PLoS Comput. Biol.* 12, e1004768
114. Wilmes, K.A. *et al.* (2017) Spike-timing dependent inhibitory plasticity to learn a selective gating of backpropagating action potentials. *Eur. J. Neurosci.* 45, 1032–1043
115. Paille, V. *et al.* (2013) GABAergic circuits control spike-timing-dependent plasticity. *J. Neurosci.* 33, 9353–9363
116. Letzkus, J.J. *et al.* (2015) Disinhibition, a circuit mechanism for associative learning and nemory. *Neuron* 88, 264–276
117. Krabbe, S. *et al.* (2019) Adaptive disinhibitory gating by VIP interneurons permits associative learning. *Nat. Neurosci.* 22, 1834–1843
118. Canto-Bustos, M. *et al.* (2022) Disinhibitory circuitry gates associative synaptic plasticity in olfactory cortex. *J. Neurosci.* 42, 2942–2950
119. Adler, A. *et al.* (2019) Somatostatin-expressing interneurons enable and maintain learning-dependent sequential activation of pyramidal neurons. *Neuron* 102, 202–216
120. Williams, L.E. and Holtmaat, A. (2019) Higher-order thalamocortical inputs gate synaptic long-term potentiation via disinhibition. *Neuron* 101, 91–102
121. Chiu, C.Q. *et al.* (2013) Compartmentalization of GABAergic inhibition by dendritic spines. *Science* 340, 759–763
122. Hayama, T. *et al.* (2013) GABA promotes the competitive selection of dendritic spines by controlling local Ca2+ signaling. *Nat. Neurosci.* 16, 1409–1416
123. Müllner, F.E. *et al.* (2015) Precision of Inhibition Dendritic inhibition: dendritic inhibition by individual GABAergic synapses on hippocampal pyramidal cells is confined in space and time. *Neuron* 87, 576–589
124. Agnes, E.J. and Vogels, T.P. (2021) Interacting synapses stabilise both learning and neuronal dynamics in biological networks. *bioRxiv* Published online April 4, 2021. https://doi.org/10.1101/2021.04.01.437962
125. Hiratani, N. and Fukai, T. (2017) Detailed dendritic excitatory/inhibitory balance through heterosynaptic spike-timing-dependent plasticity. *J. Neurosci.* 37, 12106–12122
126. Mikulasch, F.A. *et al.* (2021) Local dendritic balance enables learning of efficient representations in networks of spiking neurons. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2021925118
127. Zhang, S. *et al.* (2014) Long-range and local circuits for top-down modulation of visual cortex processing. *Science* 345, 660–665
128. Maes, A. *et al.* (2020) Learning spatiotemporal signals using a recurrent spiking network that discretizes time. *PLoS Comput. Biol.* 16, e1007606
129. Buzsáki, G. (2010) Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* 68, 362–385
130. Carrillo-Reid, L. and Yuste, R. (2020) Playing the piano with the cortex: role of neuronal ensembles and pattern completion in perception and behavior. *Curr. Opin. Neurobiol.* 64, 89–95
131. Fino, E. and Yuste, R. (2011) Dense inhibitory connectivity in neocortex. *Neuron* 69, 1188–1203
132. Znamenskiy, P. *et al.* (2018) Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. *bioRxiv* Published online April 4, 2018. https://doi.org/10.1101/294835

133. Miehl, C. *et al.* (2022) Formation and computational implications of assemblies in neural circuits. *J. Physiol.* Published online September 6, 2022. https://doi.org/10.1113/JP282750
134. Mackwood, O. *et al.* (2021) Learning excitatory-inhibitory neuronal assemblies in recurrent networks. *eLife* 10, e59715
135. Rost, T. *et al.* (2018) Winnerless competition in clustered balanced networks: inhibitory assemblies do the trick. *Biol. Cybern.* 112, 81–98
136. Najafi, F. *et al.* (2020) Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. *Neuron* 105, 165–179
137. Herpich, J. and Tetzlaff, C. (2019) Principles underlying the input-dependent formation and organization of memories. *Netw. Neurosci.* 3, 606–634
138. Barron, H.C. *et al.* (2016) Unmasking latent inhibitory connections in human cortex to reveal dormant cortical memories. *Neuron* 90, 191–203
139. Barron, H.C. *et al.* (2017) Inhibitory engrams in perception and memory. *Proc. Natl. Acad. Sci.* 114, 6666–6674
140. Koolschijn, R.S. *et al.* (2019) The hippocampus and neocortical inhibitory engrams protect against memory interference. *Neuron* 101, 528–541
141. Barron, H.C. (2021) Neural inhibition for continual learning and memory. *Curr. Opin. Neurobiol.* 67, 85–94
142. Ramaswami, M. (2014) Network plasticity in adaptive filtering and behavioral habituation. *Neuron* 82, 1216–1229
143. Natan, R.G. *et al.* (2017) Cortical interneurons differentially shape frequency tuning following adaptation. *Cell Rep.* 21, 878–890
144. Schulz, A. *et al.* (2021) The generation of cortical novelty responses through inhibitory plasticity. *eLife* 10, e65309
145. Ulanovsky, N. *et al.* (2003) Processing of low-probability sounds by cortical neurons. *Nat. Neurosci.* 6, 391–398
146. Chen, S.X. *et al.* (2015) Subtype-specific plasticity of inhibitory circuits in motor cortex during motor learning. *Nat. Neurosci.* 18, 1109–1115
147. Song, S. *et al.* (2022) Input-specific inhibitory plasticity improves decision accuracy under noise. *bioRxiv* Published online May 25, 2022. https://doi.org/10.1101/2022.05.24.493332
148. Freund, T.F. and Katona, I. (2007) Perisomatic inhibition. *Neuron* 56, 33–42
149. Abs, E. *et al.* (2018) Learning-related plasticity in dendrite-targeting layer 1 interneurons. *Neuron* 100, 684–699
150. Loomba, S. *et al.* (2022) Connectomic comparison of mouse and human cortex. *Science* 377, eabo0924
151. Kim, R. and Sejnowski, T.J. (2021) Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. *Nat. Neurosci.* 24, 129–139
152. Mongillo, G. *et al.* (2018) Inhibitory connectivity defines the realm of excitatory plasticity. *Nat. Neurosci.* 21, 1463–1470
153. Wilmes, K.A. and Clopath, C. (2019) Inhibitory microcircuits for top-down plasticity of sensory rep-resentations. *Nat. Commun.* 10, 5055
154. Confavreux, B. *et al.* (2020) A meta-learning approach to (re) discover plasticity rules that carve a desired function into a neural network. *bioRxiv* Published online October 25, 2020. https://doi.org/10.1101/2020.10.24.353409
155. Keijser, J. and Sprekeler, H. (2022) Optimizing interneuron circuits for compartment-specific feedback inhibition. *PLoS Comput. Biol.* 18, 1–21
156. Soldado-Magraner, S. *et al.* (2021) Orchestrated excitatory and inhibitory learning rules lead to the unsupervised emergence of self-sustained and inhibition-stabilized dynamics. *bioRxiv* Published online September 13, 2021. https://doi.org/10.1101/2020.12.30.424888
157. Pfeffer, C.K. *et al.* (2013) Inhibition of inhibition in visual cortex the logic of connections between molecularly distinct interneurons. *Nat. Neurosci.* 16, 1068–1076
158. Jiang, X. *et al.* (2015) Principles of connectivity among morphologically defined cell types in adult neocortex. *Science* 350, aac9462
159. Hartung, J. and Letzkus, J.J. (2021) Inhibitory plasticity in layer 1 – dynamic gatekeeper of neocortical associations. *Curr. Opin. Neurobiol.* 67, 26–33
160. Fu, Y. *et al.* (2014) A cortical circuit for gain control by behavioral state. *Cell* 156, 1139–1152
161. Attinger, A. *et al.* (2017) Visuomotor coupling shapes the functional development of mouse visual cortex. *Cell* 169, 1291–1302.e14
162. Campagnola, L. *et al.* (2022) Local connectivity and synaptic dynamics in mouse and human neocortex. *Science* 375, eabj5861

# IV. Nonlinear transient amplification in recurrent neural networks with short-term plasticity

Wu, Y. K. & Zenke, F. Nonlinear transient amplification in recurrent neural networks with short-term plasticity. *eLife* **10**, e71263 (2021). https://doi.org/10.7554/eLife.71263

RESEARCH ARTICLE

# Nonlinear transient amplification in recurrent neural networks with short-term plasticity

**Yue Kris Wu**[1,2,3,4], **Friedemann Zenke**[1,2]*

[1]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland; [2]Faculty of Natural Sciences, University of Basel, Basel, Switzerland; [3]Max Planck Institute for Brain Research, Frankfurt, Germany; [4]School of Life Sciences, Technical University of Munich, Freising, Germany

**Abstract** To rapidly process information, neural circuits have to amplify specific activity patterns transiently. How the brain performs this nonlinear operation remains elusive. Hebbian assemblies are one possibility whereby strong recurrent excitatory connections boost neuronal activity. However, such Hebbian amplification is often associated with dynamical slowing of network dynamics, non-transient attractor states, and pathological run-away activity. Feedback inhibition can alleviate these effects but typically linearizes responses and reduces amplification gain. Here, we study nonlinear transient amplification (NTA), a plausible alternative mechanism that reconciles strong recurrent excitation with rapid amplification while avoiding the above issues. NTA has two distinct temporal phases. Initially, positive feedback excitation selectively amplifies inputs that exceed a critical threshold. Subsequently, short-term plasticity quenches the run-away dynamics into an inhibition-stabilized network state. By characterizing NTA in supralinear network models, we establish that the resulting onset transients are stimulus selective and well-suited for speedy information processing. Further, we find that excitatory-inhibitory co-tuning widens the parameter regime in which NTA is possible in the absence of persistent activity. In summary, NTA provides a parsimonious explanation for how excitatory-inhibitory co-tuning and short-term plasticity collaborate in recurrent networks to achieve transient amplification.

***For correspondence:** friedemann.zenke@fmi.ch

**Competing interest:** The authors declare that no competing interests exist.

**Funding:** See page 30

**Preprinted:** 10 June 2021
**Received:** 14 June 2021
**Accepted:** 10 December 2021
**Published:** 13 December 2021

**Reviewing Editor:** Timothy O'Leary, University of Cambridge, United Kingdom



## Editor's evaluation

Many brain circuits, particularly those found in mammalian sensory cortices, need to respond rapidly to stimuli while at the same time avoiding pathological, runaway excitation. Over several years, many theoretical studies have attempted to explain how cortical circuits achieve these goals through interactions between inhibitory and excitatory cells. This study adds to this literature by showing how synaptic short-term depression can stabilise strong positive feedback in a circuit under a variety of plausible scenarios, allowing strong, rapid and stimulus-specific responses.

## Introduction

Perception in the brain is reliable and strikingly fast. Recognizing a familiar face or locating an animal in a picture only takes a split second (***Thorpe et al., 1996***). This pace of processing is truly remarkable since it involves several recurrently connected brain areas each of which has to selectively amplify or suppress specific signals before propagating them further. This processing is mediated through circuits with several intriguing properties. First, excitatory-inhibitory (EI) currents into individual neurons are commonly correlated in time and co-tuned in stimulus space (***Wehr and Zador, 2003***; ***Froemke et al.,***

*2007*; *Okun and Lampl, 2008*; *Hennequin et al., 2017*; *Rupprecht and Friedrich, 2018*; *Znamenskiy et al., 2018*). Second, neural responses to stimulation are shaped through diverse forms of short-term plasticity (STP) (*Tsodyks and Markram, 1997*; *Markram et al., 1998*; *Zucker and Regehr, 2002*; *Pala and Petersen, 2015*). Finally, mounting evidence suggests that amplification rests on neuronal ensembles with strong recurrent excitation (*Marshel et al., 2019*; *Peron et al., 2020*), whereby excitatory neurons with similar tuning preferentially form reciprocal connections (*Ko et al., 2011*; *Cossell et al., 2015*). Such predominantly symmetric connectivity between excitatory cells is consistent with the notion of Hebbian cell assemblies (*Hebb, 1949*), which are considered an essential component of neural circuits and the putative basis of associative memory (*Harris, 2005*; *Josselyn and Tonegawa, 2020*). Computationally, Hebbian cell assemblies can amplify specific activity patterns through positive feedback, also referred to as Hebbian amplification. Based on these principles, several studies have shown that Hebbian amplification can drive persistent activity that outlasts a preceding stimulus (*Hopfield, 1982*; *Amit and Brunel, 1997*; *Yakovlev et al., 1998*; *Wong and Wang, 2006*; *Zenke et al., 2015*; *Gillary et al., 2017*), comparable to selective delay activity observed in the prefrontal cortex when animals are engaged in working memory tasks (*Funahashi et al., 1989*; *Romo et al., 1999*).

However, in most brain areas, evoked responses are transient and sensory neurons typically exhibit pronounced stimulus *onset* responses, after which the circuit dynamics settle into a low-activity steady-state even when the stimulus is still present (*DeWeese et al., 2003*; *Mazor and Laurent, 2005*; *Bolding and Franks, 2018*). Preventing run-away excitation and multi-stable attractor dynamics in recurrent networks requires powerful and often finely tuned feedback inhibition resulting in EI balance (*Amit and Brunel, 1997*; *Compte et al., 2000*; *Litwin-Kumar and Doiron, 2012*; *Ponce-Alvarez et al., 2013*; *Mazzucato et al., 2019*), However, strong feedback inhibition tends to linearize steady-state activity (*van Vreeswijk and Sompolinsky, 1996*; *Baker et al., 2020*). *Murphy and Miller, 2009* proposed *balanced amplification* which reconciles transient amplification with strong recurrent excitation by tightly balancing recurrent excitation with strong feedback inhibition (*Goldman, 2009*; *Hennequin et al., 2012*; *Hennequin et al., 2014*; *Bondanelli and Ostojic, 2020*; *Gillett et al., 2020*). Importantly, balanced amplification was formulated for linear network models of excitatory and inhibitory neurons. Due to linearity, it intrinsically lacks the ability to nonlinearly amplify stimuli which limits its capabilities for pattern completion and pattern separation. Further, how balanced amplification relates to nonlinear neuronal activation functions and nonlinear synaptic transmission as, for instance, mediated by STP (*Tsodyks and Markram, 1997*; *Markram et al., 1998*; *Zucker and Regehr, 2002*; *Pala and Petersen, 2015*), remains elusive. This begs the question of whether there are alternative nonlinear amplification mechanisms and how they relate to existing theories of recurrent neural network processing.

Here, we address this question by studying an alternative mechanism for the emergence of transient dynamics that relies on recurrent excitation, supralinear neuronal activation functions, and STP. Specifically, we build on the notion of ensemble synchronization in recurrent networks with STP (*Loebel and Tsodyks, 2002*; *Loebel et al., 2007*) and study this phenomenon in analytically tractable network models with rectified quadratic activation functions (*Ahmadian et al., 2013*; *Rubin et al., 2015*; *Hennequin et al., 2018*; *Kraynyukova and Tchumatchenko, 2018*) and STP. We first characterize the conditions under which individual neuronal ensembles with supralinear activation functions and recurrent excitatory connectivity succumb to explosive run-away activity in response to external stimulation. We then show how STP effectively mitigates this instability by re-stabilizing ensemble dynamics in an inhibition-stabilized network (ISN) state, but only after generating a pronounced stimulus-triggered onset transient. We call this mechanism NTA and show that it yields selective onset responses that carry more relevant stimulus information than the subsequent steady-state. Finally, we characterize the functional benefits of inhibitory co-tuning, a feature that is widely observed in the brain (*Wehr and Zador, 2003*; *Froemke et al., 2007*; *Okun and Lampl, 2008*; *Rupprecht and Friedrich, 2018*) and readily emerges in computational models endowed with activity-dependent plasticity of inhibitory synapses (*Vogels et al., 2011*). We find that co-tuning prevents persistent attractor states but does not preclude NTA from occurring. Importantly, NTA purports that, following transient amplification, neuronal ensembles settle into a stable ISN state, consistent with recent studies suggesting that inhibition stabilization is a ubiquitous feature of cortical networks (*Sanzeni et al., 2020*). In summary, our work indicates that NTA is ideally suited to amplify stimuli rapidly through the interaction of strong recurrent excitation with STP.

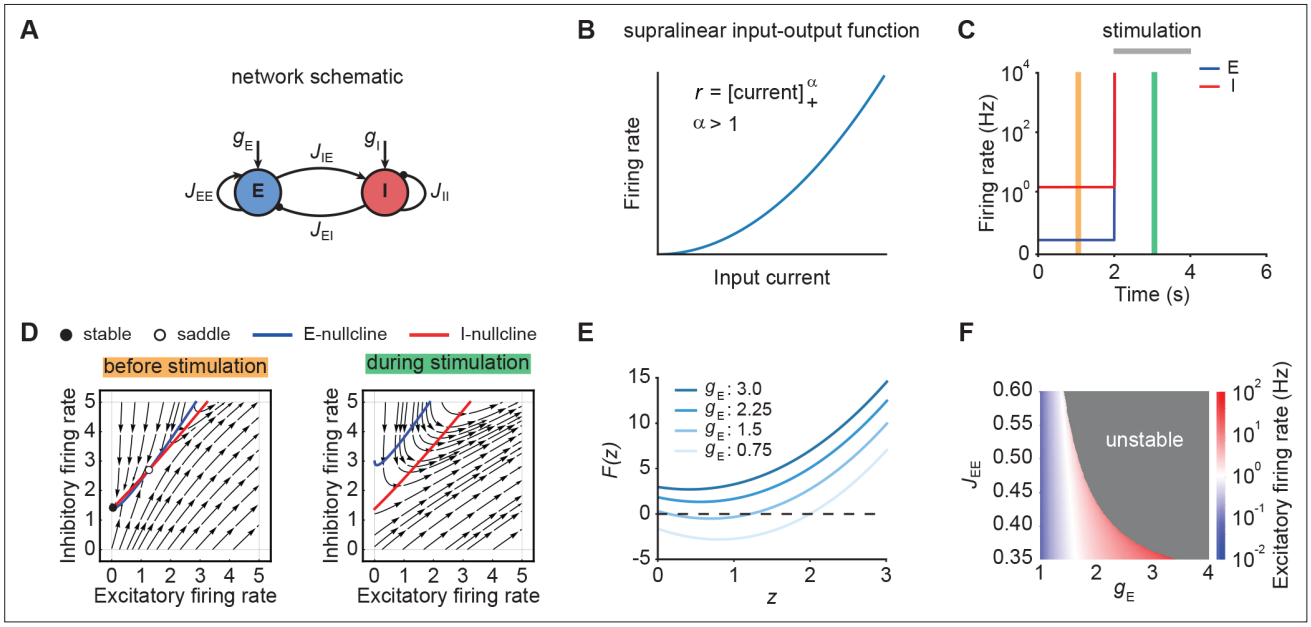


**Figure 1.** Neuronal ensembles nonlinearly amplify inputs above a critical threshold. (**A**) Schematic of the recurrent ensemble model consisting of an excitatory (blue) and an inhibitory population (red). (**B**) Supralinear input-output function given by a rectified power law with exponent $\alpha = 2$. (**C**) Firing rates of the excitatory (blue) and inhibitory population (red) in response to external stimulation during the interval from 2 to 4 s (gray bar). The stimulation was implemented by temporarily increasing the input $g_E$. (**D**) Phase portrait of the system before stimulation (left; C orange) and during stimulation (right; C green). (**E**) Characteristic function $F(z)$ for varying input strength $g_E$. Note that the function loses its zero crossings, which correspond to fixed points of the system for increasing external input. (**F**) Heat map showing the evoked firing rate of the excitatory population for different parameter combinations $J_{EE}$ and $g_E$. The gray region corresponds to the parameter regime with unstable dynamics.

The online version of this article includes the following figure supplement(s) for figure 1:

**Figure supplement 1.** Unstable ensemble dynamics can be triggered by additional stimulation in supralinear networks with negative determinant even in the presence of substantial feedforward inhibition.

## Results

To understand the emergence of transient responses in recurrent neural networks, we studied rate-based population models with a supralinear, power law input-output function (***Figure 1A and B***; ***Ahmadian et al., 2013***; ***Hennequin et al., 2018***), which captures essential aspects of neuronal activation (***Priebe et al., 2004***), while also being analytically tractable. We first considered an isolated neuronal ensemble consisting of one excitatory (E) and one inhibitory (I) population (***Figure 1A***).

The dynamics of this network are given by

$$\tau_E \frac{dr_E}{dt} = -r_E + \left[J_{EE} r_E - J_{EI} r_I + g_E\right]_+^{\alpha_E} , \tag{1}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[J_{IE} r_E - J_{II} r_I + g_I\right]_+^{\alpha_I} , \tag{2}$$

where $r_E$ and $r_I$ are the firing rates of the excitatory and inhibitory population, $\tau_E$ and $\tau_I$ represent the corresponding time constants, $J_{XY}$ denotes the synaptic strength from the population $Y$ to the population $X$, where $X, Y \in \{E, I\}$, $g_E$ and $g_I$ are the external inputs to the respective populations. Finally, $\alpha_E$ and $\alpha_I$, the exponents of the respective input-output functions, are fixed at two unless mentioned otherwise. For ease of notation, we further define the weight matrix $\mathbf{J}$ of the compound system as follows:

$$\mathbf{J} = \begin{bmatrix} J_{EE} & -J_{EI} \\ J_{IE} & -J_{II} \end{bmatrix} . \tag{3}$$

We were specifically interested in networks with strong recurrent excitation that can generate positive feedback dynamics in response to external inputs $g_E$. Therefore, we studied networks with

$$\det(\mathbf{J}) = -J_{EE}J_{II} + J_{IE}J_{EI} < 0 \quad . \tag{4}$$

In contrast, networks in which recurrent excitation is met by strong feedback inhibition such that $\det(\mathbf{J}) > 0$ are unable to generate positive feedback dynamics provided that inhibition is fast enough (***Ahmadian et al., 2013***). Importantly, we assumed that most inhibition originates from recurrent connections (***Franks et al., 2011***; ***Large et al., 2016***) and, hence, we kept the input to the inhibitory population $g_I$ fixed unless mentioned otherwise.

## Nonlinear amplification of inputs above a critical threshold

We initialized the network in a stable low-activity state in the absence of external stimulation, consistent with spontaneous activity in cortical networks (***Figure 1C***). However, an input $g_E$ of sufficient strength, destabilized the network (***Figure 1C***). Importantly, this behavior is distinct from linear network models in which the network stability is independent of inputs (Materials and methods). The transition from stable to unstable dynamics can be understood by examining the phase portrait of the system (***Figure 1D***). Before stimulation, the system has a stable and an unstable fixed point (***Figure 1D***, left). However, both fixed points disappear for an input $g_E$ above a critical stimulus strength (***Figure 1D***, right).

To further understand the system's bifurcation structure, we consider the characteristic function

$$F(z) = J_{EE}\left[z\right]_+^{\alpha_E} - J_{EI}\left[\det(\mathbf{J}) \cdot J_{EI}^{-1}\left[z\right]_+^{\alpha_E} + J_{EI}^{-1}J_{II}z - J_{EI}^{-1}J_{II}g_E + g_I\right]_+^{\alpha_I} - z + g_E \quad , \tag{5}$$

where $z$ denotes the total current into the excitatory population and $\det(\mathbf{J})$ represents the determinant of the weight matrix (***Kraynyukova and Tchumatchenko, 2018***; Materials and methods). The characteristic function reduces the original two-dimensional system to one dimension, whereby the zero crossings of the characteristic function correspond to the fixed points of the original system (***Eq. (1)-(2)***). We use this correspondence to visualize how the fixed points of the system change with the input $g_E$. Increasing $g_E$ shifts $F(z)$ upwards, which eventually leads to all zero crossings disappearing and the ensuing unstable dynamics (***Figure 1E***; Materials and methods). Importantly, for any weight matrix $\mathbf{J}$ with negative determinant, there exists a critical input $g_E$ at which all fixed points disappear (Materials and methods). While for weak recurrent E-to-E connection strength $J_{EE}$, the transition from stable dynamics to unstable is gradual, in that it happens at higher firing rates (***Figure 1F***), it becomes more abrupt for stronger $J_{EE}$. Thus, our analysis demonstrates that individual neuronal ensembles with negative determinant $\det(\mathbf{J})$ nonlinearly amplify inputs above a critical threshold by switching from initially stable to unstable dynamics.

## Short-term plasticity, but not spike-frequency adaptation, can re-stabilize ensemble dynamics

Since unstable dynamics are not observed in neurobiology, we wondered whether neuronal spike frequency adaptation (SFA) or STP could re-stabilize the ensemble dynamics while keeping the nonlinear amplification character of the system. Specifically, we considered SFA of excitatory neurons, E-to-E short-term depression (STD), and E-to-I short-term facilitation (STF). We focused on these particular mechanisms because they are ubiquitously observed in the brain. Most pyramidal cells exhibit SFA (***Barkai and Hasselmo, 1994***) and most synapses show some form of STP (***Markram et al., 1998***; ***Zucker and Regehr, 2002***; ***Pala and Petersen, 2015***). Moreover, the time scales of these mechanisms are well-matched to typical timescales of perception, ranging from milliseconds to seconds (***Tsodyks and Markram, 1997***; ***Fairhall et al., 2001***; ***Pozzorini et al., 2013***).

When we simulated our model with SFA (***Eqs. (21)–(23)***), we observed different network behaviors depending on the adaptation strength. When adaptation strength was weak, SFA was unable to stabilize run-away excitation (***Figure 2A***; Materials and methods). Increasing the adaptation strength eventually prevented run-away excitation, but to give way to oscillatory ensemble activity (***Figure 2—figure supplement 1***). Finally, we confirmed analytically that SFA cannot stabilize excitatory run-away dynamics at a stable fixed point (Materials and methods). In particular, while the input is present, strong SFA creates a stable limit cycle with associated oscillatory ensemble activity (***Figure 2—figure***

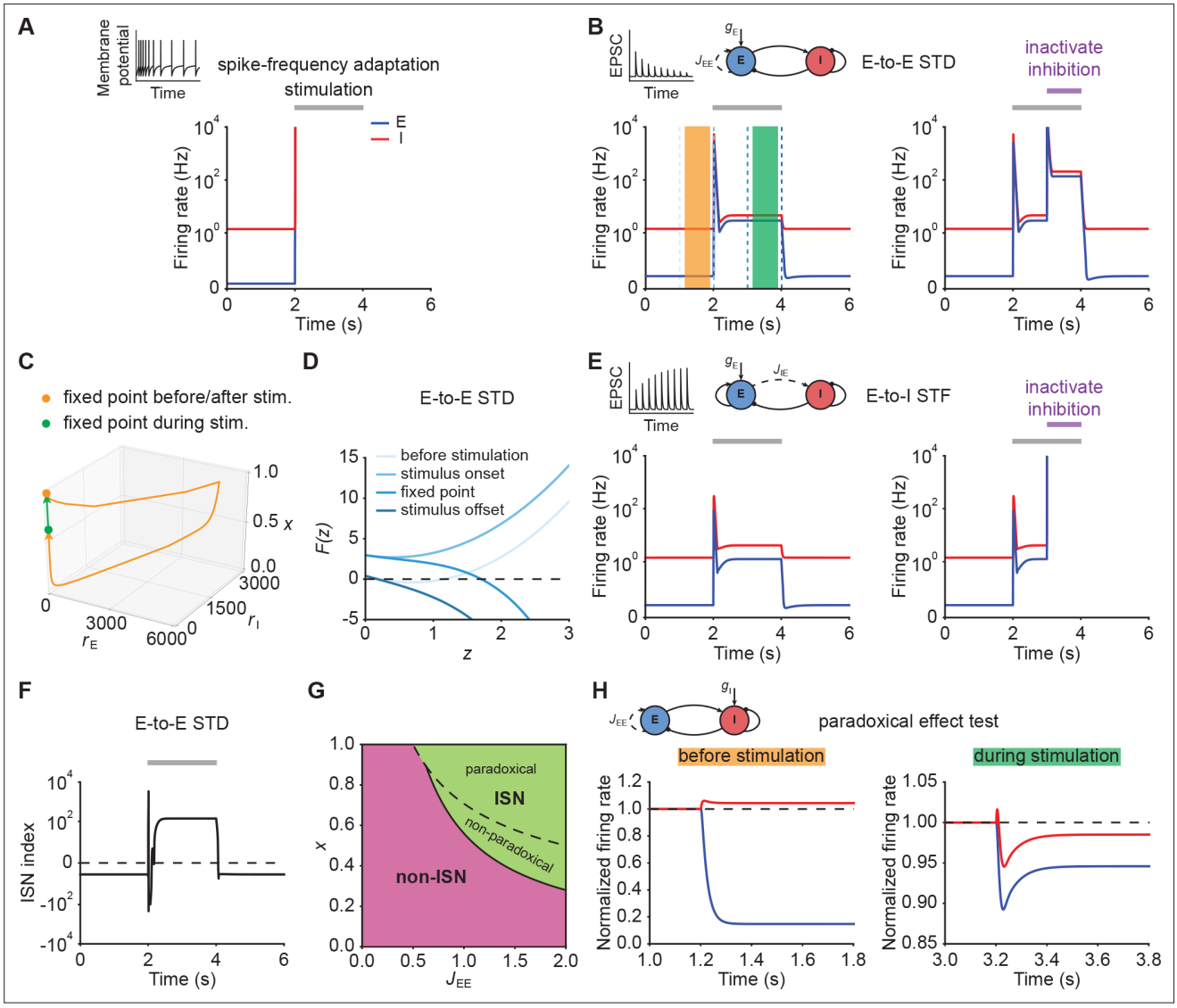


**Figure 2.** Short-term plasticity, but not spike-frequency adaptation, re-stabilizes ensemble dynamics. (**A**) Firing rates of the excitatory (blue) and inhibitory population (red) in the presence of spike-frequency adaptation (SFA). During stimulation (gray bar) additional input is injected into the excitatory population. The inset shows a cartoon of how SFA affects spiking neuronal dynamics in response to a step current input. (**B**) Left: Same as (**A**) but in the presence of E-to-E short-term depression (STD). Right: Same as left but inactivating inhibition in the period marked in purple. (**C**) 3D plot of the excitatory activity $r_E$, inhibitory activity $r_I$, and the STD variable $x$ of the network in B left. The orange and green points mark the fixed points before/after and during stimulation. (**D**) Characteristic function $F(z)$ in networks with E-to-E STD. Different brightness levels correspond to different time points in B left. (**E**) Same as (**B**) but in the presence of E-to-I short-term facilitation (STF). (**F**) Inhibition-stabilized network (ISN) index, which corresponds to the largest real part of the eigenvalues of the Jacobian matrix of the E-E subnetwork with STD, as a function of time for the network with E-to-E STD in B left. For values above zero (dashed line), the ensemble is an ISN. (**G**) Analytical solution of non-ISN (magenta), ISN (green), paradoxical, and non-paradoxical regions for different parameter combinations $J_{EE}$ and the STD variable $x$. The solid line separates the non-ISN and ISN regions, whereas the dashed line separates the non-paradoxical and paradoxical regions. (**H**) The normalized firing rates of the excitatory (blue) and inhibitory population (red) when injecting additional excitatory current into the inhibitory population before stimulation (left; orange bar in B), and during stimulation (right; green bar in B). Initially, the ensemble is in the non-ISN regime and injecting excitatory current into the inhibitory population increases its firing rate. During stimulation, however, the ensemble is an ISN. In this case, excitatory current injection into the inhibitory population results in a reduction of its firing rate, also known as the *paradoxical effect*.

The online version of this article includes the following figure supplement(s) for figure 2:

**Figure supplement 1.** Ensemble dynamics in supralinear networks with strong SFA.

**Figure supplement 2.** Dependence of peak amplitude and fixed point activity on input $g_E$ and E-to-E connection strength $J_{EE}$.

*Figure 2 continued on next page*

*Figure 2 continued*

**Figure supplement 3.** Comparisons of amplification ability between NTA and linear networks, and between NTA and SSNs.

**Figure supplement 4.** Dependence of peak amplitude and fixed point activity on STP parameters.

**Figure supplement 5.** Networks initially in the ISN regime can exhibit strong NTA.

**Figure supplement 6.** ISN index and paradoxical effect test for networks with E-to-I STF.

**Figure supplement 7.** Inhibition stabilization does not imply paradoxical response in networks with E-to-E STD.

**Figure supplement 8.** Transition from non-ISN to ISN indicating by frozen inhibition test.

**Figure supplement 9.** Similar qualitative behavior in rate-based models with maximal firing rate capped at 300 Hz.

**Figure supplement 10.** Similar qualitative behavior in spiking neural networks.

**Figure supplement 11.** Unstable dynamics can emerge in supralinear networks with positive determinant and slow inhibition.

**Figure supplement 12.** Networks with substantial feedforward inhibition can exhibit strong NTA.

***supplement 1***; Materials and methods), which was also shown in previous modeling studies (***van Vreeswijk and Hansel, 2001***), but is not typically observed in sensory systems (***DeWeese et al., 2003***; ***Rupprecht and Friedrich, 2018***).

Next, we considered STP, which is capable of saturating the effective neuronal input-output function (***Mongillo et al., 2012***; ***Zenke et al., 2015***; ***Eqs. (37)–(39)***, ***Eqs. (41)–(43)***). We first analyzed the stimulus-evoked network dynamics when we added STD to the recurrent E-to-E connections. Strong depression of synaptic efficacy resulted in a brief onset transient after which the ensemble dynamics quickly settled into a stimulus-evoked steady-state with slightly higher activity than the baseline (***Figure 2B***, left). After stimulus removal, the ensemble activity returned back to its baseline level (***Figure 2B***, left; ***Figure 2C***). Notably, the ensemble dynamics settled at a stable steady state with a much higher firing rate, when inhibition was inactivated during stimulus presentation (***Figure 2B***, right). This shows that STP is capable of creating a stable high-activity fixed point, which is fundamentally different from the SFA dynamics discussed above. This difference in ensemble dynamics can be readily understood by analyzing the stability of the three-dimensional dynamical system (Materials and methods). We can gain a more intuitive understanding by considering self-consistent solutions of the characteristic function $F(z)$. Initially, the ensemble is at the stable low activity fixed point. But the stimulus causes this fixed point to disappear, thus giving way to positive feedback which creates the leading edge of the onset transient (***Figure 2B***). However, because E-to-E synaptic transmission is rapidly reduced by STD, the curvature of $F(z)$ changes and a stable fixed point is created, thereby allowing excitatory run-away dynamics to terminate and the ensemble dynamics settle into a steady-state at low activity levels (***Figure 2D***). We found that E-to-I STF leads to similar dynamics (***Figure 2E***, left; Appendix 1) with the only difference that this configuration requires inhibition for network stability (***Figure 2E***, right), whereas E-to-E STD stabilizes activity even without inhibition, albeit at physiologically implausibly high activity levels. Importantly, the re-stabilization through either form of STP did not impair an ensemble's ability to amplify stimuli during the initial onset phase.

Crucially, transient amplification in supralinear networks with STP occurs above a critical threshold (***Figure 2—figure supplement 2***), and requires recurrent excitation $J_{EE}$ to be sufficiently strong (***Figure 2—figure supplement 2C, D***). To quantify the amplification ability of these networks, we calculated the ratio of the evoked peak firing rate to the input strength, henceforth called the 'Amplification index'. We found that amplification in STP-stabilized supralinear networks can be orders of magnitude larger than in linear networks with equivalent weights and comparable stabilized supralinear networks (SSNs) without STP (***Figure 2—figure supplement 3***). We stress that the resulting firing rates are parameter-dependent (***Figure 2—figure supplement 4***) and their absolute value can be high due to the high temporal precision of the onset peak and its short duration. In experiments, such high rates manifest themselves as precisely time-locked spikes with millisecond resolution (***DeWeese et al., 2003***; ***Wehr and Zador, 2003***; ***Bolding and Franks, 2018***; ***Gjoni et al., 2018***).

Recent studies suggest that cortical networks operate as inhibition-stabilized networks (ISNs) (***Sanzeni et al., 2020***; ***Sadeh and Clopath, 2021***), in which the excitatory network is unstable in the absence of feedback inhibition (***Tsodyks et al., 1997***; ***Ozeki et al., 2009***). To that end, we investigated how ensemble re-stabilization relates to the network operating regime at baseline and during stimulation. Whether a network is an ISN or not is mathematically determined by the real part of the

leading eigenvalue of the Jacobian of the excitatory-to-excitatory subnetwork (***Tsodyks et al., 1997***). We computed the leading eigenvalue in our model incorporating STP and referred to it as 'ISN index' (Materials and methods; Appendix 2). We found that in networks with STP the ISN index can switch sign from negative to positive during external stimulation, indicating that the ensemble can transition from a non-ISN to an ISN (***Figure 2F***). Notably, this behavior is distinct from linear network models in which the network operating regime is independent of the input (Materials and methods). Whether this switch between non-ISN to ISN occurred, however, was parameter dependent and we also found network configurations that were already in the ISN regime at baseline and remained ISNs during stimulation (***Figure 2—figure supplement 5***). Thus, re-stabilization was largely unaffected by the network state and consistent with experimentally observed ISN states (***Sanzeni et al., 2020***).

Theoretical studies have shown that one defining characteristic of ISNs in static excitatory and inhibitory networks is that injecting excitatory (inhibitory) current into inhibitory neurons decreases (increases) inhibitory firing rates, which is also known as the paradoxical effect (***Tsodyks et al., 1997***; ***Miller and Palmigiano, 2020***). Yet, it is unclear whether in networks with STP, inhibitory stabilization implies paradoxical response and vice versa. We therefore analyzed the condition of being an ISN and the condition of having paradoxical response in networks with STP (Materials and methods; Appendix 2; Appendix 3). Interestingly, we found that in networks with E-to-E STD, the paradoxical effect implies inhibitory stabilization, whereas inhibitory stabilization does not necessarily imply paradoxical response (***Figure 2G***; Materials and methods), suggesting that having paradoxical effect is a sufficient but not necessary condition for being an ISN. In contrast, in networks with E-to-I STF, inhibitory stabilization and paradoxical effect imply each other (Appendix 2; Appendix 3). Therefore, paradoxical effect can be exploited as a proxy for inhibition stabilization for networks with STP we considered here. By injecting excitatory current into the inhibitory population, we found that the network did not exhibit the paradoxical effect before stimulation (***Figure 2H***, left; ***Figure 2—figure supplement 6***). In contrast, injecting excitatory inputs into the inhibitory population during stimulation reduced their activity (***Figure 2H***, right; ***Figure 2—figure supplement 6***). As demonstrated in our analysis, non-paradoxical response does not imply non-ISN (***Figure 2—figure supplement 7***; Materials and methods). We therefore examined the inhibition stabilization property of the ensemble by probing the ensemble behavior when a small transient perturbation to excitatory population activity is introduced while inhibition is frozen before stimulation and during stimulation. Before stimulation, the firing rate of the excitatory population slightly increases and then returns to its baseline after the transient perturbation (***Figure 2—figure supplement 8***). During stimulation, however, the transient perturbation leads to a transient explosion of the excitatory firing rate (***Figure 2—figure supplement 8***). These results further confirm that the ensemble shown in our example is initially a non-ISN before stimulation and can transition to an ISN with stimulation. By elevating the input level at the baseline in the model, the ensemble can be initially an ISN (***Figure 2—figure supplement 5***), resembling recent studies revealing that cortical circuits in the mouse V1 operate as ISNs in the absence of sensory stimulation (***Sanzeni et al., 2020***).

Despite the fact that the supralinear input-output function of our framework captures some aspects of intracellular recordings (***Priebe et al., 2004***), it is unbounded and thus allows infinitely high firing rates. This is in contrast to neurobiology where firing rates are bounded due to neuronal refractory effects. While this assumption permitted us to analytically study the system and therefore to gain a deeper understanding of the underlying ensemble dynamics, we wondered whether our main conclusions were also valid when we limited the maximum firing rates. To that end, we carried out the same simulations while capping the firing rate at 300 Hz. In the absence of additional SFA or STP mechanisms, the firing rate saturation introduced a stable high-activity state in the ensemble dynamics which replaced the unstable dynamics in the uncapped model. As above, the ensemble entered this high-activity steady-state when stimulated with an external input above a critical threshold and exhibited persistent activity after stimulus removal (***Figure 2—figure supplement 9***). While weak SFA did not change this behavior, strong SFA resulted in oscillatory behavior during stimulation consistent with previous analytical work (***Figure 2—figure supplement 9***, ***van Vreeswijk and Hansel, 2001***), but did not in stable steady-states commonly observed in biological circuits. In the presence of E-to-E STD or E-to-I STF, however, the ensemble exhibited transient evoked activity at stimulation onset that was comparable to the uncapped case. Importantly, the ensemble did not show persistent activity after the stimulation (***Figure 2—figure supplement 9***). Finally, we confirmed that all of these findings were

qualitatively similar in a realistic spiking neural network model (***Figure 2—figure supplement 10***; Materials and methods).

In summary, we found that neuronal ensembles can rapidly, nonlinearly, and transiently amplify inputs by briefly switching from stable to unstable dynamics before being re-stabilized through STP mechanisms. We call this mechanism nonlinear transient amplification (NTA) which, in contrast to balanced amplification (***Murphy and Miller, 2009***; ***Hennequin et al., 2012***), arises from population dynamics with supralinear neuronal activation functions interacting with STP. While we acknowledge that there may be other nonlinear transient amplification mechanisms, in this article we restrict our analysis to the definition above. NTA is characterized by a large onset response, a subsequent ISN steady-state while the stimulus persists, and a return to a unique baseline activity state after the stimulus is removed. Thus, NTA is ideally suited to rapidly and nonlinearly amplify sensory inputs through recurrent excitation, like reported experimentally (***Ko et al., 2011***; ***Cossell et al., 2015***), while avoiding persistent activity.

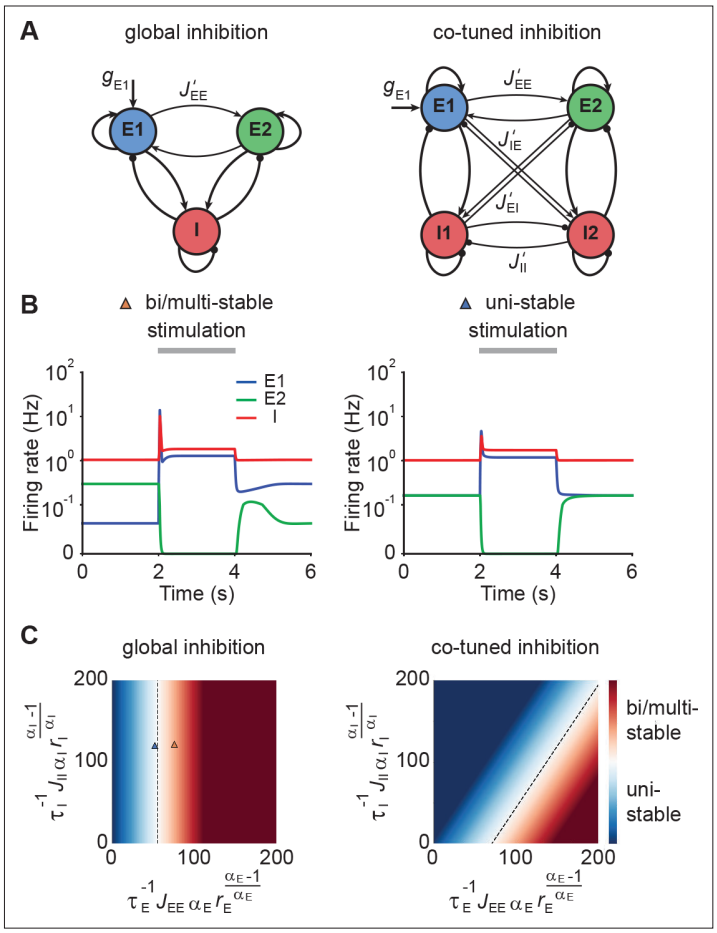


**Figure 3.** Co-tuned inhibition broadens the parameter regime of NTA in the absence of persistent activity. (**A**) Schematic of two neuronal ensembles with global inhibition (left) and with co-tuned inhibition (right). (**B**) Firing rate dynamics of bi/multi-stable ensemble dynamics (left) and uni-stable (right). In both cases, additional excitatory inputs are injected into excitatory ensemble E1 during the period marked in gray. (**C**) Analytical solution of uni- and bi/multi-stability regions for global inhibition (left) and co-tuned inhibition (right). Co-tuning results in a larger parameter regime of uni-stability. The triangles correspond to the two examples in B.

The online version of this article includes the following figure supplement(s) for figure 3:

**Figure supplement 1.** Ensembles with co-tuned inhibition exhibit weaker — but still strong — NTA in comparison to ensembles with global inhibition.

## Co-tuned inhibition broadens the parameter regime of NTA in the absence of persistent activity

Up to now, we have focused on a single neuronal ensemble. However, to process information in the brain, several ensembles with different stimulus selectivity presumably coexist and interact in the same circuit. This coexistence creates potential problems. It can lead to multi-stable persistent attractor dynamics, which are not commonly observed and could have adverse effects on the processing of subsequent stimuli. One solution to this issue could be EI co-tuning, which arises in network models with plastic inhibitory synapses (***Vogels et al., 2011***) and has been observed experimentally in several sensory systems (***Wehr and Zador, 2003***; ***Froemke et al., 2007***; ***Okun and Lampl, 2008***; ***Rupprecht and Friedrich, 2018***).

To characterize the conditions under which neuronal ensembles nonlinearly amplify stimuli without persistent activity, we analyzed the case of two interacting ensembles. More specifically, we considered networks with two excitatory ensembles and distinguished between global and co-tuned inhibition (***Figure 3A***). In the case of global inhibition, one inhibitory population non-specifically inhibits both excitatory populations (***Figure 3A***, left). In contrast, in networks with co-tuned inhibition, each ensemble is formed by a dedicated pair of an excitatory and an inhibitory population which can have cross-over connections, for instance, due to overlapping ensembles (***Figure 3A***, right).

Global inhibition supports winner-take-all competition and is therefore often associated with multi-stable attractor dynamics (***Wong and Wang, 2006***; ***Mongillo et al., 2008***). We first illustrated this effect in a network model with global inhibition. When the recurrent excitatory connections within each ensemble were sufficiently strong, small amounts of noise in the initial condition led to one of the ensembles spontaneously activating at elevated firing rates, while the other ensemble's activity remained low (***Figure 3B***, left). A specific external stimulation could trigger a switch from one state to the other in which the other ensemble was active at a high firing rate. Importantly, this change persisted even after the stimulus had been removed, a hallmark of multi-stable dynamics. In contrast, uni-stable systems have a global symmetric state in which both ensembles have the same activity in the absence of stimulation. While the stimulated ensemble showed elevated firing rates in response to the stimulus, its activity returned to the baseline level after the stimulus is removed (***Figure 3B***, right), consistent with experimental observations (***DeWeese et al., 2003***; ***Rupprecht and Friedrich, 2018***; ***Bolding and Franks, 2018***). Note that the only difference between these two models is that $J_{EE}$ is larger in the multi-stable example than in the uni-stable one.

Symmetric baseline activity is most consistent with activity observed in sensory areas. Hence, we sought to understand which inhibitory connectivity would be most conducive to maintain it. To that end, we analytically identified the uni-stability conditions, which are determined by the leading eigenvalue of the Jacobian matrix of the system, for networks with varying degrees of EI co-tuning (Materials and methods). We found that a broader parameter regime underlies uni-stability in networks with co-tuned inhibition than global inhibition (***Figure 3C***). Notably, this conclusion is general and extends to networks with an arbitrary number of ensembles (Materials and methods). In comparison to the ensemble with global inhibition, the ensemble with co-tuned inhibition exhibits weaker — but still strong — NTA (***Figure 3—figure supplement 1***). Thus, co-tuned inhibition broadens the parameter regime in which NTA is possible while simultaneously avoiding persistent attractor dynamics.

## NTA provides better pattern completion than fixed points while retaining stimulus selectivity

Neural circuits are capable of generating stereotypical activity patterns in response to partial cues and forming distinct representations in response to different stimuli (***Carrillo-Reid et al., 2016***; ***Marshel et al., 2019***; ***Bolding et al., 2020***; ***Vinje and Gallant, 2000***; ***Cayco-Gajic and Silver, 2019***). To test whether NTA achieves pattern completion while retaining stimulus selectivity, we analyzed the transient onset activity in our models and compared it to the fixed point activity.

To investigate pattern completion and stimulus selectivity in our model, we considered a co-tuned network with E-to-E STD and two distinct excitatory ensembles $E1$ and $E2$. We gave additional input $g_{E1}$ to a Subset 1, consisting of 75% of the neurons in ensemble $E1$ (***Figure 4A***). We then measured the evoked activity in the remaining 25% of the excitatory neurons in $E1$ to quantify pattern completion. To assess stimulus selectivity, we injected additional input $g_{E1}$ into the entire $E1$ ensemble during the second stimulation phase (***Figure 4A***) while measuring the activity of $E2$. We found that neurons

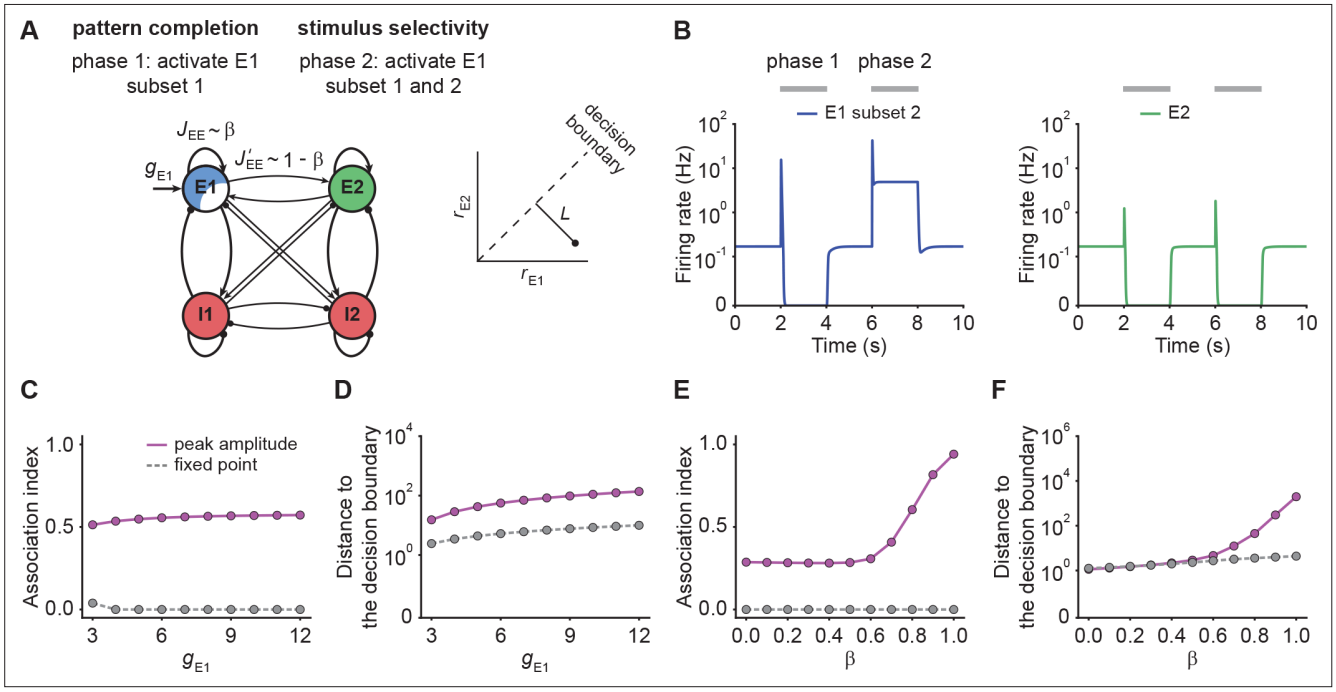


**Figure 4.** NTA yields stronger pattern completion than fixed points while retaining stimulus selectivity. (**A**) Schematic of the network setup used to probe pattern completion and stimulus selectivity. To assess the effect on pattern completion, 75% of the neurons (Subset 1) in ensemble E1 received additional input $g_{E1}$ during Phase one (2–4 s), while we recorded the firing rate of the remaining 25% (Subset 2) in the excitatory ensemble E1. To evaluate the impact on stimulus selectivity, all neurons in E1 received additional inputs $g_{E1}$ in Phase two (6–8 s) while the firing rate of E2 was measured. A downstream neuron's ability to discriminate between E1 or E2 being active depends on whether their activity is well separated by a symmetric decision boundary (inset). (**B**) Examples of firing rates of Subset 2 of E1 (left, blue) and E2 (right, green) with E-to-E STD. (**C**) Association index as a function of input $g_{E1}$ for the onset peak amplitude (magenta solid line) and fixed point activity (gray dashed line) for E-to-E STD. (**D**) Distance to the decision boundary (see panel A, inset) as a function of input $g_{E1}$ for the onset peak amplitude (magenta solid line) and fixed point activity (gray dashed line) for E-to-E STD. (**E and F**) Same as C and D but as a function of β, which controls the inner- and inter-ensemble connection strength.

The online version of this article includes the following figure supplement(s) for figure 4:

**Figure supplement 1.** Change in steady state activity for unstimulated co-tuned neurons in the rate-based model.

**Figure supplement 2.** Quantification of pattern completion and stimulus selectivity in networks with E-to-I STF.

in Subset 2, which did not receive additional input, showed large onset responses, their steady-state activity was largely suppressed (***Figure 4B***). Despite the fact that inputs to $E1$ caused increased transient onset responses in $E2$, the amount of increase was orders of magnitude smaller than in $E1$ (***Figure 4B***). To quantify pattern completion, we defined the

$$\text{Association index} = 1 + \frac{r_{E1_2} - r_{E1_1}}{r_{E1_2} + r_{E1_1}} \quad . \tag{6}$$

Here, $r_{E1_1}$ and $r_{E1_2}$ correspond to the subpopulation activities of $E1$, respectively. By definition, the Association index ranges from zero to one, with larger values indicating stronger associativity. In addition, to quantify the selectivity between $E1$ and $E2$, we considered a symmetric binary classifier (***Figure 4A***, inset) and measured the distance to the decision boundary (Materials and methods). Note that the Association index was computed during Phase one and the distance to the decision boundary during Phase two in this simulation paradigm (***Figure 4B***).

With these definitions, we ran simulations with different input strengths $g_{E1}$. We found that the onset peaks showed stronger association than the fixed point activity (***Figure 4C***). Note that the Association index at the fixed point remained zero, a direct consequence of $r_{E1_2}$ being suppressed to zero. Furthermore, we found that the distance between the transient onset response and the decision boundary was always greater than for the fixed point activity (***Figure 4D***) showing that onset responses retain stimulus selectivity. While the fixed point activity of the unstimulated co-tuned neurons is zero in the given example, stimulating a subset of neurons in one ensemble can lead to an increase in

the fixed point activity of the unstimulated neurons in the same ensemble under certain conditions (***Figure 4—figure supplement 1***; Appendix 4), which is consistent with pattern completion experiments (***Carrillo-Reid et al., 2016***; ***Marshel et al., 2019***) showing that unstimulated neurons from the same ensemble can remain active throughout the whole stimulation period.

To investigate how the recurrent excitatory connectivity affects both pattern completion and stimulus selectivity, we introduced the parameter $\beta$ which controls recurrent excitatory tuning by trading off within-ensemble E-to-E strength $J_{EE}$ relative to the inter-ensemble strength $J'_{EE}$ (***Figure 4A***) such that $J_{EE} = \beta J_{tot}$ and $J'_{EE} = (1-\beta)J_{tot}$. These definitions ensure that the total weight $J_{tot} = J_{EE} + J'_{EE}$ remains constant for any choice of $\beta$. Notably, the overall recurrent excitation strength within an ensemble $J_{EE}$ increases with increasing $\beta$. When $\beta$ is larger than 0.5, the excitatory connection strength within the ensemble $J_{EE}$ exceeds the one between ensembles $J'_{EE}$.

We found that pattern completion ability monotonically increases with $\beta$ with a pronounced onset for $\beta > 0.6$ where NTA takes hold (***Figure 4E***). Moreover, in this regime the two stimulus representations are well separated (***Figure 4F***) which ensures stimulus selectivity also during onset transients. Together, these findings recapitulate the point that recurrent excitatory tuning is a key determinant of network dynamics. Finally, we confirmed that our findings were also valid in networks with E-to-I STF (***Figure 4—figure supplement 2***), which is commonly observed in the brain (***Markram et al., 1998***; ***Zucker and Regehr, 2002***; ***Pala and Petersen, 2015***). In summary, NTA's transient onset responses maintain stimulus selectivity and result in overall better pattern completion than fixed point activity.

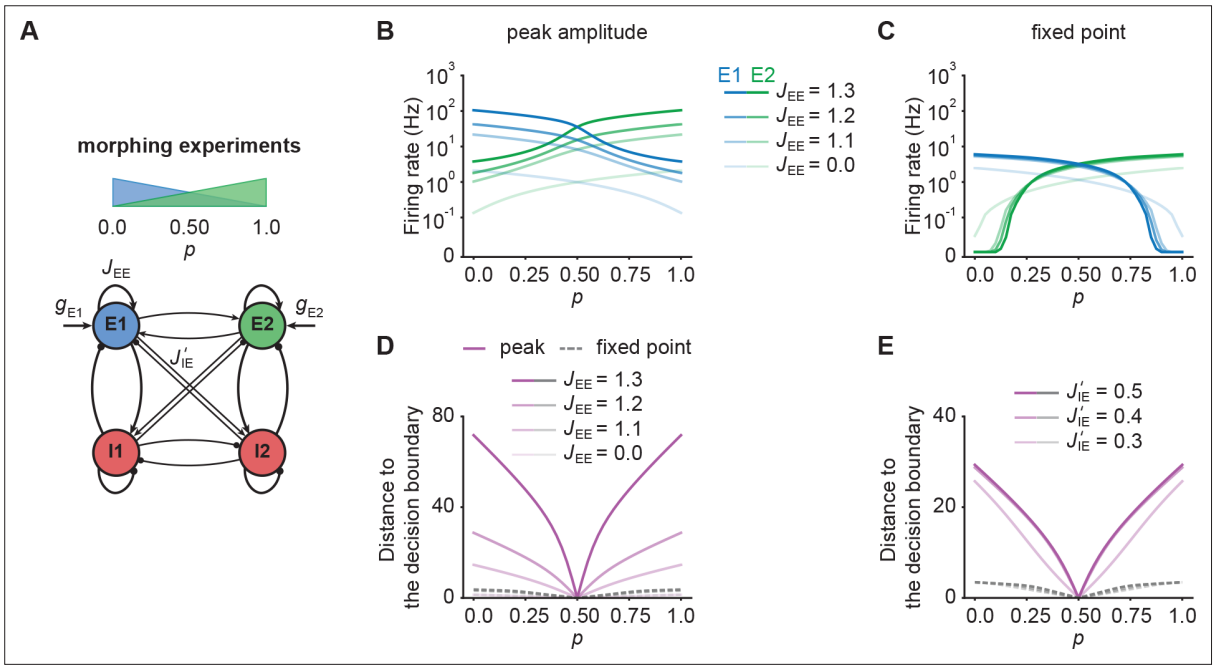


**Figure 5.** NTA provides stronger amplification and pattern separation in morphing experiments than fixed point activity. (**A**) Schematic of the morphing stimulation paradigm. The fraction of the additional inputs into the two excitatory ensembles is controlled by the parameter p. (**B**) Peak amplitude of E1 (blue) and E2 (green) as a function of p for E-to-E STD. Brightness levels represent different recurrent E-to-E connection strengths $J_{EE}$ . (**C**) Same as in B but for fixed point activity. (**D**) Distance to the decision boundary as a function of p for the peak onset response (magenta solid line) and fixed point activity (gray dashed line) for E-to-E STD in a network with $J'_{IE} = 0.4$. (**E**) Same as D but with different E-to-I connection strengths $J'_{IE}$ across ensembles for a network with $J_{EE} = 1.2$.

The online version of this article includes the following figure supplement(s) for figure 5:

**Figure supplement 1.** Quantification of pattern separation in morphing experiments using a normalized measure.

**Figure supplement 2.** Quantification of pattern separation in morphing experiments for networks with E-to-I STF.

## NTA provides higher amplification and pattern separation in morphing experiments

So far, we only considered input to one ensemble. To examine how representations in our model are affected by ambiguous inputs to several ensembles, we performed additional morphing experiments (***Freedman et al., 2001***; ***Niessing and Friedrich, 2010***). To that end, we introduced the parameter $p$ which interpolates between two input stimuli which target $E1$ and $E2$ respectively. When $p$ is zero, all additional input is injected into $E1$. For $p$ equal to one, all additional input is injected into $E2$. Finally, $p$ equal to 0.5 corresponds to the symmetric case in which $E1$ and $E2$ receive the same amount of additional input (***Figure 5A***).

First, we investigated how the recurrent excitatory connection strength within each ensemble $J_{EE}$ affects the onset peak amplitude and fixed point activity. We found that the peak amplitudes depend strongly on $J_{EE}$, whereas the fixed point activity was only weakly dependent on $J_{EE}$ (***Figure 5B and C***). When we disconnected the ensembles by completely eliminating all recurrent excitatory connections, activity was noticeably decreased (***Figure 5B and C***). This illustrates, that recurrent excitation does play an important role in selectively amplifying specific stimuli similar to experimental observations (***Marshel et al., 2019***; ***Peron et al., 2020***), but that amplification is highest at the onset.

Further, we examined the impact of competition through lateral inhibition as a function of the E-to-I inter-ensemble strength $J_{IE}^{'}$ (Materials and methods). As above, we quantified its impact by measuring the representational distance to the decision boundary for the transient onset responses and fixed point activity. We found that regardless of the specific STP mechanism, the distance was larger for the onset responses than for the fixed point activity, consistent with the notion that the onset dynamics separate stimulus identity reliably (***Figure 5D and E***). Since the absolute activity levels between onset and fixed point differed substantially, we further computed the relative pattern Separation index $(r_{E2} - r_{E1})/(r_{E1} + r_{E2})$ and found that the onset transient provides better pattern separation ability for ambiguous stimuli with $p$ close to 0.5 (***Figure 5—figure supplement 1***) provided that the E-to-I connection strength across ensembles $J_{IE}^{'}$ is strong enough. All the while separability for the onset transient was slightly decreased for distinct inputs with $p \in \{0, 1\}$ in comparison to the fixed point. In contrast, fixed points clearly separated such pure stimuli while providing weaker pattern separation for ambiguous input combinations. Importantly, these findings qualitatively held for networks with NTA mediated by E-to-I STF (***Figure 5—figure supplement 2***). Thus, NTA provides stronger amplification and pattern separation than fixed point activity in response to ambiguous stimuli.

## NTA in spiking neural networks

Thus far, our analysis relied on power law neuronal input-output functions in the interest of analytical tractability. To test whether our findings also qualitatively apply to more realistic network models, we built a spiking neural network consisting of randomly connected 800 excitatory and 200 inhibitory neurons, in which the E-to-E synaptic connections were subject to STD (Materials and methods). Here, we defined five overlapping ensembles, each corresponding to 200 randomly selected excitatory neurons. During an initial simulation phase (0–22 s), we consecutively stimulated each ensemble by giving additional input to their excitatory neurons, whereas the input to other neurons remained unchanged (***Figure 6A***). In addition, we also tested pattern completion by stimulating only 75% (Subset 1) of the neurons belonging to Ensemble 5 (22–24 s; ***Figure 6A***). We quantified each ensemble's activity by calculating the population firing rate of the ensemble (Materials and methods). As in the case of the rate-based model, the neuronal ensembles in the spiking model generated pronounced transient onset responses. We then measured the difference of peak ensemble activity and fixed point activity between the stimulated ensemble and the remaining unstimulated ensembles (Materials and methods). As for the rate-based networks, this difference was consistently larger for the onset peak than for the fixed point (***Figure 6B and C***). Thus, transient onset responses allow better stimulus separation than fixed points also in spiking neural network models.

Finally, to visualize the neural activity, we projected the binned spiking activity during the first 10 s of our simulation onto its first two principal components. Notably, the PC trajectory does not exhibit a pronounced rotational component (***Figure 6D***) as activity is confined to one specific ensemble, consistent with experiments (***Marshel et al., 2019***). Furthermore, we computed the fifth ensemble's activity for Subset 1 and 2 during the time interval 16–26 s. In agreement with our rate models, neurons in Subset 2 which did not receive additional inputs showed a strong response at the onset (***Figure 6E***),

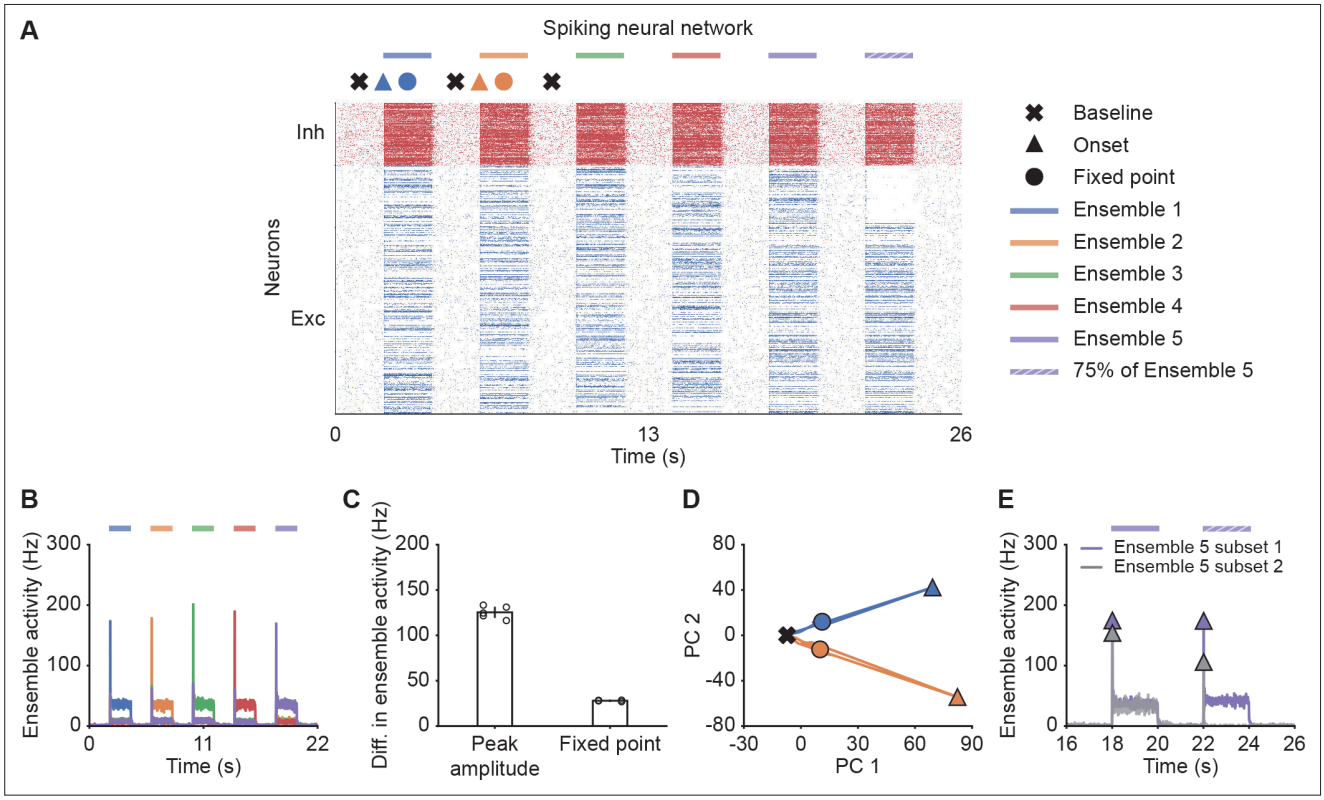


**Figure 6.** Spiking neural network simulations qualitatively reproduce NTA dynamics of rate models. (**A**) Spiking activity of excitatory (blue) and inhibitory (red) neurons in a spiking neural network. From 2 to 20 s, Ensembles 1–5 individually received additional input for 2 s each (colored bars). From 22 to 24 s, 75% of Ensemble 5 neurons (Subset 1) received additional input, whereas the rest 25% of Ensemble 5 neurons (Subset 2) did not receive additional input. The symbols at the top designate the different simulation phases of baseline activity, the onset transients, and the fixed point activity. Different colors correspond to the distinct stimulation periods. (**B**) Ensemble activity (colors). (**C**) Difference in ensemble activity between the stimulated ensemble with the remaining ensembles for the transient onset peak and the fixed point. Points correspond to the different stimulation periods. (**D**) Spiking activity during the interval 0–10 s represented in the PCA basis spanned by the first two principal components which captured approximately 40% of the total variance. The colored lines represent the PC trajectories of the first two stimuli shown in A and B. Triangles, points and crosses correspond to the onset peak, fixed point, and baseline activity, respectively. (**E**) Ensemble activity of Subset 1 (purple) and Subset 2 (gray) of Ensemble 5 from 16 to 26 s. Onset peaks are marked by triangles.

The online version of this article includes the following figure supplement(s) for figure 6:

**Figure supplement 1.** Change in steady state activity for unstimulated co-tuned neurons in spiking neural networks.

but not at the fixed point, suggesting that the strongest pattern completion occurs during the initial amplification phase. Finally, we also observed higher-than-baseline fixed point activity in unstimulated neurons of Subset 2 in spiking neural networks (***Figure 6—figure supplement 1***). Thus, the key characteristics of NTA are preserved across rate-based and more realistic spiking neural network models.

## Discussion

In this study, we demonstrated that neuronal ensemble models with recurrent excitation and suitable forms of STP exhibit nonlinear transient amplification (NTA), a putative mechanism underlying selective amplification in recurrent circuits. NTA combines a supralinear neuronal transfer function, recurrent excitation between neurons with similar tuning, and pronounced STP. Using analytical and numerical methods, we showed that NTA generates rapid transient onset responses during which optimal stimulus separation occurs rather than at steady-states. Additionally, we showed that co-tuned inhibition is conducive to prevent the emergence of persistent activity, which could otherwise interfere with processing subsequent stimuli. In contrast to balanced amplification (***Murphy and Miller, 2009***), NTA is an intrinsically nonlinear mechanism for which only stimuli above a critical threshold are

amplified effectively. While the precise threshold value is parameter-dependent, it can be arbitrarily low provided the excitatory recurrent connections are sufficiently strong (***Figure 1F***). Importantly, such a critical activation threshold offers a possible explanation for sensory perception experiments which show similar threshold behavior (***Marshel et al., 2019***; ***Peron et al., 2020***). Following transient amplification, ensemble dynamics are inhibition-stabilized, which renders our model compatible with existing work on SSNs (***Ahmadian et al., 2013***; ***Rubin et al., 2015***; ***Hennequin et al., 2018***; ***Kraynyukova and Tchumatchenko, 2018***; ***Echeveste et al., 2020***). Thus, NTA provides a parsimonious explanation for why sensory systems may rely upon neuronal ensembles with recurrent excitation in combination with EI co-tuning, and pronounced STP dynamics.

Several theoretical studies approached the problem of transient amplification in recurrent neural network models. ***Loebel and Tsodyks, 2002*** have described an NTA-like mechanism as a driver for powerful ensemble synchronization in rate-based networks and in spiking neural network models of auditory cortex (***Loebel et al., 2007***). Here, we generalized this work to both E-to-E STD and E-to-I STF and provide an in-depth characterization of its amplification capabilities, pattern completion properties, and the resulting network states with regard to their inhibition-stabilization properties. Moreover, we showed that SFA cannot provide similar network stabilization and explored how EI co-tuning interacts with NTA. Finally, we contrasted NTA to alternative transient amplification mechanisms. Balanced amplification is a particularly well-studied transient amplification mechanism (***Murphy and Miller, 2009***; ***Goldman, 2009***; ***Hennequin et al., 2014***; ***Bondanelli and Ostojic, 2020***; ***Gillett et al., 2020***; ***Christodoulou et al., 2021***) that relies on non-normality of the connectivity matrix to selectively and rapidly amplify stimuli. Importantly, balanced amplification occurs in networks in which strong recurrent excitation is appropriately balanced by strong recurrent inhibition. It is capable of generating rich transient activity in linear network models (***Hennequin et al., 2014***), and selectively amplifies specific activity patterns, but without a specific activation threshold. In addition, in spiking neural networks, strong input can induce synchronous firing at the population level which is subsequently stabilized by strong feedback inhibition without the requirement for STP mechanisms (***Stern et al., 2018***). These properties contrast with NTA, which has a nonlinear activation threshold and intrinsically relies on STP to stabilize otherwise unstable run-away dynamics. Due to the switch of the network's dynamical state, NTA's amplification can be orders of magnitudes larger than balanced amplification (***Figure 2—figure supplement 3***). Interestingly, after the transient amplification phase, ensemble dynamics settle in an inhibitory-stabilized state, which renders NTA compatible with previous work on SSNs but in the presence of STP. Finally, although NTA and balanced amplification rely on different amplification mechanisms, they are not mutually exclusive and could, in principle, co-exist in biological networks.

NTA's requirement to generate positive feedback dynamics through recurrent excitation, motivated our focus on networks with $\det(\mathbf{J}) < 0$. As demonstrated in previous work (***Ahmadian et al., 2013***), supralinear networks with $\det(\mathbf{J}) > 0$ and instantaneous inhibition ($\tau_I/\tau_E \to 0$) are always stable for any given input, they are thus unable to generate positive feedback dynamics. In addition, networks with $\det(\mathbf{J}) > 0$ can exhibit a range of interesting behaviors, for example, oscillatory dynamics and persistent activity (***Kraynyukova and Tchumatchenko, 2018***). It is worth noting, however, that for delayed or slow inhibition, stimulation can still lead to unstable network dynamics in networks with $\det(\mathbf{J}) > 0$. Nevertheless, our simulations suggest that our main conclusions about the stabilization mechanisms still hold (***Figure 2—figure supplement 11***).

NTA shares some properties with the notion of network criticality in the brain, like synchronous activation of cell ensembles (***Plenz and Thiagarajan, 2007***) and STP which can tune networks to a critical state (***Levina et al., 2007***). However, in contrast to most models of criticality, in NTA an ensemble briefly transitions to supercritical dynamics in a controlled, stimulus-dependent manner rather than spontaneously. Yet, how the two paradigms are connected at a more fundamental level, is an intriguing question left for future work. Furthermore, recurrent co-tuned inhibition is essential for NTA to ensure uni-stability and selectivity through the suppression of ensembles with different tuning. This requirement is similar in flavor to semi-balanced networks characterized by excess inhibition to some excitatory ensembles while others are balanced (***Baker et al., 2020***). However, the theory of semi-balanced networks has, so far, only been applied to steady-state dynamics while ignoring transients and STP. EI co-tuning prominently features in several models and was shown to support network stability (***Vogels et al., 2011***; ***Hennequin et al., 2017***; ***Znamenskiy et al., 2018***), efficient coding (***Denève and Machens, 2016***), novelty detection (***Schulz et al., 2021***), changes in neuronal variability

(*Hennequin et al., 2018*; *Rost et al., 2018*), and correlation structure (*Wu et al., 2020*). Moreover, some studies have argued that EI balance and co-tuning could increase robustness to noise in the brain (*Rubin et al., 2017*). The present work mainly highlights its importance for preventing multi-stability and delay activity in circuits not requiring such long-timescale dynamics.

NTA is consistent with several experimental findings. First, our model recapitulates the key findings of *Shew et al., 2015* who showed ex vivo that strong sensory inputs cause a transient shift to a supercritical state, after which adaptive changes rapidly tune the network to criticality. Second, NTA requires strong recurrent excitatory connectivity between neurons with similar tuning, which has been reported in experiments (*Ko et al., 2011*; *Cossell et al., 2015*; *Peron et al., 2020*). Third, ensemble activation in our model depends on a critical stimulus strength in line with recent all-optical experiments in the visual cortex, which further link ensemble activation with a perceptual threshold (*Marshel et al., 2019*). Fourth, sensory networks are uni-stable in that they return to a non-selective activity state after the removal of the stimulus and usually do not show persistent activity (*DeWeese et al., 2003*; *Mazor and Laurent, 2005*; *Rupprecht and Friedrich, 2018*). Fifth, our work shows that NTA's onset responses encode stimulus identity better than the fixed point activity, consistent with experiments in the locust antennal lobe (*Mazor and Laurent, 2005*) and research supporting that the brain relies on coactivity on short timescales to represent information (*Stopfer et al., 1997*; *Engel et al., 2001*; *Harris et al., 2003*; *El-Gaby et al., 2021*). Yet, it remains to be seen whether these findings are also coherent with data on the temporal evolution in other sensory systems. Finally, EI co-tuning, which is conducive for NTA, has been found ubiquitously in different sensory circuits (*Wehr and Zador, 2003*; *Froemke et al., 2007*; *Okun and Lampl, 2008*; *Rupprecht and Friedrich, 2018*; *Znamenskiy et al., 2018*).

In our model, we made several simplifying assumptions. For instance, we kept the input to inhibitory neurons fixed and only varied the input to the excitatory population. This step was motivated by experiments in the piriform cortex where the total inhibition is dominated by feedback inhibition (*Franks et al., 2011*). Nevertheless, significant feedforward inhibition was observed in other areas (*Bissière et al., 2003*; *Cruikshank et al., 2007*; *Ji et al., 2016*; *Miska et al., 2018*). While an in-depth comparison for different origins of inhibition was beyond the scope of the present study, we found that increasing the inputs to the excitatory population and inhibitory population by the same amount can still lead to NTA (*Figure 1—figure supplement 1*; *Figure 2—figure supplement 12*; Materials and methods), suggesting that our main findings can remain unaffected in the presence of substantial feedforward inhibition. In addition, we limited our analysis to only a few overlapping ensembles. It will be interesting future work to study NTA in the case of many interacting and potentially overlapping ensembles and to determine the maximum storage capacity above which performance degrades. Finally, we anticipate that temporal differences in excitatory and inhibitory synaptic transmission may be important to preserve NTA's stimuli selectivity.

Our model makes several predictions. In contrast to balanced amplification, in which the network operating regime depends solely on the connectivity, an ensemble involved in NTA can transition from a non-ISN to an ISN state. Such a transition is consistent with noise variability observed in sensory cortices (*Hennequin et al., 2018*) and could be tested experimentally by probing the paradoxical effect under different stimulation conditions (*Figure 2G–H*; *Figure 2—figure supplement 6*). Moreover, NTA predicts that onset activity provides a better stimulus encoding and its activity is correlated with the fixed point activity. This signature is different from purely non-normal amplification mechanisms which would involve a wave of neuronal activity across several distinct ensembles similar to a synfire chain (*Abeles, 1991*). The difference should be clearly discernible in data. Since NTA relies on recurrent excitation between ensemble neurons, it suggests normal dynamics in which distinct ensembles first activate and then inactivate. The resulting dynamics have weak rotational components (*Figure 6D*) as seen in some experiments (*Marshel et al., 2019*). Strong non-normal amplification, on the other hand, relies on sequential activation associated with pronounced rotational dynamics (*Hennequin et al., 2014*; *Gillett et al., 2020*), as for instance observed in motor areas (*Churchland et al., 2012*). Although both non-normal mechanisms and NTA are likely to co-exist in the brain, we speculate that strong NTA is best suited for, and thus most like to be found in, sensory systems.

In summary, we introduced a general theoretical framework of selective transient signal amplification in recurrent networks. Our approach derives from the minimal assumptions of a nonlinear neuronal transfer function, recurrent excitation within neuronal ensembles, and STP. Importantly, our

analysis revealed the functional benefits of STP and EI co-tuning, both pervasively found in sensory circuits. Finally, our work suggests that transient onset responses rather than steady-state activity are ideally suited for coactivity-based stimulus encoding and provides several testable predictions.

# Materials and methods

## Stability conditions for supralinear networks

The dynamics of a neuronal ensemble consisting of one excitatory and one inhibitory population with a supralinear, power law input-output function can be described as follows:

$$\tau_E \frac{dr_E}{dt} = -r_E + \left[J_{EE} r_E - J_{EI} r_I + g_E\right]_+^{\alpha_E} \tag{7}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[J_{IE} r_E - J_{II} r_I + g_I\right]_+^{\alpha_I} \tag{8}$$

The Jacobian $\mathbf{M}$ of the system is given by

$$\mathbf{M} = \begin{bmatrix} \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & -\tau_E^{-1} J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} \\ \tau_I^{-1} J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} & -\tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) \end{bmatrix} \tag{9}$$

To ensure that the system is stable, the product of $\mathbf{M}$'s eigenvalues $\lambda_1\lambda_2$, which is equivalent to the determinant of $\mathbf{M}$, has to be positive. In addition, the sum of the two eigenvalues $\lambda_1 + \lambda_2$, which corresponds to $\mathrm{tr}(\mathbf{M})$, has to be negative. We therefore obtained the following two stability conditions

$$\lambda_1\lambda_2 = -\tau_E^{-1}\tau_I^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) + \tau_E^{-1}\tau_I^{-1} J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} > 0 \tag{10}$$

$$\lambda_1 + \lambda_2 = \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) < 0 \tag{11}$$

Notably, the stability conditions depend on the firing rate of the excitatory population $r_E$ and the inhibitory population $r_I$. Since firing rates are input-dependent, the stability of supralinear networks is input-dependent. In contrast, in linear networks in which $\alpha_E = \alpha_I = 1$, the conditions can be simplified to

$$\lambda_1\lambda_2 = -\tau_E^{-1}\tau_I^{-1}(J_{EE} - 1)(1 + J_{II}) + \tau_E^{-1}\tau_I^{-1} J_{EI} J_{IE} > 0 \tag{12}$$

$$\lambda_1 + \lambda_2 = \tau_E^{-1}(J_{EE} - 1) - \tau_I^{-1}(1 + J_{II}) < 0 \tag{13}$$

and are thus input-independent.

## ISN index for supralinear networks

If an ensemble is unstable without feedback inhibition, then the ensemble is an ISN (***Tsodyks et al., 1997***). To determine whether a given system is an ISN, we analyzed the stability of the E-E subnetwork, which is determined by the real part of the leading eigenvalue of the Jacobian of the E-E subnetwork. In the following, we call this leading eigenvalue the 'ISN index', which is defined as follows:

$$\text{ISN index} = \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) \tag{14}$$

A positive ISN index indicates the system is an ISN. Otherwise, the system is non-ISN. For supralinear networks in which $\alpha_E > 1$, the ISN index depends on the firing rates, inputs can therefore switch the network from non-ISN to ISN. In contrast, $\alpha_E = 1$ for linear networks which renders the ISN index firing rate independent.

## Characteristic function

To investigate how network stability changes with input, we trace the steps of ***Kraynyukova and Tchumatchenko, 2018*** and define the characteristic function $F(z)$ as follows:

$$F(z) = J_{EE}\left[z\right]_+^{\alpha_E} - J_{EI}\left[\det(\mathbf{J}) \cdot J_{EI}^{-1}\left[z\right]_+^{\alpha_E} + J_{EI}^{-1} J_{II} z - J_{EI}^{-1} J_{II} g_E + g_I\right]_+^{\alpha_I} - z + g_E \tag{15}$$

where

$$z = J_{EE}r_E - J_{EI}r_I + g_E \tag{16}$$

is the current into the excitatory population. The characteristic function simplifies the original two-dimensional system to a one-dimensional system, and the zero crossings of $F(z)$ correspond to the fixed points of the original system. For $z \geq 0$, we note:

$$\frac{dF(z)}{dz} = J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - J_{EI}\alpha_I\left(\det(\mathbf{J})\cdot J_{EI}^{-1}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} + J_{EI}^{-1}J_{II}\right)r_I^{\frac{\alpha_I-1}{\alpha_I}} - 1 = -\tau_E\tau_I\lambda_1\lambda_2 \tag{17}$$

Therefore, if the derivative of $F(z)$ evaluated at one of its roots is positive, the corresponding fixed point is a saddle point. Note that as $r_E$ and $r_I$ increase, the term in parenthesis becomes dominant. To ensure that $\lambda_1\lambda_2$ is negative also for large $r_E$ and $r_I$, the determinant of the weight matrix $\det(\mathbf{J})$ has to be positive. Therefore, $\det(\mathbf{J})$ has a decisive impact on the curvature of $F(z)$. In systems with negative determinant, $F(z)$ bends upwards for large $z$. In contrast, $F(z)$ asymptotically bends downwards in systems with positive determinant. Hence, the high-activity steady-state of systems with negative determinant is unstable. In addition, we can simplify the above condition to the determinant of the weight matrix which is a necessary condition for network stability at any firing rate:

$$\det(\mathbf{J}) = -J_{EE}J_{II} + J_{IE}J_{EI} > 0 \tag{18}$$

To investigate how the network stability changes with input $g_E$, we examined how $F(z)$ varies with changing input $g_E$ by calculating the derivative of $F(z)$ with respect to $g_E$,

$$\frac{dF(z)}{dg_E} = \alpha_I J_{II}\left[\det(\mathbf{J})\cdot J_{EI}^{-1}\left[z\right]_+^{\alpha_E} + J_{EI}^{-1}J_{II}z - J_{EI}^{-1}J_{II}g_E + g_I\right]_+^{\alpha_I-1} + 1 \tag{19}$$

Since $\frac{dF(z)}{dg_E}$ is positive, increasing $g_E$ always shifts $F(z)$ upwards, eventually leading to the vanishing of all roots and, thus, unstable dynamics in supralinear networks with negative $\det(\mathbf{J})$. In scenarios in which feedforward input to the inhibitory population also changes, we have

$$\begin{aligned}\frac{dF(z)}{dt} =& \frac{\partial F(z)}{\partial g_E}\frac{dg_E}{dt} + \frac{\partial F(z)}{\partial g_I}\frac{dg_I}{dt}\\ =& \left(\alpha_I J_{II}\left[\det(\mathbf{J})\cdot J_{EI}^{-1}\left[z\right]_+^{\alpha_E} + J_{EI}^{-1}J_{II}z - J_{EI}^{-1}J_{II}g_E + g_I\right]_+^{\alpha_I-1} + 1\right)\Delta g_E\\ &- \alpha_I J_{EI}\left[\det(\mathbf{J})\cdot J_{EI}^{-1}\left[z\right]_+^{\alpha_E} + J_{EI}^{-1}J_{II}z - J_{EI}^{-1}J_{II}g_E + g_I\right]_+^{\alpha_I-1}\Delta g_I\end{aligned} \tag{20}$$

When the change in stimulation strength into the excitatory ($\Delta g_E$) and the inhibitory population ($\Delta g_I$) are the same, $\frac{dF(z)}{dt}$ is always positive provided $J_{II}$ is greater than $J_{EI}$. Hence, depending on the value of $\frac{J_{II}}{J_{EI}}$, stimulation can lead to unstable network dynamics even when the input to the inhibitory population increases more than to the excitatory population.

## Spike-frequency adaptation (SFA)

We modeled SFA of excitatory neurons as an activity-dependent negative feedback current (***Benda and Herz, 2003***; ***Brette and Gerstner, 2005***):

$$\tau_E\frac{dr_E}{dt} = -r_E + \left[J_{EE}r_E - J_{EI}r_I + g_E\right]_+^{\alpha_E} - a \tag{21}$$

$$\tau_I\frac{dr_I}{dt} = -r_I + \left[J_{IE}r_E - J_{II}r_I + g_I\right]_+^{\alpha_I} \tag{22}$$

$$\tau_a\frac{da}{dt} = -a + br_E \tag{23}$$

where $a$ is the adaptation variable, $\tau_a$ is the adaptation time constant, and $b$ is the adaptation strength.

### Stability conditions in networks with SFA

The Jacobian $\mathbf{M_{SFA}}$ of the system with SFA is given by

$$\mathbf{M_{SFA}} = \begin{bmatrix} \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & -\tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} & -\tau_E^{-1} \\ \tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} & -\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) & 0 \\ \tau_a^{-1}b & 0 & -\tau_a^{-1} \end{bmatrix} \tag{24}$$

The characteristic polynomial of the system with SFA can be written as follows (***Horn and Johnson, 1985***):

$$\lambda^3 - \mathrm{tr}(\mathbf{M_{SFA}})\lambda^2 + (A_{11}+A_{22}+A_{33})\lambda - \det(\mathbf{M_{SFA}}) = 0 \tag{25}$$

where $\mathrm{tr}(\mathbf{M_{SFA}})$ and $\det(\mathbf{M_{SFA}})$ are the trace and the determinant of the Jacobian matrix $\mathbf{M_{SFA}}$, $A_{11}$, $A_{22}$, and $A_{33}$ are the matrix cofactors. More specifically,

$$\mathrm{tr}(\mathbf{M_{SFA}}) = \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) - \tau_a^{-1} \tag{26}$$

$$A_{11} = \begin{vmatrix} -\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) & 0 \\ 0 & -\tau_a^{-1} \end{vmatrix} = \tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})\tau_a^{-1} \tag{27}$$

$$A_{22} = \begin{vmatrix} \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & -\tau_E^{-1} \\ \tau_a^{-1}b & -\tau_a^{-1} \end{vmatrix} = -\tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_a^{-1} + \tau_a^{-1}b\tau_E^{-1} \tag{28}$$

$$\begin{aligned} A_{33} =& \begin{vmatrix} \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & -\tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} \\ \tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} & -\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) \end{vmatrix} \\ =& -\tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) + \tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}\tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \end{aligned} \tag{29}$$

$$\begin{aligned} A_{11}+A_{22}+A_{33} =& \tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})\tau_a^{-1} - \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_a^{-1} + \tau_a^{-1}b\tau_E^{-1} \\ & - \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) + \tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}\tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \end{aligned} \tag{30}$$

$$\begin{aligned} \det(\mathbf{M_{SFA}}) =& \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})\tau_a^{-1} \\ & - \tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}\tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}\tau_a^{-1} - \tau_a^{-1}b\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) \end{aligned} \tag{31}$$

To ensure that the dynamics of the system are stable, the real parts of the eigenvalues of the Jacobian at the fixed point, and thus all roots of the characteristic polynomial have to be negative. Since the product of the roots is equal to $\det(\mathbf{M_{SFA}})$, $-\det(\mathbf{M_{SFA}})$ has to be positive. We then have

$$b > \frac{\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}(J_{EE} - \det(\mathbf{J})\cdot\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})}{1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}} - 1 \tag{32}$$

Since SFA does not modify the synaptic connections, the term $J_{EE} - \det(\mathbf{J})\cdot\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}$ is positive for networks with $\det(\mathbf{J}) < 0$.

In the large $r_E$ limit, if $b$ is small such that the above condition cannot be fulfilled, $\det(\mathbf{M_{SFA}})$ is then positive, suggesting that the Jacobian of the system has always at least one positive eigenvalue. Therefore, the dynamics of the system cannot be stabilized in the presence of small $b$.

In addition, $A_{11}+A_{22}+A_{33}$ is equal to $\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3$, with the roots of the characteristic polynomial $\lambda_1$, $\lambda_2$, and $\lambda_3$. If all roots are real and negative, $A_{11}+A_{22}+A_{33}$ has to be positive. If one root is real and negative and two other roots are complex conjugates, to ensure that all roots have negative real parts, one necessary condition is $A_{11}+A_{22}+A_{33} > 0$. From the $\mathrm{tr}(\mathbf{M_{SFA}})$ and $\det(\mathbf{M_{SFA}})$ conditions, we have

$$A_{11}+A_{22}+A_{33} > \tau_a^{-1}(-\tau_a^{-1} + b\tau_E^{-1}) - b\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) \tag{33}$$

As a result, if $\tau_a^{-1}(-\tau_a^{-1}+b\tau_E^{-1})-b\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})>0$, $A_{11}+A_{22}+A_{33}$ is guaranteed to be positive. We therefore have

$$b[\tau_a^{-1}\tau_E^{-1}-\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})]>\tau_a^{-2} \tag{34}$$

Note that $\tau_a$ has to be small, in other words, SFA has to be fast, so that $\tau_a^{-1}\tau_E^{-1}-\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})$ is positive for arbitrary $r_I$. For positive $\tau_a^{-1}\tau_E^{-1}-\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})$, we have

$$b>\frac{\tau_a^{-2}}{\tau_a^{-1}\tau_E^{-1}-\tau_E^{-1}\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})} \tag{35}$$

Since $\tau_a$ has to be small, the above condition cannot be satisfied for small $b$.

Next, we consider the system with large $b$. Suppose that the firing rate $r_E$ and $r_I$ in the initial network are of order 1, and $b$ is of order $K$, where $K$ is a large number. We therefore have $-\mathrm{tr}(\mathbf{M_{SFA}})\sim O(1)$, $A_{11}+A_{22}+A_{33}\sim O(K)$, and $-\det(\mathbf{M_{SFA}})\sim O(K)$. The discriminant of the characteristic polynomial is

$$\begin{aligned}&(-\mathrm{tr}(\mathbf{M_{SFA}}))^2(A_{11}+A_{22}+A_{33})^2-4(A_{11}+A_{22}+A_{33})^3-4(-\mathrm{tr}(\mathbf{M_{SFA}}))^3(-\det(\mathbf{M_{SFA}}))\\&\quad-27(-\det(\mathbf{M_{SFA}}))^2+18(-\mathrm{tr}(\mathbf{M_{SFA}}))(A_{11}+A_{22}+A_{33})(-\det(\mathbf{M_{SFA}}))\\&=(A_{11}+A_{22}+A_{33})^3\Big[\frac{(-\mathrm{tr}(\mathbf{M_{SFA}}))^2}{A_{11}+A_{22}+A_{33}}-4-\frac{4(-\mathrm{tr}(\mathbf{M_{SFA}}))^3(-\det(\mathbf{M_{SFA}}))}{(A_{11}+A_{22}+A_{33})^3}\\&\quad-\frac{27(-\det(\mathbf{M_{SFA}}))^2}{(A_{11}+A_{22}+A_{33})^3}+\frac{18(-\mathrm{tr}(\mathbf{M_{SFA}}))(-\det(\mathbf{M_{SFA}}))}{(A_{11}+A_{22}+A_{33})^2}\Big]\end{aligned} \tag{36}$$

Clearly, in the large $b$ limit, the discriminant is negative, suggesting that the characteristic polynomial has one real root and two complex conjugate roots (***Irving, 2004***).

As the input $g_E$ increases, the complex conjugate eigenvalues cross the imaginary axis when $\mathrm{tr}(\mathbf{M_{SFA}})(A_{11}+A_{22}+A_{33})$ equals $\det(\mathbf{M_{SFA}})$. As a result, the system undergoes a supercritical Hopf bifurcation. We numerically confirmed that the resulting limit cycle is stable (***Figure 2—figure supplement 1***), consistent with previous work (***van Vreeswijk and Hansel, 2001***). Thus, the system shows oscillatory behavior instead of stable steady state.

## Short-term plasticity (STP)

We modeled E-to-E STD following previous work (***Tsodyks and Markram, 1997***; ***Varela et al., 1997***):

$$\tau_E\frac{dr_E}{dt}=-r_E+\left[xJ_{EE}r_E-J_{EI}r_I+g_E\right]_+^{\alpha_E} \tag{37}$$

$$\tau_I\frac{dr_I}{dt}=-r_I+\left[J_{IE}r_E-J_{II}r_I+g_I\right]_+^{\alpha_I} \tag{38}$$

$$\frac{dx}{dt}=\frac{1-x}{\tau_x}-U_d x r_E \tag{39}$$

where $x$ is the depression variable, which is limited to the interval $(0,1)$, $\tau_x$ is the depression time constant, and $U_d$ is the depression rate. The steady-state solution $x^*$ is given by

$$x^*=\frac{1}{1+U_d r_E\tau_x} \tag{40}$$

Similarly, we modeled E-to-I STF as

$$\tau_E\frac{dr_E}{dt}=-r_E+\left[J_{EE}r_E-J_{EI}r_I+g_E\right]_+^{\alpha_E} \tag{41}$$

$$\tau_I\frac{dr_I}{dt}=-r_I+\left[uJ_{IE}r_E-J_{II}r_I+g_I\right]_+^{\alpha_I} \tag{42}$$

$$\frac{du}{dt}=\frac{1-u}{\tau_u}+U_f(U_{max}-u)r_E \tag{43}$$

where $u$ is the facilitation variable constrained to the interval $(1, U_{max})$ , $U_{max}$ is the maximal facilitation value, $\tau_u$ is the time constant of STF, and $U_f$ is the facilitation rate. The steady-state solution $u^*$ is given by

$$u^* = \frac{1+U_f U_{max} r_E \tau_u}{1+U_f r_E \tau_u} \tag{44}$$

## Stability conditions for networks with E-to-E STD

The Jacobian $\mathbf{M_{STD}}$ of the system with E-to-E STD is given by

$$\mathbf{M_{STD}} = \begin{bmatrix} \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & -\tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} & \tau_E^{-1}J_{EE}\alpha_E r_E^{\frac{2\alpha_E-1}{\alpha_E}} \\ \tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} & -\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) & 0 \\ -U_d x & 0 & -\tau_x^{-1} - U_d r_E \end{bmatrix} \tag{45}$$

and the characteristic polynomial can be written as follows:

$$\lambda^3 - \mathrm{tr}(\mathbf{M_{STD}})\lambda^2 + (A_{11}+A_{22}+A_{33})\lambda - \det(\mathbf{M_{STD}}) = 0 \tag{46}$$

where $\mathrm{tr}(\mathbf{M_{STD}})$ and $\det(\mathbf{M_{STD}})$ are the trace and the determinant of the Jacobian matrix $\mathbf{M_{STD}}$, $A_{11}$, $A_{22}$, and $A_{33}$ are the matrix cofactors. More specifically,

$$\mathrm{tr}(\mathbf{M_{STD}}) = \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) - \tau_x^{-1} - U_d r_E \tag{47}$$

In the case of unstable dynamics, $r_E$ goes to infinity due to run-away excitation. However, the depression variable $x$ approaches zero in this limit, as $\lim_{r_E\to\infty} x = \lim_{r_E\to\infty} \frac{1}{1+U_d r_E \tau_x} = 0$. Therefore, in the large $r_E$ limit, $-\mathrm{tr}(\mathbf{M_{STD}})$ is positive.

$$\begin{aligned} A_{11}+A_{22}+A_{33} =& \tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})(\tau_x^{-1}+U_d r_E) \\ &+ \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)(-\tau_x^{-1} - U_d r_E) - \tau_E^{-1}J_{EE}\alpha_E r_E^{\frac{2\alpha_E-1}{\alpha_E}}(-U_d x) \\ &- \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) + \tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}\tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \end{aligned} \tag{48}$$

Similarly, in the large $r_E$ limit, $A_{11}+A_{22}+A_{33}$ is positive.

$$\begin{aligned} \det(\mathbf{M_{STD}}) =& \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})(\tau_x^{-1}+U_d r_E) \\ &- \tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}\tau_I^{-1}J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}(\tau_x^{-1}+U_d r_E) - \tau_E^{-1}J_{EE}\alpha_E r_E^{\frac{2\alpha_E-1}{\alpha_E}} U_d x \tau_I^{-1}(1+J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) \end{aligned} \tag{49}$$

Similarly, in the large $r_E$ limit, $-\det(\mathbf{M_{STD}})$ is positive.

According to the Descartes' rule of signs, the number of positive roots is at most the number of sign changes in the sequences of polynomial's coefficients. Therefore, there are no positive roots for the above characteristic polynomial and the network dynamics can be stabilized by E-to-E STD.

## Characteristic function approximation for networks with E-to-E STD

As demonstrated above, E-to-E STD is able to restabilize the system, there exists a stable steady state for which the STD variable $x$ is constant $x = x^*$. Because $x$ changes slowly compared to the neuronal dynamics, we can approximate it as constant which results in a natural reduction to a 2D system in which the weights with STD are modified. The stability of this 2D system can be readily characterized by the characteristic function $F(z)$ (***Kraynyukova and Tchumatchenko, 2018***), which depends on the previous steady state value of $x$. The characteristic function approximation with E-to-E STD can therefore be written as follows:

$$F(z) = xJ_{EE}\left[z\right]_+^{\alpha_E} - J_{EI}\left[\det(\mathbf{J_{STD}})\cdot J_{EI}^{-1}\left[z\right]_+^{\alpha_E} + J_{EI}^{-1}J_{II}z - J_{EI}^{-1}J_{II}g_E + g_I\right]_+^{\alpha_I} - z + g_E \tag{50}$$

where

$$\det(\mathbf{J_{STD}}) = \begin{vmatrix} xJ_{EE} & -J_{EI} \\ J_{IE} & -J_{II} \end{vmatrix} = -xJ_{EE}J_{II} + J_{IE}J_{EI} \tag{51}$$

Note that $\det(\mathbf{J_{STD}})$ can now change its sign due to E-to-E STD, the characteristic function can therefore change its bending shape. We used this relation to visualize how E-to-E STD effectively changes the network stability of the reduced system in ***Figure 2D***.

## Conditions for ISN in networks with E-to-E STD

Here, we identify the condition of being in the ISN regime in supralinear networks with E-to-E STD. When the level of inhibition is frozen, the Jacobian of the system reduces to the following:

$$\mathbf{M_1} = \begin{bmatrix} \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & \tau_E^{-1}J_{EE}\alpha_E r_E^{\frac{2\alpha_E-1}{\alpha_E}} \\ -U_d x & -\tau_x^{-1} - U_d r_E \end{bmatrix} \tag{52}$$

For the system with frozen inhibition, the dynamics are stable if

$$\mathrm{tr}(\mathbf{M_1}) = \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_x^{-1} - U_d r_E < 0 \tag{53}$$

and

$$\det(\mathbf{M_1}) = \tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)(-\tau_x^{-1} - U_d r_E) + \tau_E^{-1}J_{EE}\alpha_E r_E^{\frac{2\alpha_E-1}{\alpha_E}} U_d x > 0 \tag{54}$$

Therefore, if the network is an ISN at the fixed point, the following condition has to be satisfied:

$$x > \min\left(\sqrt{\frac{1}{J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}}}, \frac{\tau_x+\tau_E+\tau_E\tau_x U_d r_E}{\tau_x J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}}\right) \tag{55}$$

Furthermore, we define the largest real part of the eigenvalues of $\mathbf{M_1}$ as the ISN index for networks with E-to-E STD. More specifically,

$$\text{ISN index} = \mathrm{Re}\left[\frac{\tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_x^{-1} - U_d r_E}{2} + \sqrt{\frac{1}{4}(\tau_E^{-1}(xJ_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) + \tau_x^{-1} + U_d r_E)^2 - \tau_E^{-1}J_{EE}\alpha_E r_E^{\frac{2\alpha_E-1}{\alpha_E}} U_d x}\right] \tag{56}$$

## Conditions for paradoxical response in networks with E-to-E STD

Next, we identify the condition of having the paradoxical effect in supralinear networks with E-to-E STD. To that end, we exploit a separation of timescales between the fast neural activity and the slow STP variable. Therefore, set the depression variable to its value at the fixed point corresponding to the fixed point value of $r_E$. The excitatory nullcline is defined as follows

$$\tau_E\frac{dr_E}{dt} = -r_E + \left[\frac{1}{1+\tau_x U_d r_E}J_{EE}r_E - J_{EI}r_I + g_E\right]_+^{\alpha_E} = 0 \tag{57}$$

For $r_{E,I} > 0$, we have

$$r_I = \frac{\frac{1}{1+\tau_x U_d r_E}J_{EE}r_E - r_E^{\frac{1}{\alpha_E}} + g_E}{J_{EI}} \tag{58}$$

The slope of the excitatory nullcline in the $r_E/r_I$ plane where $x$ axis is $r_E$ and $y$ axis is $r_I$ can be written as follows

$$k_{STD}^E = \frac{1}{J_{EI}}\left(-\frac{J_{EE}}{(1+\tau_x U_d r_E)^2}\tau_x U_d r_E + \frac{J_{EE}}{1+\tau_x U_d r_E} - \frac{1}{\alpha_E}r_E^{\frac{1}{\alpha_E}-1}\right) \tag{59}$$

Note that the slope of the excitatory nullcline is nonlinear. To have paradoxical effect, the slope of the excitatory nullcline at the fixed point of the system has to be positive. Therefore, the STD variable $x$ at the fixed point has to satisfy the following condition

$$x > \sqrt{\frac{1}{J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}}} \tag{60}$$

The inhibitory nullcline can be written as follows

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[J_{IE}r_E - J_{II}r_I + g_I\right]_+^{\alpha_I} = 0 \tag{61}$$

In the region of rates $r_{E,I} > 0$, we have

$$r_I = \frac{J_{IE}r_E - r_I^{\frac{1}{\alpha_I}} + g_I}{J_{II}} \tag{62}$$

The slope of the inhibitory nullcline can be written as follows

$$k_{STD}^I = \frac{J_{IE}}{J_{II} + \frac{1}{\alpha_I} r_I^{\frac{1-\alpha_I}{\alpha_I}}} \tag{63}$$

In addition to the positive slope of the excitatory nullcline, the slope of the inhibitory nullcline at the fixed point of the system has to be larger than the slope of the excitatory nullcline. We therefore have

$$J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \left(\tau_x^{-1} + U_d r_E\right) > \left(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}\right)\left(-\frac{J_{EE}U_d r_E}{1+\tau_x U_d r_E}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} + \frac{J_{EE}}{1+\tau_x U_d r_E}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}\left(\tau_x^{-1} + U_d r_E\right) - \left(\tau_x^{-1} + U_d r_E\right)\right) \tag{64}$$

The above condition is the same as the stability condition of the determinant of the Jacobian of the system with E-to-E STD (***Eq. (49)***). Therefore, the condition is always satisfied when the system with E-to-E STD is stable.

Based on the condition of being ISN shown in ***Eq. (55)*** and the condition of having paradoxical effect shown in ***Eq. (60)***, we therefore can conclude that in supralinear networks with E-to-E STD, the paradoxical effect implies inhibitory stabilization, whereas inhibitory stabilization does not necessarily imply paradoxical responses. This is consistent with recent work by ***Sanzeni et al., 2020***, in which threshold-linear networks with STP have been studied. Here, we showed analytically that the conclusion holds for any rectified power-law activation function with positive $\alpha$.

To visualize the conditions in a two-dimensional plane, we reduced the conditions into a function of $J_{EE}$ and $x$. For ***Figure 2G***, $r_E = 1$. In ***Figure 2—figure supplement 5*** and ***Figure 2—figure supplement 8***, the depression variable thresholds above which the network exhibits the paradoxical effect were calculated based on ***Eq. (60)***.

## Uni-stability conditions

The system is said to be 'uni-stable', when it has a single stable fixed point. We first identified the uni-stability condition for networks with global inhibition. To that end, we considered a general network with $N$ excitatory populations and $N$ inhibitory populations. To treat this problem analytically, we did not take STP into account in our analysis. The Jacobian matrix of networks with global inhibition $\mathbf{Q}$, can be written as follows,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{J}_{E\leftarrow E} & \mathbf{J}_{E\leftarrow I} \\ \mathbf{J}_{I\leftarrow E} & \mathbf{J}_{I\leftarrow I} \end{bmatrix} \tag{65}$$

where $\mathbf{J}_{E\leftarrow E}$, $\mathbf{J}_{E\leftarrow I}$, $\mathbf{J}_{I\leftarrow E}$, and $\mathbf{J}_{I\leftarrow I}$ are $N$ by $N$ block matrices defined below.

$$\mathbf{J}_{E\leftarrow E}=\begin{bmatrix} a-e & ka & \cdots & ka \\ ka & a-e & \cdots & ka \\ \vdots & \vdots & \ddots & \vdots \\ ka & ka & \cdots & a-e \end{bmatrix} \tag{66}$$

$$\mathbf{J}_{E\leftarrow I}=-b\mathbf{J}_{N,N} \tag{67}$$

$$\mathbf{J}_{I\leftarrow E}=c\mathbf{J}_{N,N} \tag{68}$$

$$\mathbf{J}_{I\leftarrow I}=\begin{bmatrix} -d-f & -d & \cdots & -d \\ -d & -d-f & \cdots & -d \\ \vdots & \vdots & \ddots & \vdots \\ -d & -d & \cdots & -d-f \end{bmatrix} \tag{69}$$

where $a=\tau_E^{-1}J_{EE}\alpha_E[z_E]_+^{\alpha_E-1}$, $b=\tau_E^{-1}J_{EI}\alpha_E[z_E]_+^{\alpha_E-1}$, $c=\tau_I^{-1}J_{IE}\alpha_I[z_I]_+^{\alpha_I-1}$, $d=\tau_I^{-1}J_{II}\alpha_I[z_I]_+^{\alpha_I-1}$, $e=\tau_E^{-1}$, and $f=\tau_I^{-1}$. Here, $z_E$ and $z_I$ denote the total current into the excitatory and inhibitory population, respectively. Note that all these parameters are non-negative. Parameter $k$ controls the excitatory connection strength across different populations. $\mathbf{J}_{N,N}$ is a $N$ by $N$ matrix of ones.

The eigenvalues of the Jacobian $\mathbf{Q}$ are roots of its characteristic polynomial,

$$\det((\mathbf{J}_{E\leftarrow E}-\lambda\mathbb{1})(\mathbf{J}_{I\leftarrow I}-\lambda\mathbb{1})-\mathbf{J}_{E\leftarrow I}\mathbf{J}_{I\leftarrow E})=0 \tag{70}$$

where $\mathbb{1}$ represents the identity matrix of size $N$. The characteristic polynomial can be expanded to:

$$\Big[(a-e-ka-\lambda)(-f-\lambda)\Big]^{N-1}\Big[(a-e+(N-1)ka-\lambda)(-Nd-f-\lambda)+N^2bc\Big]=0 \tag{71}$$

We therefore had four distinct eigenvalues:

$$\lambda_1=a-e-ka \tag{72}$$

$$\lambda_2=-f \tag{73}$$

and

$$\lambda_{3/4}=\frac{1}{2}\Big[(a-e-f-Nd+(N-1)ka)\pm\sqrt{(a-e-f-Nd+(N-1)ka)^2-4((-af+ef+kaf)-N(a-e)d-Nkaf-N(N-1)kad+N^2bc)}\Big] \tag{74}$$

Note that the eigenvalues $\lambda_1$ and $\lambda_2$ have an algebraic and geometric multiplicity of ($N$–1), whereas the eigenvalues $\lambda_3$ and $\lambda_4$ have an algebraic and geometric multiplicity of 1.

In analogy to networks with global inhibition, the Jacobian matrix of networks with co-tuned inhibition $\mathbf{R}$, can be written as

$$\mathbf{R}=\begin{bmatrix} \mathbf{J}_{E\leftarrow E} & \mathbf{J}_{E\leftarrow I} \\ \mathbf{J}_{I\leftarrow E} & \mathbf{J}_{I\leftarrow I} \end{bmatrix} \tag{75}$$

where $\mathbf{J}_{E\leftarrow E}$, $\mathbf{J}_{E\leftarrow I}$, $\mathbf{J}_{I\leftarrow E}$, and $\mathbf{J}_{I\leftarrow I}$ are $N$ by $N$ block matrices defined as follows:

$$\mathbf{J}_{E\leftarrow E}=\begin{bmatrix} a-e & ka & \cdots & ka \\ ka & a-e & \cdots & ka \\ \vdots & \vdots & \ddots & \vdots \\ ka & ka & \cdots & a-e \end{bmatrix} \tag{76}$$

$$\mathbf{J}_{E\leftarrow I}=\begin{bmatrix}-Nb+(N-1)mb & -mb & \cdots & -mb\\ -mb & -Nb+(N-1)mb & \cdots & -mb\\ \vdots & \vdots & \ddots & \vdots\\ -mb & -mb & \cdots & -Nb+(N-1)mb\end{bmatrix} \tag{77}$$

$$\mathbf{J}_{I\leftarrow E}=\begin{bmatrix}Nc-(N-1)mc & mc & \cdots & mc\\ mc & Nc-(N-1)mc & \cdots & mc\\ \vdots & \vdots & \ddots & \vdots\\ mc & mc & \cdots & Nc-(N-1)mc\end{bmatrix} \tag{78}$$

$$\mathbf{J}_{I\leftarrow I}=\begin{bmatrix}-Nd+(N-1)md-f & -md & \cdots & -md\\ -md & -Nd+(N-1)md-f & \cdots & -md\\ \vdots & \vdots & \ddots & \vdots\\ -md & -md & \cdots & -Nd+(N-1)md-f\end{bmatrix} \tag{79}$$

where $m$ controls the degree of co-tuning in the network. If $m=0$, the network decouples into $N$ independent ensembles and inhibition is perfectly co-tuned with excitation. In the case $m=1$, inhibition is global and the block matrices become identical to the above case of global inhibition.

The eigenvalues of the matrix $\mathbf{R}$ are given as the roots of the characteristic polynomial defined by:

$$\det((\mathbf{J}_{E\leftarrow E}-\lambda\mathbb{1})(\mathbf{J}_{I\leftarrow I}-\lambda\mathbb{1})-\mathbf{J}_{E\leftarrow I}\mathbf{J}_{I\leftarrow E})=0 \tag{80}$$

which yields the following expression:

$$\begin{aligned}&\Big[\lambda^2-(a-e-ka-Nd+Nmd-f)\lambda-(a-e-ka)(Nd-Nmd-f)\\&+N^2bc(1-m)^2\Big]^{N-1}\Big[(a-e+(N-1)ka-\lambda)(-Nd-f-\lambda)+N^2bc\Big]=0\end{aligned} \tag{81}$$

We therefore had four distinct eigenvalues:

$$\lambda_{1/2}^{'}=\frac{1}{2}\Big[(a-e-ka-Nd+Nmd-f)\pm\sqrt{(a-e-ka+Nd-Nmd+f)^2-4N^2bc(1-m)^2}\Big] \tag{82}$$

$$\lambda_{3/4}^{'}=\frac{1}{2}\Big[(a-e-f-Nd+(N-1)ka)\pm\sqrt{(a-e-f-Nd+(N-1)ka)^2-4((-af+ef+kaf)-N(a-e)d-Nkaf-N(N-1)kad+N^2bc)}\Big] \tag{83}$$

The eigenvalues $\lambda_1^{'}$ and $\lambda_2^{'}$ have an algebraic and geometric multiplicity of ($N$–1), whereas the eigenvalues $\lambda_3^{'}$ and $\lambda_4^{'}$ have an algebraic and geometric multiplicity of 1. We noted that $\lambda_3=\lambda_3^{'}$, $\lambda_4=\lambda_4^{'}$.

To compare under which conditions networks with different structures are uni-stable, we examined the different eigenvalues derived above. As $\lambda_2<0$, and $\lambda_1^{'}>\lambda_2^{'}$, we only had to compare $\lambda_1^{'}$ to $\lambda_1$. For networks with co-tuned inhibition, we have $m<1$,

$$\begin{aligned}\lambda_1^{'}=&\frac{1}{2}\Big[(a-e-ka-Nd+Nmd-f)+\sqrt{(a-e-ka+Nd-Nmd+f)^2-4N^2bc(1-m)^2}\Big]\\<&\frac{1}{2}\Big[(a-e-ka-Nd+Nmd-f)+\sqrt{(a-e-ka+Nd-Nmd+f)^2}\Big]=a-e-ka=\lambda_1\end{aligned} \tag{84}$$

The inequality, $\lambda_1^{'}<\lambda_1$, indicates that networks with co-tuned inhibition have a broad parameter regime in which they are uni-stable than networks with global inhibition. Note that in the absence of a saturating nonlinearity of the input-output function and in the absence of any additional stabilization mechanisms, systems with positive eigenvalues of the Jacobian are unstable. In this case, networks with co-tuned inhibition have a broad parameter regime of being stable than networks with global inhibition.

To visualize the conditions in a two-dimensional plane, we reduced the conditions into a function of $a$ and $d$. For ***Figure 3C***, $k = 0.1$, $m = 0.5$ and $bc = 0.9ad$.

## Distance to the decision boundary

To calculate the distance to the decision boundary in ***Figures 4 and 5***, ***Figure 4—figure supplement 2*** and ***Figure 5—figure supplement 2***, we first projected the excitatory activity in Phase two onto a two-dimensional Cartesian coordinate system in which the horizontal axis is the activity of the first excitatory ensemble $r_{E1}$ and the vertical axis is the activity of the second excitatory ensemble $r_{E2}$. We denote the location of the projected data point in the Cartesian coordinate system by $(x, y)$, where $x$ and $y$ equal $r_{E1}$ and $r_{E2}$, respectively. The distance $L$ between the projected data and the decision boundary which corresponds to the diagonal line in the coordinate system can be expressed as follows:

$$L = \sqrt{x^2 + y^2}\sin(|45^o - \arcsin(\frac{x}{\sqrt{x^2+y^2}})|) \tag{85}$$

Note that the inverse trigonometric function arcsin gives the value of the angle in degrees.

## Inhibitory feedback pathways for suppressing unwanted neural activation

To identify the important neural pathways for the suppression of unwanted neural activation, we analyzed how the activity of the second excitatory ensemble $r_{E2}$ changes with the input to the first excitatory ensemble $g_{E1}$. To that end, we considered a general weight matrix for networks with two interacting ensembles

$$\mathbf{J} = \begin{bmatrix} J_{E1E1} & J_{E1E2} & -J_{E1I1} & -J_{E1I2} \\ J_{E2E1} & J_{E2E2} & -J_{E2I1} & -J_{E2I2} \\ J_{I1E1} & J_{I1E2} & -J_{I1I1} & -J_{I1I2} \\ J_{I2E1} & J_{I2E2} & -J_{I2I1} & -J_{I2I2} \end{bmatrix} \tag{86}$$

We can write the change in firing rate of the excitatory population in the second ensemble $\delta r_{E2}$ as a function of the change in the input to the other $\delta g_{E1}$:

$$\begin{aligned}\delta r_{E2} = &\frac{1}{\det(\mathbb{1} - \mathbf{FJ})}\Big[(-f'_{E2}J_{E2E1})f'_{I1}J_{I1I2}f'_{I2}J_{I2I1} + f'_{E2}J_{E2I1}(-f'_{I1}J_{I1E1})(1 + f'_{I2}J_{I2I2}) + f'_{E2}J_{E2I2}(1 + f'_{I1}J_{I1I1})(-f'_{I2}J_{I2E1}) \\ &- (-f'_{E2}J_{E2E1})(1 + f'_{I1}J_{I1I1})(1 + f'_{I2}J_{I2I2}) - f'_{E2}J_{E2I1}f'_{I1}J_{I1I2}(-f'_{I2}J_{I2E1}) - f'_{E2}J_{E2I2}(-f'_{I1}J_{I1E1})f'_{I2}J_{I2I1}\Big]f'_{E1}\delta g_{E1}\end{aligned} \tag{87}$$

where $\mathbb{1}$ is the identity matrix. And $\mathbf{F}$ is given by

$$\mathbf{F} = \begin{bmatrix} f'_{E1} & 0 & 0 & 0 \\ 0 & f'_{E2} & 0 & 0 \\ 0 & 0 & f'_{I1} & 0 \\ 0 & 0 & 0 & f'_{I2} \end{bmatrix} \tag{88}$$

where $f'_{E1}$, $f'_{E2}$, $f'_{I1}$ and $f'_{I2}$ are the derivatives of the input-output functions evaluated at the fixed point.

Assuming that $J_{E1E1} = J_{E2E2} = J_{EE}$, $J_{I1E1} = J_{I2E2} = J_{IE}$, $J_{E1I1} = J_{E2I2} = J_{EI}$, $J_{I1I1} = J_{I2I2} = J_{II}$, $J_{E1E2} = J_{E2E1} = J'_{EE}$, $J_{I1E2} = J_{I2E1} = J'_{IE}$, $J_{E1I2} = J_{E2I1} = J'_{EI}$ and $J_{I1I2} = J_{I2I1} = J'_{II}$, we find

$$\begin{aligned}\delta r_{E2} = &\frac{1}{\det(\mathbb{1} - \mathbf{FJ})}\Big[(-f'_{E2}J'_{EE})f'_{I1}J'_{II}f'_{I2}J'_{II} + f'_{E2}J'_{EI}(-f'_{I1}J_{IE})(1 + f'_{I2}J_{II}) + f'_{E2}J_{EI}(1 + f'_{I1}J_{II})(-f'_{I2}J'_{IE}) \\ &- (-f'_{E2}J'_{EE})(1 + f'_{I1}J_{II})(1 + f'_{I2}J_{II}) - f'_{E2}J'_{EI}f'_{I1}J'_{II}(-f'_{I2}J'_{IE}) - f'_{E2}J_{EI}(-f'_{I1}J_{IE})f'_{I2}J'_{II}\Big]f'_{E1}\delta g_{E1}\end{aligned} \tag{89}$$

By further assuming that the weight strengths across ensembles are weak and ignoring the corresponding higher-order terms, we get

$$\begin{aligned}\delta r_{E2} \approx \frac{1}{\det(\mathbb{1}-\mathbf{FJ})}\Big[&f'_{E2}J'_{EI}(-f'_{I1}J_{IE})(1+f'_{I2}J_{II})+f'_{E2}J_{EI}(1+f'_{I1}J_{II})(-f'_{I2}J'_{IE})\\&-(-f'_{E2}J'_{EE})(1+f'_{I1}J_{II})(1+f'_{I2}J_{II})-f'_{E2}J_{EI}(-f'_{I1}J_{IE})f'_{I2}J'_{II}\Big]f'_{E1}\delta g_{E1}\\=\frac{1}{\det(\mathbb{1}-\mathbf{FJ})}\Bigg[&\left(\frac{J'_{II}}{J'_{EI}}f'_{I2}-\left(\frac{1}{J_{EI}}+f'_{I2}\frac{J_{II}}{J_{EI}}\right)\right)J'_{EI}J_{EI}J_{IE}f'_{E2}f'_{I1}\\&+\left(\frac{J'_{EE}}{J'_{IE}}(1+J_{II}f'_{I2})-J_{EI}f'_{I2}\right)J'_{IE}f'_{E2}(1+f'_{I1}J_{II})\Bigg]f'_{E1}\delta g_{E1}\end{aligned} \tag{90}$$

Note that $\frac{J'_{EE}}{J'_{IE}}$ and $\frac{J'_{II}}{J'_{EI}}$ are terms regulating the respective excitatory and inhibitory input from one ensemble to the excitatory and inhibitory population in another ensemble. The term $\det(\mathbb{1}-\mathbf{FJ})$ is positive to ensure the stability of the system.

To suppress the activity of the excitatory population in the second ensemble $r_{E2}$, in other words, to ensure that $\delta r_{E2}<0$, $J'_{IE}$ or/and $J'_{EI}$ have to be large. Therefore, we identified $J'_{IE}$ and $J'_{EI}$ as important synaptic connections which lead to suppression of the unwanted neural activation, suggesting that inhibition can be provided via $J'_{IE}$ through the $E1$-$I2$-$E2$ pathway or via $J'_{EI}$ through the $E1$-$I1$-$E2$ pathway.

For ***Figures 4 and 5***, the rate-based model consists of two ensembles, each of which is composed of 100 excitatory and 25 inhibitory neurons with all-to-all connectivity.

## Spiking neural network model

The spiking neural network model was composed of $N_E$ excitatory and $N_I$ inhibitory leaky integrate-and-fire neurons. Neurons were randomly connected with probability of 20%. The dynamics of membrane potential of neuron $i$, $U_i$, as defined by ***Zenke et al., 2015***:

$$\tau^m\frac{dU_i}{dt}=(U^{\text{rest}}-U_i)+g_i^{\text{ext}}(t)(U^{\text{exc}}-U_i)+g_i^{\text{inh}}(t)(U^{\text{inh}}-U_i) \tag{91}$$

Here, $\tau^m$ is the membrane time constant and $U^{\text{rest}}$ is the resting potential. Spikes are triggered when the membrane potential reaches the spiking threshold $U^{\text{thr}}$. After a spike is emitted, the membrane potential is reset to $U^{\text{rest}}$ and the neuron enters a refractory period of $\tau^{\text{ref}}$. Inhibitory neurons obeyed the same integrate-and-fire formalism but with a shorter membrane time constant.

Excitatory synapses contain a fast AMPA component and a slow NMDA component. The dynamics of the excitatory conductance are described by:

$$\tau^{\text{ampa}}\frac{dg_i^{\text{ampa}}}{dt}=-g_i^{\text{ampa}}+\sum_{j\in\text{exc}}J_{ij}S_j(t) \tag{92}$$

$$\tau^{\text{nmda}}\frac{dg_i^{\text{nmda}}}{dt}=-g_i^{\text{nmda}}+g_i^{\text{ampa}} \tag{93}$$

$$g_i^{\text{exc}}(t)=\xi g_i^{\text{ampa}}(t)+(1-\xi)g_i^{\text{nmda}}(t) \tag{94}$$

Here, $J_{ij}$ denotes the synaptic strength from neuron $j$ to neuron $i$. If the connection does not exist, $J_{ij}$ was set to 0. $S_j(t)$ is the spike train of neuron $j$, which is defined as $S_j(t)=\sum_k\delta(t-t_j^k)$, where $\delta$ is the Dirac delta function and $t_j^k$ the spikes times $k$ of neuron $j$. $\xi$ is a weighting parameter. The dynamics of inhibitory conductances are governed by:

$$\tau^{\text{gaba}}\frac{dg_i^{\text{inh}}}{dt}=-g_i^{\text{inh}}+\sum_{j\in\text{inh}}J_{ij}S_j(t) \tag{95}$$

In the spiking neural network models, SFA of excitatory neurons is modeled as follows,

$$\tau^m\frac{dU_i}{dt}=(U^{\text{rest}}-U_i)+g_i^{\text{ext}}(t)(U^{\text{exc}}-U_i)+(g_i^{\text{inh}}(t)+a_i(t))(U^{\text{inh}}-U_i) \tag{96}$$

$$\frac{da_i}{dt}=-\frac{a_i}{\tau_a}+bS_i(t) \tag{97}$$

where $i$ is the index of excitatory neurons.
The dynamics of E-to-E STD are given by

$$\frac{dx_{ij}}{dt} = \frac{1 - x_{ij}}{\tau_x} - U_d x_{ij} S_j(t) \tag{98}$$

$$\tau^{\text{ampa}} \frac{dg_i^{\text{ampa}}}{dt} = -g_i^{\text{ampa}} + \sum_{j \in \text{exc}} x_{ij} J_{ij} S_j(t) \tag{99}$$

where $i$ represents the index of excitatory neurons.
The dynamics of E-to-I STF are governed by

$$\frac{du_{ij}}{dt} = \frac{1 - u_{ij}}{\tau_u} + U_f(U_{max} - u_{ij}) S_j(t) \tag{100}$$

$$\tau^{\text{ampa}} \frac{dg_i^{\text{ampa}}}{dt} = -g_i^{\text{ampa}} + \sum_{j \in \text{exc}} u_{ij} J_{ij} S_j(t) \tag{101}$$

where $i$ denotes the index of inhibitory neurons.

For ***Figure 6***, each excitatory and inhibitory neuron received external excitatory input from 300 neurons firing with Poisson statistics at an average firing rate of 0.1 Hz at baseline. During stimulation, the excitatory neurons corresponding to the activated ensemble received external excitatory input from 300 neurons firing with Poisson statistics at an average firing rate of 0.5 Hz. The ensemble activity is computed from the instantaneous firing rates of the respective ensembles with 10ms bin size. The difference in ensemble activity for the peak amplitude is calculated by subtracting the average maximal ensemble activity of the unstimulated ensembles from the maximal ensemble activity of the activated ensemble. Similarly, the difference in ensemble activity for the fixed point is calculated by subtracting the average ensemble activity of the unstimulated ensembles at the fixed point from the ensemble activity of the activated ensemble at the fixed point. Fixed point activity is computed by averaging the activity of the middle 1 second within the 2-second stimulation period.

For ***Figure 2—figure supplement 10***, each excitatory and inhibitory neuron received external excitatory input from 300 neurons firing with Poisson statistics at an average firing rate of 0.1 Hz at

**Table 1.** Parameters for ***Figure 1C–E***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 1.8 | - | E-to-E connection strength |
| $J_{IE}$ | 1.0 | - | E-to-I connection strength |
| $J_{EI}$ | 1.0 | - | I-to-E connection strength |
| $J_{II}$ | 0.6 | - | I-to-I connection strength |
| $\alpha_E$ | 2 | - | Power of excitatory input-output function |
| $\alpha_I$ | 2 | - | Power of inhibitory input-output function |
| $\tau_E$ | 20 | ms | Time constant of excitatory firing dynamics |
| $\tau_I$ | 10 | ms | Time constant of inhibitory firing dynamics |
| $g_E^{bs}$ | 1.55 | - | Input to the *E* population at baseline |
| $g_E^{stim}$ | 3.0 | - | Input to the *E* population during stimulation |
| $g_I$ | 2.0 | - | Input to the *I* population |
| Parameters for ***Figure 1F*** | | | |
| $J_{IE}$ | 0.45 | - | E-to-I connection strength |
| $J_{EI}$ | 1.0 | - | I-to-E connection strength |
| $J_{II}$ | 1.5 | - | I-to-I connection strength |

**Table 2.** Parameters for ***Figure 2***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $\tau_a$ | 200 | ms | Time constant of SFA |
| $b$ | 1.0 | - | Strength of SFA |
| $\tau_x$ | 200 | ms | Time constant of STD |
| $U_d$ | 1.0 | - | Depression rate |
| $\tau_u$ | 200 | ms | Time constant of STF |
| $U_f$ | 1.0 | - | Facilitation rate |
| $U_{max}$ | 6.0 | - | Maximal facilitation value |

Note that these values are also applied elsewhere unless mentioned otherwise.

the baseline. During stimulation, each excitatory neuron received external excitatory input from 300 neurons firing with Poisson statistics at an average firing rate of 0.3 Hz.

For ***Figure 6—figure supplement 1***, the firing rates of 300 neurons are varying from 4/15 Hz to 7/15 Hz.

## Simulations

Simulations were performed in Python and Mathematica. All differential equations were implemented by Euler integration with a time step of 0.1 ms. All simulation parameters are listed in ***Tables 1–5*** and ***Appendix 5—Tables 1–10***. The simulation source code to reproduce the figures is publicly available at https://github.com/fmi-basel/gzenke-nonlinear-transient-amplification (***Wu, 2021*** copy archived at swh:1:rev:6ff6ff10b9f4994a0f948a987a66cc82f98451e1).

**Table 3.** Parameters for ***Figure 3*** bi/multi-stable example.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 1.4 | - | Within-ensemble E-to-E connection strength |
| $J_{IE}$ | 0.6 | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | 1.0 | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | 0.6 | - | Within-ensemble I-to-I connection strength |
| $J'_{EE}$ | 0.14 | - | Inter-ensemble E-to-E connection strength |
| $J'_{IE}$ | 0.6 | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | 1.0 | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | 0.6 | - | Inter-ensemble I-to-I connection strength |
| $g_{E1}^{bs}$ | 2.2 | - | Input to the *E*1 population at baseline |
| $g_{E1}^{stim}$ | 3.0 | - | Input to the *E*1 population during stimulation |
| $g_{E2}$ | 2.2 | - | Input to the *E*2 population |
| $g_I$ | 2.0 | - | Input to the *I* population |
| Parameters for ***Figure 3*** uni-stable example | | | |
| $J_{EE}$ | 1.3 | - | Within-ensemble E-to-E connection strength |
| $J'_{EE}$ | 0.13 | - | Inter-ensemble E-to-E connection strength |

**Table 4.** Parameters for ***Figures 4 and 5***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $N_E$ | 200 | - | Number of excitatory neurons |
| $N_I$ | 50 | - | Number of inhibitory neurons |
| $N$ | 2 | - | Number of ensembles |
| $J_{EE}$ | $1.2/(N_E/2-1)$ | - | Within-ensemble E-to-E connection strength |
| $J_{IE}$ | $1.0/(N_E/2)$ | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | $1.0/(N_I/2)$ | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | $1.0/(N_I/2-1)$ | - | Within-ensemble I-to-I connection strength |
| $J'_{EE}$ | $0.36/(N_E/2-1)$ | - | Inter-ensemble E-to-E connection strength |
| $J'_{IE}$ | $0.4/(N_E/2)$ | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | $0.1/(N_I/2)$ | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | $0.1/(N_I/2)$ | - | Inter-ensemble I-to-I connection strength |
| $g_I$ | 2.0 | - | Input to the *I* population |
| Parameters for ***Figure 4*** | | | |
| $g_{E1}^{bs}$ | 1.35 | - | Input to the *E*1 population |
| $g_{E1}^{stim}$ | 4.0 | - | Input to the *E*1 population during stimulation |
| $g_{E2}$ | 1.35 | - | Input to the *E*2 population |
| Parameters for ***Figure 5*** | | | |
| $g_{E1}^{bs}$ | 1.35 | - | Input to the *E*1 population at baseline |
| $g_{E1}^{stim}$ | 1.35 + (4.0–1.35) (1-$p$) | - | Input to the *E*1 population during stimulation |
| $g_{E2}^{bs}$ | 1.35 | - | Input to the *E*2 population at baseline |
| $g_{E2}^{stim}$ | 1.35 + (4.0–1.35)$p$ | - | Input to the *E*2 population during stimulation |
| Here, $p$ is a parameter between 0 and 1 controlling the additional inputs to *E1* and *E2*. | | | |

**Table 5.** Parameters for ***Figure 6***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $N_E$ | 400 | - | Number of excitatory neurons |
| $N_I$ | 100 | - | Number of inhibitory neurons |
| $U^{\text{rest}}$ | –70 | mV | Resting membrane potential |
| $U^{\text{exc}}$ | 0 | mV | Excitatory reversal potential |
| $U^{\text{inh}}$ | –80 | mV | Inhibitory reversal potential |
| $\tau^{\text{ref}}$ | 3 | ms | Duration of refractory period |
| $\tau^{m}_{\text{exc}}$ | 20 | ms | Membrane time constant of excitatory neurons |
| $\tau^{m}_{\text{inh}}$ | 10 | ms | Membrane time constant of inhibitory neurons |
| $\tau^{\text{ampa}}$ | 5 | ms | Time constant of AMPA receptor |
| $\tau^{\text{gaba}}$ | 10 | ms | Time constant of GABA receptor |
| $\tau^{\text{nmda}}$ | 100 | ms | Time constant of NMDA receptor |
| $\xi$ | 0.5 | - | Receptor weighting factor |
| $J_{EE}$ | 0.19 | - | Within-ensemble E-to-E connection strength |
| $J_{IE}$ | 0.10 | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | 0.10 | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | 0.06 | - | Within-ensemble I-to-I connection strength |
| $J'_{EE}$ | 0.019 | - | Inter-ensemble E-to-E connection strength |
| $J'_{IE}$ | 0.05 | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | 0.04 | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | 0.006 | - | Inter-ensemble I-to-I connection strength |

## Acknowledgements

We thank Rainer W Friedrich, Claire Meissner-Bernard, William F Podlaski, and members of the Zenke Group for comments and discussions. This work was supported by the Novartis Research Foundation.

## Additional information

### Funding

| Funder | Grant reference number | Author |
|---|---|---|
| Novartis Foundation | | Yue Kris Wu<br>Friedemann Zenke |

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

### Author contributions

Yue Kris Wu, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing; Friedemann Zenke, Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review and editing

**Author ORCIDs**
Yue Kris Wu http://orcid.org/0000-0002-9804-2537
Friedemann Zenke http://orcid.org/0000-0003-1883-644X

**Decision letter and Author response**
Decision letter https://doi.org/10.7554/eLife.71263.sa1
Author response https://doi.org/10.7554/eLife.71263.sa2

## Additional files

### Supplementary files

- Transparent reporting form

### Data availability

This project is a theory project without data. All simulation code has been deposited on GitHub under https://github.com/fmi-basel/gzenke-nonlinear-transient-amplification, (copy archived at swh:1:rev:6ff6ff10b9f4994a0f948a987a66cc82f98451e1).

## References

**Abeles M**. 1991. Corticonics. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511574566

**Ahmadian Y**, Rubin DB, Miller KD. 2013. Analysis of the stabilized supralinear network. *Neural Computation* **25**:1994–2037. DOI: https://doi.org/10.1162/NECO_a_00472, PMID: 23663149

**Amit DJ**, Brunel N. 1997. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex* **7**:237–252. DOI: https://doi.org/10.1093/cercor/7.3.237, PMID: 9143444

**Baker C**, Zhu V, Rosenbaum R. 2020. Nonlinear stimulus representations in neural circuits with approximate excitatory-inhibitory balance. *PLOS Computational Biology* **16**:1008192. DOI: https://doi.org/10.1371/journal.pcbi.1008192, PMID: 32946433

**Barkai E**, Hasselmo ME. 1994. Modulation of the input/output function of rat piriform cortex pyramidal cells. *Journal of Neurophysiology* **72**:644–658. DOI: https://doi.org/10.1152/jn.1994.72.2.644, PMID: 7983526

**Benda J**, Herz AVM. 2003. A universal model for spike-frequency adaptation. *Neural Computation* **15**:2523–2564. DOI: https://doi.org/10.1162/089976603322385063, PMID: 14577853

**Bissière S**, Humeau Y, Lüthi A. 2003. Dopamine gates LTP induction in lateral amygdala by suppressing feedforward inhibition. *Nature Neuroscience* **6**:587–592. DOI: https://doi.org/10.1038/nn1058, PMID: 12740581

**Bolding KA**, Franks KM. 2018. Recurrent cortical circuits implement concentration-invariant odor coding. *Science* **361**:eaat6904. DOI: https://doi.org/10.1126/science.aat6904, PMID: 30213885

**Bolding KA**, Nagappan S, Han BX, Wang F, Franks KM. 2020. Recurrent circuitry is required to stabilize piriform cortex odor representations across brain states. *eLife* **9**:e53125. DOI: https://doi.org/10.7554/eLife.53125, PMID: 32662420

**Bondanelli G**, Ostojic S. 2020. Coding with transient trajectories in recurrent neural networks. *PLOS Computational Biology* **16**:e1007655. DOI: https://doi.org/10.1371/journal.pcbi.1007655, PMID: 32053594

**Brette R**, Gerstner W. 2005. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology* **94**:3637–3642. DOI: https://doi.org/10.1152/jn.00686.2005, PMID: 16014787

**Carrillo-Reid L**, Yang W, Bando Y, Peterka DS, Yuste R. 2016. Imprinting and recalling cortical ensembles. *Science* **353**:691–694. DOI: https://doi.org/10.1126/science.aaf7560, PMID: 27516599

**Cayco-Gajic NA**, Silver RA. 2019. Re-evaluating Circuit Mechanisms Underlying Pattern Separation. *Neuron* **101**:584–602. DOI: https://doi.org/10.1016/j.neuron.2019.01.044, PMID: 30790539

**Christodoulou G**, Vogels TP, Agnes EJ. 2021. Regimes and Mechanisms of Transient Amplification in Abstract and Biological Networks. [bioRxiv]. DOI: https://doi.org/10.1101/2021.04.01.437964

**Churchland MM**, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV. 2012. Neural population dynamics during reaching. *Nature* **487**:51–56. DOI: https://doi.org/10.1038/nature11129, PMID: 22722855

**Compte A**, Brunel N, Goldman-Rakic PS, Wang XJ. 2000. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex* **10**:910–923. DOI: https://doi.org/10.1093/cercor/10.9.910, PMID: 10982751

**Cossell L**, Iacaruso MF, Muir DR, Houlton R, Sader EN, Ko H, Hofer SB, Mrsic-Flogel TD. 2015. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518**:399–403. DOI: https://doi.org/10.1038/nature14182, PMID: 25652823

**Cruikshank SJ**, Lewis TJ, Connors BW. 2007. Synaptic basis for intense thalamocortical activation of feedforward inhibitory cells in neocortex. *Nature Neuroscience* **10**:462–468. DOI: https://doi.org/10.1038/nn1861, PMID: 17334362

**Denève S**, Machens CK. 2016. Efficient codes and balanced networks. *Nature Neuroscience* **19**:375–382. DOI: https://doi.org/10.1038/nn.4243, PMID: 26906504
**DeWeese MR**, Wehr M, Zador AM. 2003. Binary spiking in auditory cortex. *The Journal of Neuroscience* **23**:7940–7949. DOI: https://doi.org/10.1523/JNEUROSCI.23-21-07940.2003, PMID: 12944525
**Echeveste R**, Aitchison L, Hennequin G, Lengyel M. 2020. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience* **23**:1138–1149. DOI: https://doi.org/10.1038/s41593-020-0671-1, PMID: 32778794
**El-Gaby M**, Reeve HM, Lopes-Dos-Santos V, Campo-Urriza N, Perestenko PV, Morley A, Strickland LAM, Lukács IP, Paulsen O, Dupret D. 2021. An emergent neural coactivity code for dynamic memory. *Nature Neuroscience* **24**:694–704. DOI: https://doi.org/10.1038/s41593-021-00820-w, PMID: 33782620
**Engel AK**, Fries P, Singer W. 2001. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience* **2**:704–716. DOI: https://doi.org/10.1038/35094565, PMID: 11584308
**Fairhall AL**, Lewen GD, Bialek W, de Ruyter Van Steveninck RR. 2001. Efficiency and ambiguity in an adaptive neural code. *Nature* **412**:787–792. DOI: https://doi.org/10.1038/35090500, PMID: 11518957
**Franks KM**, Russo MJ, Sosulski DL, Mulligan AA, Siegelbaum SA, Axel R. 2011. Recurrent circuitry dynamically shapes the activation of piriform cortex. *Neuron* **72**:49–56. DOI: https://doi.org/10.1016/j.neuron.2011.08.020, PMID: 21982368
**Freedman DJ**, Riesenhuber M, Poggio T, Miller EK. 2001. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**:312–316. DOI: https://doi.org/10.1126/science.291.5502.312, PMID: 11209083
**Froemke RC**, Merzenich MM, Schreiner CE. 2007. A synaptic memory trace for cortical receptive field plasticity. *Nature* **450**:425–429. DOI: https://doi.org/10.1038/nature06289, PMID: 18004384
**Funahashi S**, Bruce CJ, Goldman-Rakic PS. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* **61**:331–349. DOI: https://doi.org/10.1152/jn.1989.61.2.331, PMID: 2918358
**Gillary G**, Heydt R, Niebur E. 2017. Short-term depression and transient memory in sensory cortex. *Journal of Computational Neuroscience* **43**:273–294. DOI: https://doi.org/10.1007/s10827-017-0662-8, PMID: 29027605
**Gillett M**, Pereira U, Brunel N. 2020. Characteristics of sequential activity in networks with temporally asymmetric Hebbian learning. *PNAS* **117**:29948–29958. DOI: https://doi.org/10.1073/pnas.1918674117, PMID: 33177232
**Gjoni E**, Zenke F, Bouhours B, Schneggenburger R. 2018. Specific synaptic input strengths determine the computational properties of excitation-inhibition integration in a sound localization circuit. *The Journal of Physiology* **596**:4945–4967. DOI: https://doi.org/10.1113/JP276012, PMID: 30051910
**Goldman MS**. 2009. Memory without feedback in a neural network. *Neuron* **61**:621–634. DOI: https://doi.org/10.1016/j.neuron.2008.12.012, PMID: 19249281
**Harris KD**, Csicsvari J, Hirase H, Dragoi G, Buzsáki G. 2003. Organization of cell assemblies in the hippocampus. *Nature* **424**:552–556. DOI: https://doi.org/10.1038/nature01834, PMID: 12891358
**Harris KD**. 2005. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience* **6**:399–407. DOI: https://doi.org/10.1038/nrn1669, PMID: 15861182
**Hebb DO**. 1949. The Organization of Behavior: A Neuropsychological Theory. Wiley.
**Hennequin G**, Vogels TP, Gerstner W. 2012. Non-normal amplification in random balanced neuronal networks. *Physical Review E* **86**:1–12. DOI: https://doi.org/10.1103/PhysRevE.86.011909, PMID: 23005454
**Hennequin G**, Vogels TP, Gerstner W. 2014. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron* **82**:1394–1406. DOI: https://doi.org/10.1016/j.neuron.2014.04.045, PMID: 24945778
**Hennequin G**, Agnes EJ, Vogels TP. 2017. Inhibitory Plasticity: Balance, Control, and Codependence. *Annual Review of Neuroscience* **40**:557–579. DOI: https://doi.org/10.1146/annurev-neuro-072116-031005, PMID: 28598717
**Hennequin G**, Ahmadian Y, Rubin DB, Lengyel M, Miller KD. 2018. The Dynamical regime of sensory cortex: stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron* **98**:846–860. DOI: https://doi.org/10.1016/j.neuron.2018.04.017, PMID: 29772203
**Hopfield JJ**. 1982. Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**:2554–2558. DOI: https://doi.org/10.1073/pnas.79.8.2554, PMID: 6953413
**Horn RA**, Johnson CR. 1985. Matrix Analysis. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511810817
**Irving RS**. 2004. Integers, Polynomials, and Rings: A Course in Algebra. Springer. DOI: https://doi.org/10.1007/b97633
**Ji XY**, Zingg B, Mesik L, Xiao Z, Zhang LI, Tao HW. 2016. Thalamocortical Innervation Pattern in Mouse Auditory and Visual Cortex: Laminar and Cell-Type Specificity. *Cerebral Cortex* **26**:2612–2625. DOI: https://doi.org/10.1093/cercor/bhv099, PMID: 25979090
**Josselyn SA**, Tonegawa S. 2020. Memory engrams: Recalling the past and imagining the future. *Science* **367**:eaaw4325. DOI: https://doi.org/10.1126/science.aaw4325, PMID: 31896692
**Ko H**, Hofer SB, Pichler B, Buchanan KA, Sjöström PJ, Mrsic-Flogel TD. 2011. Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**:87–91. DOI: https://doi.org/10.1038/nature09880, PMID: 21478872
**Kraynyukova N**, Tchumatchenko T. 2018. Stabilized supralinear network can give rise to bistable, oscillatory, and persistent activity. *PNAS* **115**:3464–3469. DOI: https://doi.org/10.1073/pnas.1700080115, PMID: 29531035

**Large AM**, Vogler NW, Mielo S, Oswald AMM. 2016. Balanced feedforward inhibition and dominant recurrent inhibition in olfactory cortex. *PNAS* **113**:2276–2281. DOI: https://doi.org/10.1073/pnas.1519295113, PMID: 26858458

**Levina A**, Herrmann JM, Geisel T. 2007. Dynamical synapses causing self-organized criticality in neural networks. *Nature Physics* **3**:857–860. DOI: https://doi.org/10.1038/nphys758

**Litwin-Kumar A**, Doiron B. 2012. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature Neuroscience* **15**:1498–1505. DOI: https://doi.org/10.1038/nn.3220, PMID: 23001062

**Loebel A**, Tsodyks M. 2002. Computation by ensemble synchronization in recurrent networks with synaptic depression. *Journal of Computational Neuroscience* **13**:111–124. DOI: https://doi.org/10.1023/a:1020110223441, PMID: 12215725

**Loebel A**, Nelken I, Tsodyks M. 2007. Processing of sounds by population spikes in a model of primary auditory cortex. *Frontiers in Neuroscience* **1**:197–209. DOI: https://doi.org/10.3389/neuro.01.1.1.015.2007, PMID: 18982129

**Markram H**, Wang Y, Tsodyks M. 1998. Differential signaling via the same axon of neocortical pyramidal neurons. *PNAS* **95**:5323–5328. DOI: https://doi.org/10.1073/pnas.95.9.5323, PMID: 9560274

**Marshel JH**, Kim YS, Machado TA, Quirin S, Benson B, Kadmon J, Raja C, Chibukhchyan A, Ramakrishnan C, Inoue M, Shane JC, McKnight DJ, Yoshizawa S, Kato HE, Ganguli S, Deisseroth K. 2019. Cortical layer-specific critical dynamics triggering perception. *Science* **365**:eaaw5202. DOI: https://doi.org/10.1126/science.aaw5202, PMID: 31320556

**Mazor O**, Laurent G. 2005. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* **48**:661–673. DOI: https://doi.org/10.1016/j.neuron.2005.09.032, PMID: 16301181

**Mazzucato L**, La Camera G, Fontanini A. 2019. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nature Neuroscience* **22**:787–796. DOI: https://doi.org/10.1038/s41593-019-0364-9, PMID: 30936557

**Miller KD**, Palmigiano A. 2020. Generalized Paradoxical Effects in Excitatory/Inhibitory Networks. [bioRxiv]. DOI: https://doi.org/10.1101/2020.10.13.336727

**Miska NJ**, Richter LM, Cary BA, Gjorgjieva J, Turrigiano GG. 2018. Sensory experience inversely regulates feedforward and feedback excitation-inhibition ratio in rodent visual cortex. *eLife* **7**:e38846. DOI: https://doi.org/10.7554/eLife.38846, PMID: 30311905

**Mongillo G**, Barak O, Tsodyks M. 2008. Synaptic theory of working memory. *Science* **319**:1543–1546. DOI: https://doi.org/10.1126/science.1150769, PMID: 18339943

**Mongillo G**, Hansel D, van Vreeswijk C. 2012. Bistability and spatiotemporal irregularity in neuronal networks with nonlinear synaptic transmission. *Physical Review Letters* **108**:158101. DOI: https://doi.org/10.1103/PhysRevLett.108.158101, PMID: 22587287

**Murphy BK**, Miller KD. 2009. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* **61**:635–648. DOI: https://doi.org/10.1016/j.neuron.2009.02.005, PMID: 19249282

**Niessing J**, Friedrich RW. 2010. Olfactory pattern classification by discrete neuronal network states. *Nature* **465**:47–52. DOI: https://doi.org/10.1038/nature08961, PMID: 20393466

**Okun M**, Lampl I. 2008. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature Neuroscience* **11**:535–537. DOI: https://doi.org/10.1038/nn.2105, PMID: 18376400

**Ozeki H**, Finn IM, Schaffer ES, Miller KD, Ferster D. 2009. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* **62**:578–592. DOI: https://doi.org/10.1016/j.neuron.2009.03.028, PMID: 19477158

**Pala A**, Petersen CCH. 2015. In vivo measurement of cell-type-specific synaptic connectivity and synaptic transmission in layer 2/3 mouse barrel cortex. *Neuron* **85**:68–75. DOI: https://doi.org/10.1016/j.neuron.2014.11.025, PMID: 25543458

**Peron S**, Pancholi R, Voelcker B, Wittenbach JD, Ólafsdóttir HF, Freeman J, Svoboda K. 2020. Recurrent interactions in local cortical circuits. *Nature* **579**:256–259. DOI: https://doi.org/10.1038/s41586-020-2062-x, PMID: 32132709

**Plenz D**, Thiagarajan TC. 2007. The organizing principles of neuronal avalanches: cell assemblies in the cortex? *Trends in Neurosciences* **30**:101–110. DOI: https://doi.org/10.1016/j.tins.2007.01.005, PMID: 17275102

**Ponce-Alvarez A**, Thiele A, Albright TD, Stoner GR, Deco G. 2013. Stimulus-dependent variability and noise correlations in cortical MT neurons. *PNAS* **110**:13162–13167. DOI: https://doi.org/10.1073/pnas.1300098110, PMID: 23878209

**Pozzorini C**, Naud R, Mensi S, Gerstner W. 2013. Temporal whitening by power-law adaptation in neocortical neurons. *Nature Neuroscience* **16**:942–948. DOI: https://doi.org/10.1038/nn.3431, PMID: 23749146

**Priebe NJ**, Mechler F, Carandini M, Ferster D. 2004. The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature Neuroscience* **7**:1113–1122. DOI: https://doi.org/10.1038/nn1310, PMID: 15338009

**Romo R**, Brody CD, Hernández A, Lemus L. 1999. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**:470–473. DOI: https://doi.org/10.1038/20939, PMID: 10365959

**Rost T**, Deger M, Nawrot MP. 2018. Winnerless competition in clustered balanced networks: inhibitory assemblies do the trick. *Biological Cybernetics* **112**:81–98. DOI: https://doi.org/10.1007/s00422-017-0737-7, PMID: 29075845

**Rubin DB**, Van Hooser SD, Miller KD. 2015. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* **85**:402–417. DOI: https://doi.org/10.1016/j.neuron.2014.12.026, PMID: 25611511

**Rubin R**, Abbott LF, Sompolinsky H. 2017. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *PNAS* **114**:E9366–E9375. DOI: https://doi.org/10.1073/pnas.1705841114, PMID: 29042519

**Rupprecht P**, Friedrich RW. 2018. Precise Synaptic Balance in the Zebrafish Homolog of Olfactory Cortex. *Neuron* **100**:669–683. DOI: https://doi.org/10.1016/j.neuron.2018.09.013, PMID: 30318416

**Sadeh S**, Clopath C. 2021. Inhibitory stabilization and cortical computation. *Nature Reviews Neuroscience* **22**:21–37. DOI: https://doi.org/10.1038/s41583-020-00390-z, PMID: 33177630

**Sanzeni A**, Akitake B, Goldbach HC, Leedy CE, Brunel N, Histed MH. 2020. Inhibition stabilization is a widespread property of cortical networks. *eLife* **9**:e54875. DOI: https://doi.org/10.7554/eLife.54875, PMID: 32598278

**Schulz A**, Miehl C, Berry MJ, Gjorgjieva J. 2021. The generation of cortical novelty responses through inhibitory plasticity. *eLife* **10**:e65309. DOI: https://doi.org/10.7554/eLife.65309, PMID: 34647889

**Shew WL**, Clawson WP, Pobst J, Karimipanah Y, Wright NC, Wessel R. 2015. Adaptation to sensory input tunes visual cortex to criticality. *Nature Physics* **11**:659–663. DOI: https://doi.org/10.1038/nphys3370

**Stern M**, Bolding KA, Abbott LF, Franks KM. 2018. A transformation from temporal to ensemble coding in a model of piriform cortex. *eLife* **7**:e34831. DOI: https://doi.org/10.7554/eLife.34831, PMID: 29595470

**Stopfer M**, Bhagavan S, Smith BH, Laurent G. 1997. Impaired odour discrimination on desynchronization of odour-encoding neural assemblies. *Nature* **390**:70–74. DOI: https://doi.org/10.1038/36335, PMID: 9363891

**Thorpe S**, Fize D, Marlot C. 1996. Speed of processing in the human visual system. *Nature* **381**:520–522. DOI: https://doi.org/10.1038/381520a0, PMID: 8632824

**Tsodyks MV**, Markram H. 1997. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *PNAS* **94**:719–723. DOI: https://doi.org/10.1073/pnas.94.2.719, PMID: 9012851

**Tsodyks MV**, Skaggs WE, Sejnowski TJ, McNaughton BL. 1997. Paradoxical effects of external modulation of inhibitory interneurons. *The Journal of Neuroscience* **17**:4382–4388. DOI: https://doi.org/10.1523/JNEUROSCI.17-11-04382.1997, PMID: 9151754

**van Vreeswijk C**, Sompolinsky H. 1996. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**:1724–1726. DOI: https://doi.org/10.1126/science.274.5293.1724, PMID: 8939866

**van Vreeswijk C**, Hansel D. 2001. Patterns of synchrony in neural networks with spike adaptation. *Neural Computation* **13**:959–992. DOI: https://doi.org/10.1162/08997660151134280, PMID: 11359640

**Varela JA**, Sen K, Gibson J, Fost J, Abbott LF, Nelson SB. 1997. A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. *The Journal of Neuroscience* **17**:7926–7940. DOI: https://doi.org/10.1523/JNEUROSCI.17-20-07926.1997, PMID: 9315911

**Vinje WE**, Gallant JL. 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**:1273–1276. DOI: https://doi.org/10.1126/science.287.5456.1273, PMID: 10678835

**Vogels TP**, Sprekeler H, Zenke F, Clopath C, Gerstner W. 2011. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**:1569–1573. DOI: https://doi.org/10.1126/science.1211095, PMID: 22075724

**Wehr M**, Zador AM. 2003. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**:442–446. DOI: https://doi.org/10.1038/nature02116, PMID: 14647382

**Wong KF**, Wang XJ. 2006. A recurrent network mechanism of time integration in perceptual decisions. *The Journal of Neuroscience* **26**:1314–1328. DOI: https://doi.org/10.1523/JNEUROSCI.3733-05.2006, PMID: 16436619

**Wu YK**, Hengen KB, Turrigiano GG, Gjorgjieva J. 2020. Homeostatic mechanisms regulate distinct aspects of cortical circuit dynamics. *PNAS* **117**:24514–24525. DOI: https://doi.org/10.1073/pnas.1918368117, PMID: 32917810

**Wu YK**. 2021. Nonlinear Transient Amplification. swh:1:rev:6ff6ff10b9f4994a0f948a987a66cc82f98451e1. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:bcdf9ea652853e2f06b154735e693fa154943072;origin=https://github.com/fmi-basel/gzenke-nonlinear-transient-amplification;visit=swh:1:snp:6b78c97e1671014d095bb0db7eaa97aaebe0086b;anchor=swh:1:rev:6ff6ff10b9f4994a0f948a987a66cc82f98451e1

**Yakovlev V**, Fusi S, Berman E, Zohary E. 1998. Inter-trial neuronal activity in inferior temporal cortex: a putative vehicle to generate long-term visual associations. *Nature Neuroscience* **1**:310–317. DOI: https://doi.org/10.1038/1131, PMID: 10195165

**Zenke F**, Agnes EJ, Gerstner W. 2015. Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature Communications* **6**:1–13. DOI: https://doi.org/10.1038/ncomms7922, PMID: 25897632

**Znamenskiy P**, Kim HM, Muir DR, Iacaruso MC, Hofer SB, Mrsic-Flogel TD. 2018. Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. [bioRxiv]. DOI: https://doi.org/10.1101/294835

**Zucker RS**, Regehr WG. 2002. Short-term synaptic plasticity. *Annual Review of Physiology* **64**:355–405. DOI: https://doi.org/10.1146/annurev.physiol.64.092501.114547, PMID: 11826273

# Appendix 1

## Stability conditions in networks with E-to-I STF

The dynamics of supralinear networks with E-to-I STF can be described as follows:

$$\tau_E \frac{dr_E}{dt} = -r_E + \left[J_{EE} r_E - J_{EI} r_I + g_E\right]_+^{\alpha_E} \tag{102}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[u J_{IE} r_E - J_{II} r_I + g_I\right]_+^{\alpha_I} \tag{103}$$

$$\frac{du}{dt} = \frac{1-u}{\tau_u} + U_f(U_{max} - u) r_E \tag{104}$$

The Jacobian $\mathbf{M_{STF}}$ of the system with E-to-I STF is given by:

$$\mathbf{M_{STF}} = \begin{bmatrix} \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & -\tau_E^{-1} J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} & 0 \\ \tau_I^{-1} u J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} & -\tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) & \tau_I^{-1} J_{IE} r_E \alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \\ U_f(U_{max} - u) & 0 & -\tau_u^{-1} - U_f r_E \end{bmatrix} \tag{105}$$

The characteristic polynomial for the system with E-to-I STF can be written as follows:

$$\lambda^3 - \mathrm{tr}(\mathbf{M_{STF}})\lambda^2 + (A_{11} + A_{22} + A_{33})\lambda - \det(\mathbf{M_{STF}}) = 0 \tag{106}$$

where $\mathrm{tr}(\mathbf{M_{STF}})$ and $\det(\mathbf{M_{STF}})$ are the trace and the determinant of the Jacobian matrix $\mathbf{M_{SFA}}$, $A_{11}$, $A_{22}$, and $A_{33}$ are the matrix cofactors. More specifically,

$$\begin{aligned} \mathrm{tr}(\mathbf{M_{STF}}) &= \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) - \tau_u^{-1} - U_f r_E \\ &\propto \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} / r_I^{\frac{\alpha_I-1}{\alpha_I}} - r_I^{\frac{1-\alpha_I}{\alpha_I}}) - \tau_I^{-1}(r_I^{\frac{1-\alpha_I}{\alpha_I}} + J_{II}\alpha_I) - \tau_u^{-1} r_I^{\frac{1-\alpha_I}{\alpha_I}} - U_f r_E r_I^{\frac{1-\alpha_I}{\alpha_I}} \end{aligned} \tag{107}$$

Assuming that $\alpha_E = \alpha_I = \alpha$, we then have

$$\mathrm{tr}(\mathbf{M_{STF}}) \propto \tau_E^{-1}\left[J_{EE}\alpha\left(\frac{r_E}{r_I}\right)^{\frac{\alpha-1}{\alpha}} - r_I^{\frac{1-\alpha}{\alpha}}\right] - \tau_I^{-1}(r_I^{\frac{1-\alpha}{\alpha}} + J_{II}\alpha) - \tau_u^{-1} r_I^{\frac{1-\alpha}{\alpha}} - U_f r_E r_I^{\frac{1-\alpha}{\alpha}} \tag{108}$$

Substituting the firing rates with the current into excitatory population $z$, we then have

$$\begin{aligned} \mathrm{tr}(\mathbf{M_{STF}}) \propto & \tau_E^{-1}\left[J_{EE}\alpha\left(\frac{z}{\det(\mathbf{J_{STF}}) \cdot J_{EI}^{-1}[z]_+^{\alpha} + J_{EI}^{-1} J_{II} z - J_{EI}^{-1} J_{II} g_E + g_I}\right)^{\alpha-1} - r_I^{\frac{1-\alpha}{\alpha}}\right] \\ & - \tau_I^{-1}(r_I^{\frac{1-\alpha}{\alpha}} + J_{II}\alpha) - \tau_u^{-1} r_I^{\frac{1-\alpha}{\alpha}} - U_f r_E r_I^{\frac{1-\alpha}{\alpha}} \end{aligned} \tag{109}$$

$$\det(\mathbf{J_{STF}}) = \begin{vmatrix} J_{EE} & -J_{EI} \\ u J_{IE} & -J_{II} \end{vmatrix} = -J_{EE} J_{II} + u J_{IE} J_{EI} \tag{110}$$

In the large $r_E$ limit, $z$ is large, $\lim_{r_E\to\infty} u = \lim_{r_E\to\infty} \frac{1+U_f U_{max} r_E \tau_u}{1+U_f r_E \tau_u} \approx U_{max}$. Therefore, we can guarantee that $\det(\mathbf{J_{STF}})$ becomes positive for sufficiently large $U_{max}$. Since the denominator $\det(\mathbf{J_{STF}}) \cdot J_{EI}^{-1}[z]_+^{\alpha} + J_{EI}^{-1} J_{II} z - J_{EI}^{-1} J_{II} g_E + g_I$ grows faster than the numerator for $z \gg 1$, $\mathrm{tr}(\mathbf{M_{STF}})$ becomes negative for large $r_E$.

$$\begin{aligned} A_{11} + A_{22} + A_{33} =& \tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})(\tau_u^{-1} + U_f r_E) \\ & + \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)(-\tau_u^{-1} - U_f r_E) \\ & - \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) + \tau_E^{-1} J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} \tau_I^{-1} u J_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \end{aligned} \tag{111}$$

Similarly, in the large $r_E$ limit, $A_{11} + A_{22} + A_{33}$ is positive.

$$\begin{aligned}\det(\mathbf{M_{STF}}) =& \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)\tau_I^{-1}(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}})(\tau_u^{-1} + U_f r_E) \\ &+ \tau_E^{-1}J_{EI}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}}(\tau_I^{-1}uJ_{IE}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}(-\tau_u^{-1} - U_f r_E) - \tau_I^{-1}J_{IE}r_E\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}U_f(U_{max} - u))\end{aligned} \tag{112}$$

Similarly, in the large $r_E$ limit, $\det(\mathbf{M_{STF}})$ is negative.
Therefore, similar to E-to-E STD, networks dynamics can also be stabilized by E-to-I STF.

# Appendix 2

## Conditions for ISN in networks with E-to-I STF

Here, we identify the condition of being ISN in supralinear networks with E-to-I STF. If inhibition is frozen, in other words, if feedback inhibition is absent, the Jacobian of the system becomes as follows:

$$\mathbf{M_2} = \begin{bmatrix} \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) & 0 \\ U_f(U_{max} - u) & -\tau_u^{-1} - U_f r_E \end{bmatrix} \tag{113}$$

For the system with frozen inhibition, the dynamics are stable if

$$\mathrm{tr}(\mathbf{M_2}) = \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) - \tau_u^{-1} - U_f r_E < 0 \tag{114}$$

and

$$\det(\mathbf{M_2}) = \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)(-\tau_u^{-1} - U_f r_E) > 0 \tag{115}$$

Therefore, if the network is an ISN at the fixed point, the following condition has to be satisfied:

$$\tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) > 0 \tag{116}$$

Note that this condition is independent of the facilitation variable $u$ of E-to-I STF. We further define the ISN index for the system with E-to-I STF as follows:

$$\text{ISN index} = \tau_E^{-1}(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1) \tag{117}$$

# Appendix 3

## Conditions for paradoxical response in networks with E-to-I STF

Next, we identify the condition of having the paradoxical effect in supralinear networks with E-to-I STF. The excitatory nullcline is defined by

$$\tau_E \frac{dr_E}{dt} = -r_E + \left[J_{EE} r_E - J_{EI} r_I + g_E\right]_+^{\alpha_E} = 0 \tag{118}$$

For $r_{E,I} > 0$, we have

$$r_I = \frac{J_{EE} r_E - r_E^{\frac{1}{\alpha_E}} + g_E}{J_{EI}} \tag{119}$$

The slope of the excitatory nullcline in the $r_E/r_I$ plane where $x$ axis is $r_E$ and $y$ axis is $r_I$ can be written as follows

$$k_{STF}^E = \frac{J_{EE} - \frac{1}{\alpha_E} r_E^{\frac{1}{\alpha_E}-1}}{J_{EI}} \tag{120}$$

Note that the slope of the excitatory nullcline is nonlinear. To have paradoxical effect, the slope of the excitatory nullcline at the fixed point of the system has to be positive. We therefore have

$$J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1 > 0 \tag{121}$$

We exploit a separation of timescales between fast neural activity and slow short-term plasticity variable, we therefore set the facilitation variable to the value at its fixed point corresponding to the dynamical value of $r_E$. Then we can write the inhibitory nullcline as follows

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[\frac{1 + U_f U_{max} r_E \tau_u}{1 + U_f r_E \tau_u} J_{IE} r_E - J_{II} r_I + g_I\right]_+^{\alpha_I} = 0 \tag{122}$$

In the region of rates $r_{E,I} > 0$, we have

$$r_I = \frac{\frac{1+U_f U_{max} r_E \tau_u}{1+U_f r_E \tau_u} J_{IE} r_E - r_I^{\frac{1}{\alpha_I}} + g_I}{J_{II}} \tag{123}$$

The slope of the inhibitory nullcline can be written as follows

$$k_{STF}^I = \frac{\frac{1+U_f U_{max} r_E \tau_u}{1+U_f r_E \tau_u} J_{IE} + \frac{U_f U_{max} \tau_u - U_f \tau_u}{(1+U_f r_E \tau_u)^2} J_{IE} r_E}{J_{II} + \frac{1}{\alpha_I} r_I^{\frac{1-\alpha_I}{\alpha_I}}} \tag{124}$$

In addition to the positive slope of the excitatory nullcline, the slope of the inhibitory nullcline at the fixed point of the system has to be larger than the slope of the excitatory nullcline. We therefore have

$$\begin{aligned} -(J_{EE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} - 1)(1 + J_{II}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}}) + J_{IE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} \frac{1 + U_f U_{max} r_E \tau_u}{1 + U_f r_E \tau_u} J_{EI}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} \\ + J_{IE}\alpha_E r_E^{\frac{\alpha_E-1}{\alpha_E}} \frac{U_f U_{max} \tau_u - U_f \tau_u}{(1 + U_f r_E \tau_u)^2} J_{EI}\alpha_I r_I^{\frac{\alpha_I-1}{\alpha_I}} r_E > 0 \end{aligned} \tag{125}$$

The above condition is the same as the stability condition of the determinant of the Jacobian of the system with E-to-I STF (***Eq. (112)***). Therefore, the condition is always satisfied when the system with E-to-I STF is stable.

Note that the condition of being ISN shown in ***Eq. (116)*** is identical to the condition of having paradoxical effect shown in ***Eq. (121)***. Therefore, in networks with E-to-I STF alone, paradoxical

effect implies ISN and ISN implies paradoxical effect. We thus use paradoxical effect as a proxy for inhibitory stabilization.

# Appendix 4

## Change in steady-state activity of unstimulated co-tuned neurons

To analyze the pattern completion in supralinear networks, we considered a network with one excitatory population and one inhibitory population. Neurons in the excitatory population are co-tuned to the same stimulus feature and are separated into two subsets denoting by $E_{11}$ and $E_{12}$. The dynamics of the system can be described as follows:

$$\tau_E \frac{dr_{E_{11}}}{dt} = -r_{E_{11}} + \left[J_{E_{11}E_{11}} r_{E_{11}} + J_{E_{11}E_{12}} r_{E_{12}} - J_{E_{11}I} r_I + g_{E_{11}}\right]_+^{\alpha_E} \tag{126}$$

$$\tau_E \frac{dr_{E_{12}}}{dt} = -r_{E_{12}} + \left[J_{E_{12}E_{11}} r_{E_{11}} + J_{E_{12}E_{12}} r_{E_{12}} - J_{E_{12}I} r_I + g_{E_{12}}\right]_+^{\alpha_E} \tag{127}$$

$$\tau_I \frac{dr_I}{dt} = -r_I + \left[J_{IE_{11}} r_{E_{11}} + J_{IE_{12}} r_{E_{12}} - J_{II} r_I + g_I\right]_+^{\alpha_I} \tag{128}$$

The change in the firing rate of the Subset 2 in the excitatory population $\delta r_{E_{12}}$ can be written as a function of the change in the input to the Subset 1 $\delta g_{E_{11}}$:

$$\begin{aligned}\delta r_{E_{12}} &= \frac{1}{\det(\mathbb{1} - \mathbf{FJ})}[-f'_{E_{12}} J_{E_{12}I} f'_I J_{IE_{11}} - (-f'_{E_{12}} J_{E_{12}E_{11}})(1 + f'_I J_{II})] f'_{E_{11}} \delta g_{E_{11}} \\ &= \frac{1}{\det(\mathbb{1} - \mathbf{FJ})}[J_{E_{12}E_{11}} + J_{E_{12}E_{11}} J_{II} f'_I - J_{E_{12}I} J_{IE_{11}} f'_I] f'_{E_{11}} f'_{E_{12}} \delta g_{E_{11}}\end{aligned} \tag{129}$$

where $\mathbb{1}$ is the identity matrix. And $\mathbf{F}$ is given by

$$\mathbf{F} = \begin{bmatrix} f'_{E_{11}} & 0 & 0 \\ 0 & f'_{E_{12}} & 0 \\ 0 & 0 & f'_I \end{bmatrix} \tag{130}$$

where $f'_{E_{11}}$, $f'_{E_{12}}$, and $f'_I$ are the derivatives of the input-output functions evaluated at the fixed point. The term $\det(\mathbb{1} - \mathbf{FJ})$ is positive to ensure the stability of the system.

Clearly, if the term $J_{E_{12}E_{11}} + J_{E_{12}E_{11}} J_{II} f'_I - J_{E_{12}I} J_{IE_{11}} f'_I$ is positive (negative), increasing the input to the Subset 1 leads to an increase (a decrease) in the activity of neurons in the Subset 2. As the input to the Subset 1 increases, the firing rate of the inhibitory population $r_I$ and also $f'_I$ will increase. In the presence of E-to-E STD or E-to-I STF, $J_{E_{12}E_{11}}$ or $J_{IE_{11}}$ will decrease or increase with the input to the Subset 1. As a result, the sign of $J_{E_{12}E_{11}} + J_{E_{12}E_{11}} J_{II} f'_I - J_{E_{12}I} J_{IE_{11}} f'_I$ can switch from positive to negative as the input to the Subset 1 increases, indicating that the effect on the activity of the co-tuned unstimulated neurons in the same ensemble can switch from potentiation to suppression. Note that this behavior is different from linear networks in which the change is independent of the input or firing rates.

# Appendix 5

**Appendix 5—table 1.** Parameters for ***Figure 1—figure supplement 1***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 0.5 | - | E-to-E connection strength |
| $J_{IE}$ | 0.45 | - | E-to-I connection strength |
| $J_{EI}$ | 1.0 | - | I-to-E connection strength |
| $J_{II}$ | 1.5 | - | I-to-I connection strength |
| $g_E^{bs}$ | 0.5 | - | Input to the *E* population at baseline |
| $g_I^{bs}$ | 1.5 | - | Input to the I population at baseline |

**Appendix 5—table 2.** Parameters for ***Figure 2—figure supplement 2***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $g_E^{stim}$ | 2.0 | - | Input to the E population during stimulation |

Note that values of the unlisted parameters are the same as ***Tables 1–2***.

**Appendix 5—table 3.** Parameters for ***Figure 2—figure supplement 3*** SSN example.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 1.8 | - | E-to-E connection strength |
| $J_{IE}$ | 2.0 | - | E-to-I connection strength |
| $J_{EI}$ | 1.0 | - | I-to-E connection strength |
| $J_{II}$ | 1.0 | - | I-to-I connection strength |

**Appendix 5—table 4.** Parameters for ***Figure 2—figure supplement 5***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $g_E^{bs}$ | 1.8 | - | Input to the *E* population at baseline |

Note that values of the unlisted parameters are the same as ***Tables 1–2***.

**Appendix 5—table 5.** Parameters for ***Figure 2—figure supplement 10***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $N_E$ | 400 | - | Number of excitatory neurons |
| $N_I$ | 100 | - | Number of inhibitory neurons |
| $J_{EE}$ | 0.05 | - | E-to-E connection strength |
| $J_{IE}$ | 0.02 | - | E-to-I connection strength |
| $J_{EI}$ | 0.05 | - | I-to-E connection strength |
| $J_{II}$ | 0.03 | - | I-to-I connection strength |

**Appendix 5—table 6.** Parameters for ***Figure 2—figure supplement 11***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 0.9 | - | E-to-E connection strength |
| $J_{IE}$ | 1.2 | - | E-to-I connection strength |

*Appendix 5—table 6 Continued on next page*

*Appendix 5—table 6 Continued*

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EI}$ | 0.5 | - | I-to-E connection strength |
| $J_{II}$ | 0.5 | - | I-to-I connection strength |
| $\tau_E$ | 20 | ms | Time constant of excitatory firing dynamics |
| $\tau_I$ | 60 | ms | Time constant of inhibitory firing dynamics |
| $g_E^{bs}$ | 1.0 | - | Input to the *E* population at baseline |
| $g_E^{stim}$ | 2.0 | - | Input to the *E* population during stimulation |
| $g_I$ | 2.0 | - | Input to the *I* population |

Note that values of the unlisted parameters are the same as ***Tables 1–2***.

**Appendix 5—table 7.** Parameters for ***Figure 2—figure supplement 12***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 1.0 | - | E-to-E connection strength |
| $J_{IE}$ | 1.2 | - | E-to-I connection strength |
| $J_{EI}$ | 0.5 | - | I-to-E connection strength |
| $J_{II}$ | 1.0 | - | I-to-I connection strength |
| $g_E^{bs}$ | 0.5 | - | Input to the *E* population at baseline |
| $g_I^{bs}$ | 1.0 | - | Input to the *I* population at baseline |

**Appendix 5—table 8.** Parameters for ***Figure 3—figure supplement 1*** global inhibition example.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 1.6 | - | Within-ensemble E-to-E connection strength |
| $J_{IE}$ | 1.0 | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | 1.0 | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | 1.2 | - | Within-ensemble I-to-I connection strength |
| $J'_{EE}$ | 0.16 | - | Inter-ensemble E-to-E connection strength |
| $J'_{IE}$ | 1.0 | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | 1.0 | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | 1.2 | - | Inter-ensemble I-to-I connection strength |
| $g_{E1}^{bs}$ | 1.5 | - | Input to the *E*1 population at baseline |
| $g_{E2}$ | 1.5 | - | Input to the *E*2 population |
| $g_{I1}$ | 2.5 | - | Input to the *I*1 population |
| $g_{I2}$ | 2.5 | - | Input to the *I*2 population |
| Parameters for ***Figure 3—figure supplement 1*** co-tuned example | | | |
| $J_{IE}$ | 1.0 * (4/3) | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | 1.0 * (4/3) | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | 1.2 * (4/3) | - | Within-ensemble I-to-I connection strength |

*Appendix 5—table 8 Continued on next page*

*Appendix 5—table 8 Continued*

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J'_{IE}$ | 1.0 * (2/3) | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | 1.0 * (2/3) | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | 1.2 * (2/3) | - | Inter-ensemble I-to-I connection strength |

**Appendix 5—table 9.** Parameters for ***Figure 4—figure supplement 1***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | $1.5/(N_E/2-1)$ | - | Within-ensemble E-to-E connection strength |
| $J_{IE}$ | $1.0/(N_E/2)$ | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | $1.0/(N_I/2)$ | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | $1.0/(N_I/2-1)$ | - | Within-ensemble I-to-I connection strength |
| $J'_{EE}$ | $0.1/(N_E/2-1)$ | - | Inter-ensemble E-to-E connection strength |
| $J'_{IE}$ | $0.3/(N_E/2)$ | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | $0.3/(N_I/2)$ | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | $0.1/(N_I/2)$ | - | Inter-ensemble I-to-I connection strength |
| $g_{E1}^{bs}$ | 1.5 | - | Input to the *E*1 population at baseline |
| $g_{E2}$ | 1.5 | - | Input to the *E*2 population |
| $g_I$ | 2.0 | - | Input to the *I* population |

**Appendix 5—table 10.** Parameters for ***Figure 6—figure supplement 1***.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| $J_{EE}$ | 0.20 | - | Within-ensemble E-to-E connection strength |
| $J_{IE}$ | 0.09 | - | Within-ensemble E-to-I connection strength |
| $J_{EI}$ | 0.10 | - | Within-ensemble I-to-E connection strength |
| $J_{II}$ | 0.10 | - | Within-ensemble I-to-I connection strength |
| $J'_{EE}$ | 0.02 | - | Inter-ensemble E-to-E connection strength |
| $J'_{IE}$ | 0.054 | - | Inter-ensemble E-to-I connection strength |
| $J'_{EI}$ | 0.07 | - | Inter-ensemble I-to-E connection strength |
| $J'_{II}$ | 0.01 | - | Inter-ensemble I-to-I connection strength |

Note that values of the unlisted parameters are the same as ***Table 5***.

# V. Rapid and active stabilization of visual cortical firing rates across light–dark transitions

Pacheco, A. T., Tilden, E. I., Grutzner, S. M., Lane, B. J., Wu, Y., Hengen, K. B., Gjorgjieva, J. & Turrigiano, G. G.. Rapid and active stabilization of visual cortical firing rates across light–dark transitions. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 18068-18077 (2019). https://doi.org/10.1073/pnas.1906595116

# Rapid and active stabilization of visual cortical firing rates across light–dark transitions

Alejandro Torrado Pacheco[a,1], Elizabeth I. Tilden[a,1,2], Sophie M. Grutzner[a,1,3], Brian J. Lane[a], Yue Wu[b], Keith B. Hengen[a,2], Julijana Gjorgjieva[b,c], and Gina G. Turrigiano[a,4]

[a]Department of Biology, Brandeis University, Waltham, MA 02453; [b]Computation in Neural Circuits Group, Max Planck Institute for Brain Research, 60438 Frankfurt, Germany; and [c]School of Life Sciences, Technical University of Munich, 85354 Freising, Germany

Contributed by Gina G. Turrigiano, June 18, 2019 (sent for review April 17, 2019; reviewed by Chinfei Chen and Cristopher M. Niell)

**The dynamics of neuronal firing during natural vision are poorly understood. Surprisingly, mean firing rates of neurons in primary visual cortex (V1) of freely behaving rodents are similar during prolonged periods of light and darkness, but it is unknown whether this reflects a slow adaptation to changes in natural visual input or insensitivity to rapid changes in visual drive. Here, we use chronic electrophysiology in freely behaving rats to follow individual V1 neurons across many dark–light (D-L) and light–dark (L-D) transitions. We show that, even on rapid timescales (1 s to 10 min), neuronal activity was only weakly modulated by transitions that coincided with the expected 12-/12-h L-D cycle. In contrast, a larger subset of V1 neurons consistently responded to unexpected L-D and D-L transitions, and disruption of the regular L-D cycle with 60 h of complete darkness induced a robust increase in V1 firing on reintroduction of visual input. Thus, V1 neurons fire at similar rates in the presence or absence of natural stimuli, and significant changes in activity arise only transiently in response to unexpected changes in the visual environment. Furthermore, although mean rates were similar in light and darkness, pairwise correlations were significantly stronger during natural vision, suggesting that information about natural scenes in V1 may be more strongly reflected in correlations than individual firing rates. Together, our findings show that V1 firing rates are rapidly and actively stabilized during expected changes in visual input and are remarkably stable at both short and long timescales.**

visual experience | rodent vision | visual cortex | firing-rate stability

Neurons in the cerebral cortex are spontaneously active, but the function of this internally generated activity is largely unexplained. Ongoing activity has been proposed to be noise due to random fluctuations (1–3). However, other experiments have shown that spontaneous activity possesses coherent spatiotemporal structure (4–6), suggesting that it may play an important role in the processing of natural sensory stimuli (4, 7–11). In primary visual cortex (V1), spontaneous activity observed in complete darkness is similar to that evoked by visual stimulation with random noise stimuli and is only subtly modulated by natural scene viewing (8, 12). Recently, we showed that individual V1 neurons have very stable mean firing rates in freely behaving rodents and that these mean rates are indistinguishable in light and dark when averaged across many hours (13). How V1 firing can be stable across such drastic changes in the visual environment while still meaningfully encoding sensory stimuli and whether this stability is actively maintained or simply arises from intrinsic circuit dynamics remain unknown.

Regulation of individual firing rates around a stable set point is thought to be essential for proper functioning of cortical circuits in the face of developmental or experience-dependent perturbations to connectivity (14, 15). Long-term stability of individual mean firing rates has now been observed in rodent V1 (13, 16, 17) and primary motor cortex (18), suggesting that it is a general feature of neocortical networks; furthermore, perturbing V1 firing rates through prolonged sensory deprivation results in a slow but precise homeostatic regulation of firing back to an individual set point, showing that neurons actively maintain these set points over long timescales (13). This stability in mean firing rates, even across periods of light and dark, raises the question of how natural visual input is encoded by V1 activity in freely behaving animals. One possibility is that changes in visual drive result in rapid fluctuations in mean firing rates that operate over seconds to minutes. Another possibility is that firing rates are stabilized even over these short timescales, and visual information is primarily encoded in higher-order network dynamics.

To generate insight into these questions, we followed firing of individual neurons in V1 of freely behaving young rats over several days as animals experienced normal light–dark (L-D) and dark–light (D-L) transitions or transitions that were unexpectedly imposed. We found that expected transitions had a very modest effect on firing rates of both excitatory and inhibitory neurons, even when examined immediately around the time of the transition. Population activity did not change significantly across these transitions; when examined at the level of individual neurons, only a small subset (~15%) of putative excitatory neurons consistently responded and then, only during D-L transitions when animals were awake. Interestingly, randomly timed transitions throughout the L-D cycle elicited more consistent responses across sleep–wake states and at both D-L and L-D transitions, and robust and widespread responses to D-L transitions could be unmasked by exposing animals to prolonged darkness for 60 h. These results suggest that the stability normally observed at expected (circadian) L-D and D-L transitions reflects an active process of stabilization. Finally, although

## Significance

**The firing dynamics of neurons in primary visual cortex (V1) are poorly understood. Indeed, V1 neurons of freely behaving rats fire at the same mean rate in light and darkness. It is unclear how this stability is maintained and whether it is important for sensory processing. We find that transitions between light and darkness happening at expected times have only modest effects on V1 activity. In contrast, both unexpected transitions and light reexposure after extended darkness robustly increase V1 firing. Finally, pairwise correlations in neuronal spiking are significantly higher during the light when natural vision is occurring. These data show that V1 firing rates are actively stabilized while simultaneously allowing for input-dependent changes in correlations between neurons.**

Author contributions: A.T.P., E.I.T., S.M.G., B.J.L., J.G., and G.G.T. designed research; A.T.P., E.I.T., S.M.G., B.J.L., and K.B.H. performed research; K.B.H. contributed new reagents/analytic tools; A.T.P., E.I.T., S.M.G., B.J.L., and Y.W. analyzed data; and A.T.P., Y.W., J.G., and G.G.T. wrote the paper.

Reviewers: C.C., Children's Hospital; and C.M.N., University of Oregon.

The authors declare no conflict of interest.



[1]A.T.P., E.I.T., and S.M.G. contributed equally to this work.

[2]Present address: Department of Biology, Washington University in St. Louis, St. Louis, MO 63130.

[3]Present address: Department of Biology, Stanford University, Stanford, CA 94305.

[4]To whom correspondence may be addressed. Email: turrigiano@brandeis.edu.

Published online July 31, 2019.

mean rates were very similar in light and dark, the pairwise correlations between simultaneously recorded neurons were significantly higher in the light than in the dark, even when controlling for behavioral state. Together, our findings show that firing rates in V1 are actively stabilized as animals navigate dramatic changes in the visual environment. This is in contrast to the correlational structure of V1 activity, which more closely tracks visual drive.

## Results

Neurons in V1 maintain remarkably similar mean firing rates during extended periods of light and dark, but how L-D transitions affect firing on more rapid timescales in freely viewing and behaving animals is unclear. Here, we use chronic in vivo electrophysiological recordings from freely behaving rats to closely examine the activity of V1 neurons at D-L and L-D transitions in regular (12/12 h) and manipulated L-D cycles and during unexpected transitions. Using previously established methods (13), we follow individual neurons over time and across multiple light transitions. This approach allows us to analyze the dynamics of neuronal activity at different timescales in response to the appearance or disappearance of natural visual stimuli.

**The Appearance or Disappearance of Natural Visual Stimuli Has only a Modest Effect on the Mean Firing Rates of V1 Neurons.** The firing rates of V1 neurons recorded in freely behaving young rats in light and dark are very similar when averaged in 12-h periods (13). Here, we combine previously and newly acquired datasets and set out to analyze the activity of V1 neurons around the transition from presence to absence of visual input (L-D) and vice versa (D-L) (Fig. 1*A*). Recorded neurons were classified as regular spiking units (RSUs; $n = 96$) or fast-spiking (FS) cells ($n = 32$) based on waveform shape and according to established criteria (Fig. 1*B*) (16, 19). These populations are mostly composed of excitatory pyramidal neurons (RSU) and inhibitory parvalbumin-containing interneurons (FS) (20, 21).

As rats experience L-D or D-L transitions, most neurons showed little change in firing (Fig. 1*C*). We treated each transition as a separate trial and estimated the firing rate for each cell as the average of the perievent time histogram centered on the transition. We first aimed to compare activity at the population level in different stimulus conditions. To this end, we determined whether the distributions of mean firing rates averaged over 10 min on either side of the L-D and D-L transitions were similar to each other. Cumulative distributions in light and dark were indistinguishable for both RSU and FS cells in all conditions (Fig. 1 *D* and *E*) (2-sample Kolmogorov–Smirnov test; RSU, L-D: $P = 0.88$; D-L: $P = 0.99$; FS, L-D: $P = 0.99$; D-L: $P = 1.0$). Similarly, when we compared the distributions using a Wilcoxon rank sum test, we found no difference between the distributions of mean firing rates before vs. after the transitions (Wilcoxon rank sum test; RSU, L-D: $P = 0.677$; D-L: $P = 0.655$; FS, L-D: $P = 0.905$; D-L: $P = 0.827$).

Next, we took advantage of our ability to follow individual neurons across transitions to examine the data in a paired manner, where the firing rate of each neuron was compared before and after the transition. For each neuron, we computed mean firing rate in the 10 min before and after the transition time and averaged across transitions of the same type to estimate the average effect on individual neuronal firing. This analysis revealed a small but consistent change in mean RSU firing rates across both

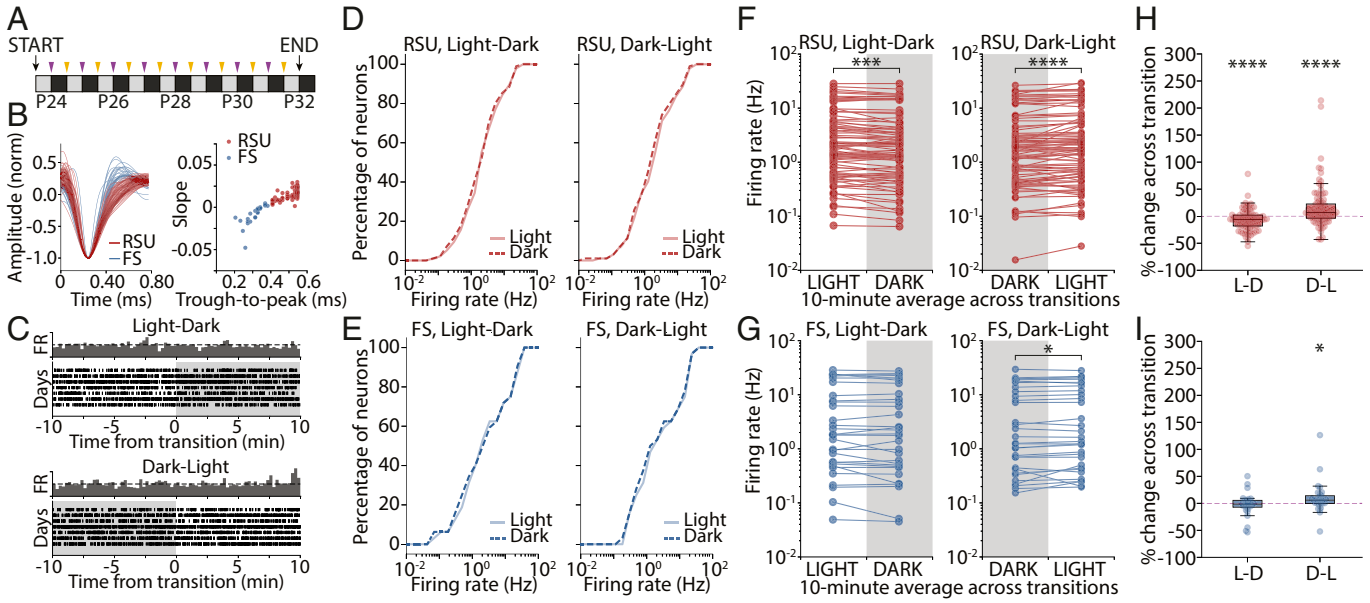


**Fig. 1.** V1 firing rates are largely stable across circadian L-D and D-L transitions. (*A*) Experimental protocol. Single-unit recordings were obtained from juvenile rats for a continuous 9-d period (P24 to P32). Throughout this period, animals were kept in a regular 12-/12-h L-D cycle and thus, underwent L-D (purple arrows) and D-L (yellow arrows) transitions at regular 12-h intervals. (*B*, *Left*) Average waveform for each continuously recorded unit identified as RSU (red) or FS cell (blue). (*B*, *Right*) Plot of trough-to-peak time vs. waveform slope 0.4 ms after trough reveals the bimodal distribution used to classify recorded units as RSU or FS. (*C*) Example raster plot of spiking activity for a recorded unit across several days, showing 20 min of activity centered on the L-D (*Upper*) and D-L (*Lower*) transitions. Dark bars represent the perievent time histogram obtained by averaging across days. (*D*) Cumulative distributions of RSU firing rates averaged over the 10 min of light (solid lines) or dark (dashed lines) around the transitions for L-D (*Left*) and D-L (*Right*) transitions (L-D, $P = 0.875$; D-L, $P = 0.99$; 2-sample Kolmogorov–Smirnov test). (*E*) As in *D* but for FS units (L-D, $P = 0.99$; D-L, $P = 1.0$; 2-sample Kolmogorov–Smirnov test). (*F*) Mean firing rate for each RSU averaged across all transitions experienced by that neuron in L-D (*Left*) and D-L (*Right*) transitions. Paired data indicate that the average FR is for the same neuron. Distributions were not significantly different (L-D, $P = 0.677$; D-L, $P = 0.655$; Wilcoxon rank sum test), but individual neurons across the whole distribution showed consistent changes at the transitions. ***$P = 0.0002$; ****$P < 0.0001$ (Wilcoxon signed rank test). (*G*) As in *F* but for FS units. Distributions were not different (L-D, $P = 0.905$; D-L, $P = 0.827$; Wilcoxon rank sum test), but individual FS units changed their firing consistently at D-L but not L-D transitions (L-D, $P = 0.318$). *$P = 0.026$ (Wilcoxon signed rank test). (*H*) Percentage of change in firing rate across transition for RSUs (L-D, $-7.09 \pm 1.99\%$; D-L, $15.60 \pm 4.00\%$; Wilcoxon signed rank test). ****$P = 0.0001$. (*I*) As in *H* for FS units. Percentage of change in FR was different from 0 in the D-L but not the L-D condition (L-D, $-2.75 \pm 3.80\%$, $P = 0.410$; D-L: $9.73 \pm 5.12\%$; Wilcoxon signed rank test). *$P = 0.017$.

L-D and D-L transitions (Fig. 1*F*) (Wilcoxon signed rank test; L-D: $P = 0.0002$; D-L: $P = 0.0001$), while the activity of FS cells only changed significantly at D-L transitions (Fig. 1*G*) (Wilcoxon signed rank test; L-D: $P = 0.318$; D-L: $P = 0.026$). The magnitude of these effects was small: on the order of 7 to 15% for RSUs (Fig. 1 *H* and *I*) (RSU, L-D: $-7.09 \pm 1.99\%$, $P = 0.0001$; D-L: $15.60 \pm 4.00\%$, $P < 0.0001$; FS, L-D: $-2.75 \pm 3.80\%$, $P = 0.410$; D-L: $9.73 \pm 5.12\%$, $P = 0.017$; Wilcoxon signed rank test).

These data show that, surprisingly, dramatic changes in visual input cause very minor changes in V1 firing rates. The distributions of mean rates in the presence and absence of natural visual stimuli are identical in the proximity of transitions. Analysis of many transitions shows that RSU firing rates are consistently affected when the visual environment changes, but this modulation is decidedly modest.

**Behavioral State Affects Sensitivity of Firing Rates to Visual Stimuli.** As rats were freely behaving throughout, we considered whether their alertness state at the light transitions could affect the activity of V1 neurons. Local field potential (LFP), electromyography (EMG), and video data were collected and used to score animals' behavioral state into either asleep or awake (13). For each animal, 20-min periods centered on the L-D and D-L transitions were considered. Only periods during which the animal remained in the same behavioral state for the entire time were analyzed. For each neuron, we plotted the mean firing rate before the transition against the mean rate after the transition. The activities of neurons proved to be strikingly similar across all transitions regardless of whether the animals were awake or asleep (Fig. 2 *A* and *B*). In either behavioral state, firing rates in light and dark were very strongly correlated, and the slope of the regression line was close to 1 (RSU, wake, L-D: slope = 0.959, $r = 0.966$, $P < 10^{-43}$; D-L: slope = 0.960, $r = 0.991$, $P < 10^{-48}$; FS, wake, L-D: slope = 1.113, $r = 0.964$, $P < 10^{-13}$; D-L: slope = 0.976, $r = 0.990$, $P < 10^{-18}$; RSU, sleep, L-D: slope = 1.147, $r = 0.941$, $P < 10^{-12}$; D-L: slope = 0.990, $r = 0.996$, $P < 10^{-13}$; FS, sleep, L-D: slope = 1.093, $r = 0.987$, $P < 10^{-13}$; D-L: slope = 1.003, $r = 0.996$, $P < 10^{-11}$).

We again looked at the data in paired form by comparing a neuron's average firing rate on either side of an L-D transition. The mean activity of RSUs in V1 changed consistently across transitions when animals were awake (Fig. 2*C*) (L-D: $P = 0.0001$; D-L: $P = 0.0457$; Wilcoxon signed rank test) but not when they were asleep (Fig. 2*D*) (L-D: $P = 0.656$; D-L: $P = 0.925$; Wilcoxon signed rank test). We observed a similar pattern in FS cells, although the data in the wake condition were not significant for L-D transitions (Fig. 2 *E* and *F*) (wake, L-D: $P = 0.689$; D-L: $P = 0.039$; sleep, L-D: 0.557; D-L: $P = 0.638$; Wilcoxon signed rank test). Once again, these effects were of small magnitude (7 to 12%). Thus, V1 neurons do not respond to expected (circadian) changes in the visual environment when animals are asleep and respond only modestly when animals are awake.

**A Subpopulation of RSUs Is Consistently Responsive to D-L Transitions.** While we only detected small changes at the population level (and no change in the population distribution), we occasionally observed neurons with activity that appeared to be consistently modulated by visual stimuli. The majority of neurons showed no spiking modulation across multiple transitions (Fig. 3*A*), but a subset of neurons showed higher activity on 1 side of the transition (Fig. 3*B*). Occasionally, neurons responded to both L-D and D-L transitions (Fig. 3*B*), but more often, neurons were only responsive to one or the other. To quantify these observations, we treated each transition independently for each neuron, averaged firing rates for 10 min before and after lights on/off, and identified neurons that changed their firing rate consistently across transitions using a paired *t* test.

Because neuronal firing rates are variable, we presumed that some of these apparent responses were spurious. To estimate the false positive rate, we performed a bootstrap analysis using random time points as dummy "transitions." We chose 9 transition points 24 h apart from each other (to match circadian transitions) and analyzed mean firing rates for each neuron as above but using these dummy transition points. This process was repeated 100 times to arrive at an estimate of the mean and 95% confidence interval (95% CI) for the percentage of responsive

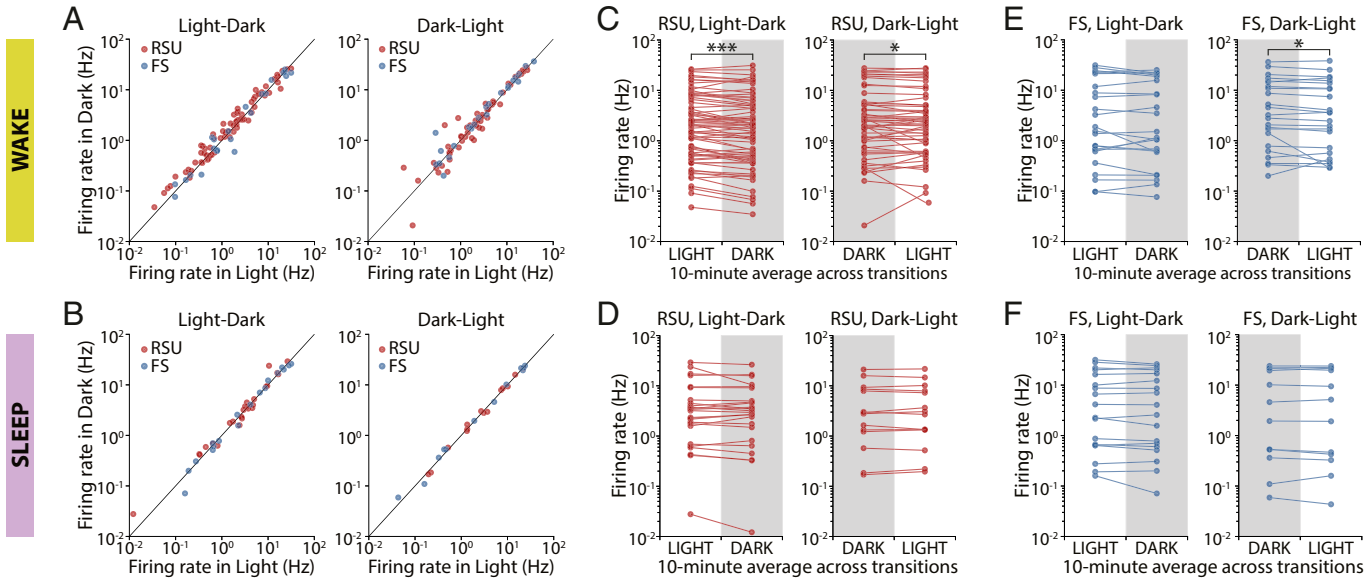


**Fig. 2.** L-D transitions modestly modulate the firing of V1 neurons during wake but not sleep. (*A*) Comparison of mean firing rates in 10-min averages around the transitions in L-D and D-L transitions when the animal was awake for the whole 20 min. Activity in light and dark was strongly correlated for both transition types. (*B*) As in *A* but for transitions during which animals were asleep for the 20-min period around the transition. Firing rates in light and dark during sleep were also strongly correlated. (*C*) Mean firing rate of individual RSUs calculated in 10-min averages around luminance transitions and averaged across all transitions during which animals were awake. Neuronal activity changed consistently at the transitions (***$P = 0.0001$; *$P = 0.0457$; Wilcoxon signed rank test). (*D*) As in *C* but for transitions during which animals were asleep. No significant change was observed (L-D, $P = 0.656$; D-L, $P = 0.925$; Wilcoxon signed rank test). (*E*) As in *C* for FS units. Cells' activity only changed significantly at D-L transitions (L-D, $P = 0.689$; D-L, *$P = 0.039$; Wilcoxon signed rank test). (*F*) As in *D* for FS cells. No significant change was observed (L-D, $P = 0.557$; D-L, $P = 0.638$; Wilcoxon signed rank test).

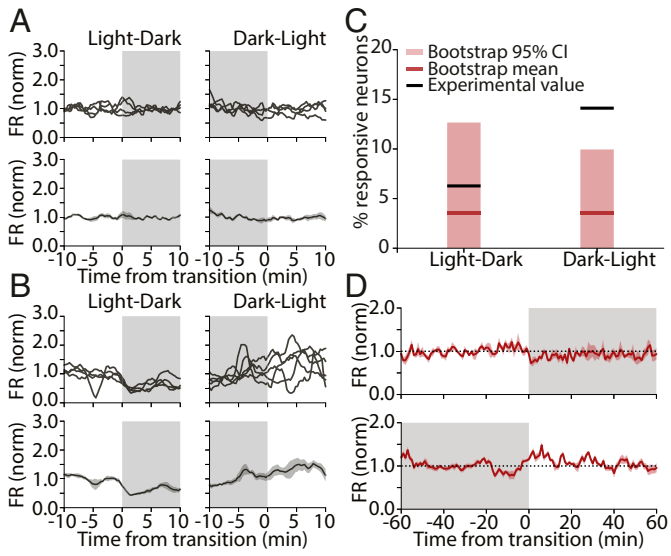


**Fig. 3.** A subset of RSUs consistently increases their firing rate in response to expected D-L transitions. (*A*) Example of an RSU unresponsive to light transitions (*Left*, L-D; *Right*, D-L). (*Upper*) Binned firing rate for each transition; (*Lower*) average across transitions. (*B*) Example RSU that responds consistently to light transitions. (*C*) Percentage of RSUs found to be responsive to L-D (*Left*) and D-L (*Right*) transitions and bootstrap control. Black lines show experimental values (actual percentages of responsive neurons); red lines show bootstrap means. Light red bars show the extent of the bootstrap 95% CI (L-D, actual value: 6.25%, bootstrap mean: 3.55%; 95% CI: 0 to 12.62%; D-L, actual value: 14.06%; bootstrap mean: 3.09%; 95% CI: 0 to 9.91%; $n = 64$). (*D*) Mean firing rate averaged across transitions for all D-L–responsive RSUs calculated for 2 h around each transition for L-D (*Upper*) and D-L (*Lower*) transitions. The transient nature of the firing-rate response is visible in *Lower*.

cells (mean [95% CI], RSU, L-D: 3.55% [0 to 12.62%]; D-L: 3.09% [0 to 9.91%]; $n = 64$).

The proportion of cells that we found to be transition responsive was within the range expected by chance for all conditions except for RSUs in D-L transitions (Fig. 3*C*). We found that 14% of RSUs in our experimental condition had significantly changing firing rates from dark to light, well outside the range expected by chance (95% CI for this group: 0 to 9.91%). In addition, most of these neurons (88.9%) showed an increase in firing rate at the onset of light, while in the bootstrap control, neurons were found to have an equal probability of increasing or decreasing their activity at a given transition point (51.8% of neurons increasing).

Finally, we examined the temporal dynamics of firing-rate changes for the subset of RSUs that were consistently responsive to D-L transitions. We plotted the mean activity within 1 h of the transition, across all transitions, and across neurons for this subpopulation (Fig. 3*D*). On average, the change in firing rate was short lived (on the order of ∼10 min) and of moderate size (∼25% increase). This analysis shows that a small subset of excitatory pyramidal neurons in V1 consistently modulates their activity in response to the expected appearance of visual input after a circadian 12-h period of darkness. This change is transient, with firing rates returning to pretransition levels within minutes.

**Light Transitions Have No Effect on Average Interspike Intervals over Short Timescales.** Our analysis so far shows that, on a timescale of tens of minutes, few V1 neurons show significant firing-rate modulation to the appearance or disappearance of natural visual stimuli. One possible explanation for this apparent lack of responsiveness is that these dramatic sensory changes trigger a rapid adaptation mechanism that quickly restores average V1 activity back to baseline. Such adaptation mechanisms within V1 have been well described and can operate on a timescale of hundreds of milliseconds to many minutes (22–24). To address this possibility, we examined neuronal firing in 1-, 10-, and 30-s intervals around L-D and D-L transitions for RSUs in our dataset. Spiking in these short time windows was sparse and variable across days (Fig. 4 *A* and *B*). We averaged the mean interspike interval (ISI) across days for each cell and compared averages in the 10 s before and after transitions. To ensure that we were not missing effects on even shorter timeframes, we also computed the mean ISI in 1-s windows around the transitions. For both the 1- and 10-s cases, we found no statistically significant effect (Fig. 4 *C*, 1 s, L-D: $P = 0.27$; D-L: $P = 0.36$; and *D*, 10 s, L-D: $P = 0.97$; D-L: $P = 0.31$; Wilcoxon signed rank test). Similar results were obtained when this analysis was carried out with 5- and 30-s intervals. This indicates that the stability of firing across transitions is not due to a short-term adaptation process that rapidly restores firing to baseline.

**Pairwise Correlations in V1 Are Significantly Higher in Light than in the Dark.** To investigate whether higher-order network properties are modified by the presence or absence of natural visual stimuli, we examined the structure of pairwise correlations in light and in dark (Fig. 5) ($n = 5$ animals). Plotting the correlation matrices of 1 animal at postnatal day 27 (P27) revealed that these correlations were higher in the light (calculated over the 12-h period at P27) than in the dark (calculated over the 12-h period at P27.5) (Fig. 5*A*). We then plotted the average correlation computed continuously over 4 d (normalized to the average correlation of each animal at P26 in light) (Fig. 5*B*). The normalized pairwise correlation showed a pronounced oscillation across light and dark periods and was consistently higher in the light. To assess the degree to which correlation of individual pairs changed, we compared the correlation of 922 pairs in light vs. dark computed for spike counts with bin sizes of 5 or 100 ms, respectively. We found that correlations in light were higher than in the dark for both bin sizes (Fig. 5 *C*, *Left*, 5 ms: $P < 10^{-70}$ and *Right*, 100 ms: $P < 10^{-125}$; Wilcoxon signed rank test). To ensure that the observed difference of correlations between light and dark was not caused by disproportionate time spent in wake or sleep, we

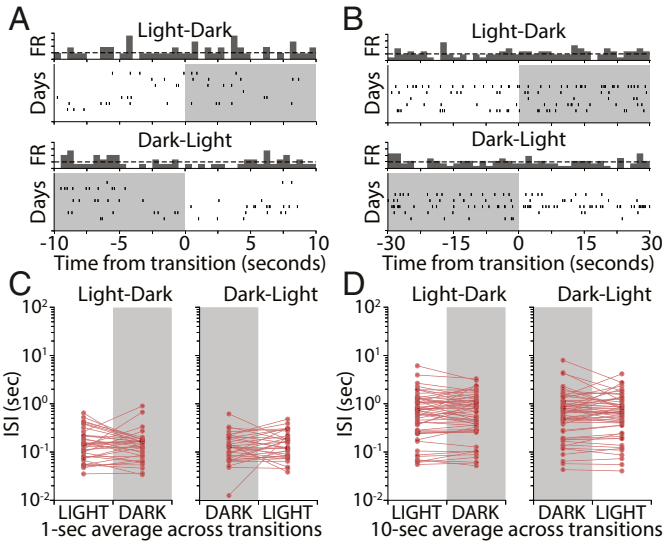


**Fig. 4.** L-D and D-L transitions have no effect on V1 firing rates over short timescales. (*A*) Raster plot showing activity of an example RSU in a 10-s interval around L-D and D-L transitions. Vertical ticks represent spikes, and rows represent transitions happening on different recording days. (*B*) A second example RSU showing a 30-s interval around transitions. (*C*) Mean ISIs for all recorded cells obtained by averaging ISIs in 1-s bins around L-D and D-L transitions for different days. Each dot represents the mean for 1 cell obtained by averaging across days. No significant change was observed (L-D, $P = 0.27$; D-L, $P = 0.36$; Wilcoxon signed rank test). (*D*) As in *C* but for 10-s averages. No significant change was observed (L-D, $P = 0.97$; D-L, $P = 0.31$; Wilcoxon signed rank test).

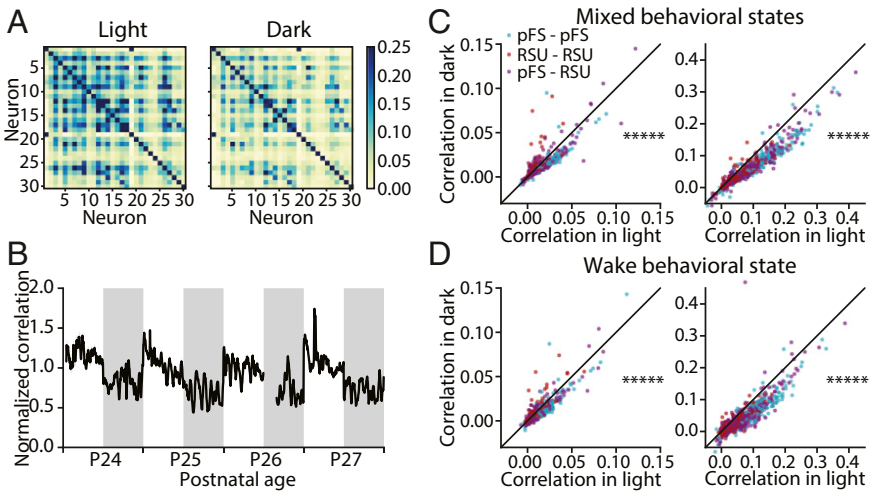


**Fig. 5.** Pairwise correlations in V1 are higher during light than during dark. (*A*) Example pairwise correlation structure of 30 neurons from a single animal during light (*Left*; calculated over the 12-h period at P27) and during dark (*Right*; calculated over the 12-h period at P27.5). (*B*) The average correlation of 922 pairs from 5 animals over 4 d normalized to the average correlation of each animal relative to P26 in light. The gap in the data at P26 corresponds to the time that animals were anesthetized for monocular deprivation, which was excluded from analysis. (*C*) Comparison of the average correlation of 922 pairs during light and during dark with bin size 5 ms (*Left*) and bin size 100 ms (*Right*). (*Left*) *****$P < 10^{-70}$ (Wilcoxon signed rank test). (*Right*) *****$P < 10^{-125}$ (Wilcoxon signed rank test). (*D*) Comparison of average correlation of 922 pairs in wake during light and during dark with bin size 5 ms (*Left*) and bin size 100 ms (*Right*). (*Left*) *****$P < 10^{-55}$ (Wilcoxon signed rank test). (*Right*) *****$P < 10^{-110}$ (Wilcoxon signed rank test). pFS, putative fast-spiking neurons.

restricted the analysis to periods of wake and again computed the average correlation. Consistent with our previous analysis, correlations in wake during light were significantly greater than in wake during dark (Fig. 5 *D*, *Left*, 5 ms: $P < 10^{-55}$ and *Right*, 100 ms: $P < 10^{-110}$; Wilcoxon signed rank test). These results indicate that the presence of natural visual stimuli increases pairwise correlations in V1.

**Noncircadian, Unexpected L-D Transitions Are More Likely to Perturb V1 Firing Rates.** All of our data so far suggest that dramatic changes in visual input at circadian L-D transitions have very subtle effects on V1 firing. We wondered if this might be due to circadian entrainment (i.e., that when L-D and D-L transitions happen at regular times, they are expected, and the response of neurons to otherwise salient stimuli is attenuated). To test this, we examined neuronal responses to stimulus transitions occurring at random points in the circadian cycle.

We recorded single-unit activity in V1 in a different subset of animals while turning the lights off (or on) for 10 min during the light (or dark) cycle (Fig. 6*A*) ($n = 6$ animals). We then calculated the number of neurons that consistently and significantly changed their firing at these unexpected transitions and again, used a bootstrap analysis to calculate the false positive rate. In marked contrast to expected transitions (Fig. 3*D*), we found that both L-D and D-L unexpected transitions caused a subset of RSUs to consistently modulate their spiking (Fig. 6 *B* and *C*). This effect was seen regardless of behavioral state (significantly changing RSUs, sleep, L-D: 21.9%, $n = 64$; D-L: 13.4%, $n = 67$; wake, L-D: 17.6%, $n = 91$; D-L: 12.7%, $n = 55$), and the proportion of significantly changing neurons was higher than expected by chance in most conditions (bootstrap mean [95% CI]; RSU, sleep, L-D: 4.42% [0 to 8.95%]; D-L: 4.31% [0 to 9.38%]; RSU, wake, L-D: 4.22% [0 to 8.79%]; D-L: 4.33% [0 to 9.09%]). These results show that more neurons respond consistently to L-D and D-L transitions when these do not line up with the circadian cycle that the animals are entrained on. However, even during these unexpected transitions, only a minority (12 to 20%) of neurons consistently changed their firing rate in response to the appearance or disappearance of natural visual stimuli.

To further examine the nature of these responses, we averaged and plotted the activity of the subpopulation of responsive units for each combination of behavioral state and transition type (Fig. 6 *D* and *E*). In the wake state, there was a net decrease in firing during L-D transitions and vice versa for D-L (Fig. 6*E*), whereas there was little net change during sleep (Fig. 6*D*). This is likely due to averaging over changes in opposite directions during sleep, as 4 of 14 cells (29%) increased their FR in L-D transitions and 5 of 9 cells (55%) increased their FR in D-L transitions. However, even in wake, these firing-rate changes were on the order of 10 to 25%, indicating that, even under the most permissive conditions, firing is only subtly modulated by brief L-D transitions.

**Prolonged Dark Exposure Enhances the Responsiveness of V1 Neurons to Natural Visual Input.** Our data show that reexposure to light after 12 h of darkness has only modest effects on V1 firing; in contrast, reexposing animals to light after a period of prolonged darkness is a standard paradigm for increasing activity-dependent gene expression in V1 (refs. 25–27; reviewed in ref. 28). We, therefore, wondered whether prolonged dark exposure might unmask robust responses to the sudden onset of visual stimuli within V1.

We began by using expression of the immediate early gene *c-fos*, which is driven by enhanced calcium influx during elevated activity (refs. 29 and 30; reviewed in ref. 31). After prolonged darkness, brief light exposure induces widespread *c-fos* expression in V1 of cats and rodents (25, 32–34). To replicate this, we placed P26 rats in the dark for 60 h (12 h + 2 d) and then, exposed them to light for 1 h before immunostaining for the c-fos protein (light exposed, $n = 28$ slices, 5 animals). We used age-matched animals either exposed to 1 h of light after a regular 12-/12-h cycle (regular control, $n = 22$ slices, 4 animals) or kept in the dark for 60 h but killed before lights on (dark control, $n = 23$ slices, 4 animals) as controls (Fig. 7 *A* and *B*). Animals in the light exposure condition showed an elevated percentage of c-fos–positive cells (Fig. 7 *C*, *Upper*) (regular control: 11.4 ± 1.6%; dark control: 6.1 ± 0.8%; light exposed: 16.8 ± 1.7%; light exposed vs. regular control $P = 0.032$; light exposed vs. dark control $P = 0.001$; 1-way ANOVA with Tukey post hoc test) as well as increased total staining intensity (Fig. 7 *C*, *Lower*) (normalized to regular control; regular control: 1.00 ± 0.06; dark control: 0.79 ± 0.05; light exposed: 1.31 ± 0.09; light exposed vs. regular control $P = 0.011$; light exposed vs. dark control $P = 0.001$; 1-way ANOVA with Tukey post hoc test). These data confirm that a 60-h period of prolonged darkness is sufficient to up-regulate c-fos in rodent V1 on light reexposure.

Next, we asked whether elevated c-fos expression was correlated with increased firing. We used the same paradigm as above but recorded continuously from V1 during the baseline, dark exposure, and light reexposure periods ($n = 4$ animals). On light reexposure, both RSUs and FS cells showed a substantial transient increase in firing rate at the time of lights on (Fig. 7*D*) (RSU: $n = 32$; FS: $n = 12$). We compared average firing rates 10 min before and after the transition for each cell. Both FS and RSU populations showed a significant increase in firing rate after light reexposure (Fig. 7 *F* and *G*) (RSU: $P < 10^{-5}$; FS: $P = 0.034$; Wilcoxon signed rank test). The percentage change in firing rate across the transition was also significantly different from 0 (Fig. 7*E*) (all cells: 87.1 ± 13.5%, $P < 10^{-7}$; RSU: 80.7 ± 14.9%, $P < 10^{-5}$; FS: 104.3 ± 29.8%, $P = 0.005$; 1-sample *t* test),

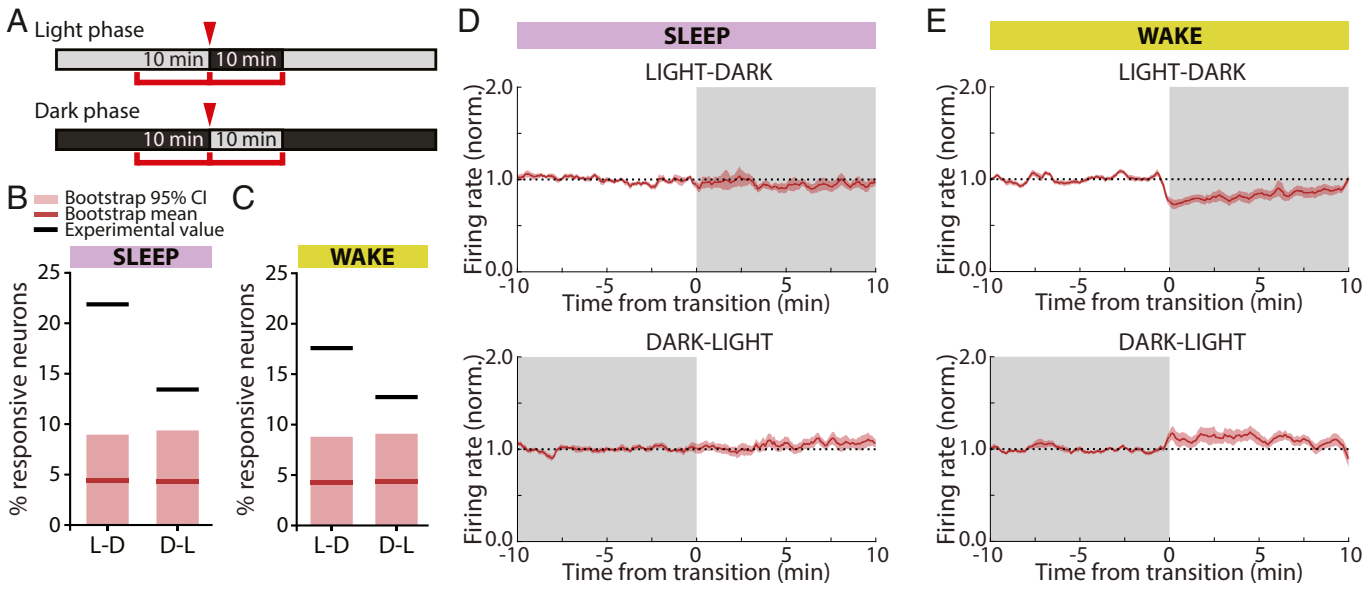


**Fig. 6.** Randomly timed L-D and D-L transitions induce consistent firing-rate changes in RSUs. (*A*) Experimental design. Animals were exposed to 10-min periods of darkness during the light phase and 10-min periods of light during the dark phase at random points throughout the L-D cycle. Mean firing rates were calculated in 10-min intervals around the transition. (*B*) Percentage of RSUs that were responsive to L-D and D-L transitions when transitions happened in epochs of sleep. Black lines shows actual experimental values; red lines show bootstrap means. Light red bars cover the bootstrap 95% CIs (sleep, L-D: 21.9%, bootstrap mean [95% CI]: 4.42% [0 to 8.95%], $n = 64$; D-L: 13.4%, bootstrap mean [95% CI]: 4.31% [0 to 9.38%], $n = 67$). (*C*) As in *B* but for transitions happening while animals were awake (wake, L-D: 17.6%, bootstrap mean [95% CI]: 4.22% [0 to 8.79%], $n = 91$; D-L: 12.7%, bootstrap mean [95% CI]: 4.33% [0 to 9.09%], $n = 55$). (*D*) Mean firing rates of units that consistently responded to visual input changes in sleep averaged across cells and transitions for L-D (*Upper*) and D-L (*Lower*) transitions. Ten minutes on either side of the transition are shown. (*E*) As in *D* for units responsive during wake transitions.

and the majority of neurons increased their activity at lights on (RSU: 31 of 32 neurons; FS: 10 of 12 neurons).

One possible cause of enhanced responsiveness at salient unexpected transitions (i.e., those happening after prolonged darkness or at noncircadian times) (Fig. 6) is that responses are normally suppressed when transitions are anticipated. If so, we would expect to see an opposite effect on firing when an expected transition does not occur. To look for such an effect, we examined firing rates during prolonged dark exposure at the times when expected (circadian) L-D transitions did not occur (Fig. 8 *A* and *B*, "missing" transitions). We found a nonsignificant trend toward reduced firing across the population at missing D-L transitions (Fig. 8 *A*, *Lower Right*) and a small but significant increase in firing at missing L-D transitions (Fig. 8 *B*, *Lower Right*). Because we had only 2 repetitions of each transition, we could not quantify the fraction of neurons that responded consistently to expected transitions that did not occur.

Finally, it has been reported that prolonged dark exposure increases firing rates in rodent V1 (35), suggesting that enhanced responsiveness to light reexposure might arise from increased excitability of V1 circuitry. To examine this, we asked how prolonged dark exposure affected RSU firing rates in freely behaving animals before light reexposure. When we compared the distribution of RSU firing rates during the first and last 12 h of the 60-h-long period of prolonged darkness, rather than an increase, we found a small but significant decrease in firing rates (Fig. 8*C*) (mean ± SEM; first 12 h: 4.00 ± 0.97 Hz, last 12 h: 2.27 ± 0.57 Hz; median; first 12 h: 1.18 Hz; last 12 h: 0.85 Hz; $P = 0.044$; Wilcoxon rank sum test). Thus, the enhanced responsiveness to restoration of natural visual stimuli is unlikely to be due to a simple global increase in circuit excitability.

In summary, these data indicate that prolonged dark exposure disrupts the normal stability of V1 firing across D-L transitions and suggest that the maintenance of this stability is dependent on visual experience.

## Discussion

How internal and external factors influence the long-term dynamics of neuronal firing in V1 is poorly understood. Here, we recorded from ensembles of single units over a period of several days in freely viewing and behaving animals and found that firing rates of both excitatory and inhibitory V1 neurons were remarkably stable even when sensory input changed abruptly and dramatically. During expected circadian L-D transitions, very few V1 neurons significantly changed their firing. A larger subset of V1 neurons was consistently responsive to unexpected L-D transitions, and disruption of the regular L-D cycle with 2 d of complete darkness induced a widespread and robust increase in V1 firing on subsequent reintroduction of visual input. These data show that most V1 neurons fire at similar rates in the presence or absence of natural visual stimuli and that significant changes in mean activity arise only in response to unexpected changes in the visual environment. While mean firing rates were not different in light and dark, pairwise correlations were significantly stronger in the light in the presence of natural visual stimuli, even when controlling for behavioral state. Taken together, our findings are consistent with a process of rapid and active stabilization of firing rates during expected changes in visual input and demonstrate that firing rates in V1 are remarkably stable at both short and long timescales.

The near absence of firing-rate modulation in response to the appearance (or disappearance) of natural visual stimuli may seem surprising, as there is a rich literature supporting the idea that V1 neurons respond to optimal stimuli by increasing their spiking (36–44). Many of these studies used anesthetized preparations, making comparisons with our results difficult, but our data are consistent with previous reports of small differences in overall activity between natural vision and complete darkness in awake animals (8) and sparse modulation of spiking in response to natural scene viewing (12, 45–47). In general, our data support the view that mean firing rates in V1 can be stabilized over both long (13) and short timescales without interfering with visual coding, which may arise through very sparse modulation of spiking and/or higher-order population dynamics (46, 48, 49).

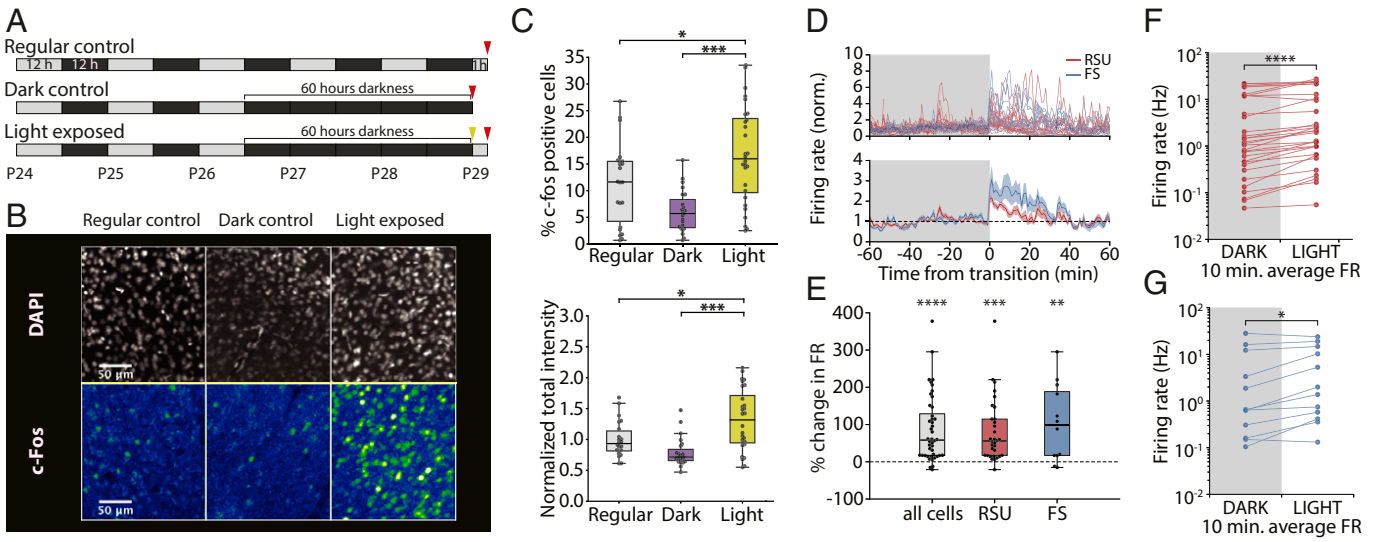


**Fig. 7.** Light reexposure after prolonged darkness results in increased c-fos expression and increased neuronal firing in V1. (*A*) Experimental protocol. Animals were exposed to a regular 12-/12-h L-D cycle and killed 1 h after lights on at P29 (regular control, $n = 22$ slices, 4 animals), exposed to darkness for 60 h starting at P26 and killed before lights on (dark control, $n = 23$ slices, 4 animals), or exposed to 60 h of darkness starting at P26 and killed 1 h after light reexposure (light exposed, $n = 28$ slices, 5 animals). (*B*) Representative images showing DAPI (*Upper*) and c-fos (*Lower*) immunostaining for regular control (*Left*), dark control (*Center*), and light-exposed (*Right*) animals. (*C*, *Upper*) Percentage of c-fos–positive cells in all 3 groups (regular control: 11.4 ± 1.6%; dark control: 6.1 ± 0.8%; light exposed: 16.8 ± 1.7%). $^{*}P = 0.032$ (1-way ANOVA with Tukey post hoc test); $^{***}P = 0.001$ (1-way ANOVA with Tukey post hoc test). (*C*, *Lower*) Total colocalized DAPI and c-fos staining intensity normalized to average of regular control group (regular control: 1.00 ± 0.06; dark control: 0.79 ± 0.05; light exposed: 1.31 ± 0.09). $^{*}P = 0.011$ (1-way ANOVA with Tukey post hoc test); $^{***}P = 0.001$ (1-way ANOVA with Tukey post hoc test). (*D*) Time course of RSU and FS spiking in a 2-h period around the time of light reexposure. Individual unit traces (*Upper*) and average across cells (*Lower*) show marked increase in firing at the time of lights on. (*E*) Percentage of change in firing rate between the 10 min before and the 10 min immediately after light reexposure (all cells, 87.1 ± 13.5%, $n = 44$; RSU, 80.7 ± 14.9%, $n = 32$; FS, 104.3 ± 29.8%, $n = 12$, 1-sample $t$ test). $^{**}P = 0.005$; $^{***}P < 10^{-5}$; $^{****}P < 10^{-7}$. (*F*) Mean firing rate in the 10 min before and after the transition for each recorded RSU. $^{****}P < 10^{-5}$ (Wilcoxon signed rank test). (*G*) As in *F* for recorded FS cells. $^{*}P = 0.034$ (Wilcoxon signed rank test).

Despite the lack of robust changes in firing rates across the population at D-L transitions, we did observe a small subset of neurons that transiently increased their firing specifically at the appearance of visual input (Fig. 3). Interestingly, many layer 4 (L4) neurons (which account for 34% of our dataset) did not show this kind of transient response at D-L transitions. The subset of responsive neurons included cells from L4 but also, all other layers of V1. Since this occurred in freely viewing animals, it is unlikely that these neurons were responding to the same specific visual stimuli on successive days. A more parsimonious explanation is that these neurons are activated by luminance changes at most D-L transitions. One possible source of drive to these neurons is from intrinsically photosensitive retinal ganglion cells (ipRGCs), which are known to exhibit prolonged changes in firing on changes in luminance (50–52). Some classes of ipRGCs have been shown to project to the dorsolateral geniculate nucleus (dLGN) of mice, where they modulate the firing of ∼20 to 30% of dLGN neurons (53–55) and thus, can influence activity in V1 (53). The firing of V1 neurons activated by L-D transitions adapted over the first several minutes, but whether all ipRGC firing adapts over a similar timescale is not known (53–55). There is evidence that the ipRGCs of the M1 class can produce persistent responses, resulting in temporal integration over several minutes (56). It is at least plausible that the activity of these cells is contributing to the coding of light levels in V1. Unfortunately, it is difficult to directly test the role of ipRGCs in V1 responses in rats, as transgenic animals that would allow the selective activation of ipRGCs without activating rods or cones (as for mouse) (54, 55, 57) are not currently available. While it is possible that this small subset of responsive neurons represents sparse coding for the transition event, it is equally possible that this is a simple reflection of upstream changes in activity and that V1 does not explicitly code for sharp light transitions.

Interestingly, we detected a greater proportion of transition-responsive cells when light transitions happened randomly throughout the L-D cycle, including a population of neurons that transiently responded to noncircadian L-D transitions by decreasing their firing rate (Fig. 6). Thus, unexpected changes in visual drive unmask robust and bidirectional changes in firing in a small subset (15 to 20%) of V1 neurons. We observed a similar effect when expected L-D transitions did not occur, which unmasked an increase in firing. There are several potential explanations for this effect. It is possible that the responsive neurons are specialized to represent this "unexpectedness" as an error signal, such as has been proposed in some models of predictive coding (58, 59). Alternatively, it could be the result of modulation by other brain areas that encode the surprise signal, akin to that seen in response to attention or reward cues (60, 61) or during modulation of V1 by locomotion (62).

We were able to disrupt the normal conservation of firing rates across D-L transitions even more dramatically by using a prolonged dark-exposure paradigm, which induced a network-wide enhancement of firing on light reexposure. This paradigm is thought to induce metaplastic changes within V1 that increase α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) quantal amplitude onto L2/3 pyramidal neurons (35, 63, 64), but the impact of these changes on overall V1 function and excitation/inhibition balance is unclear. A previous study in anesthetized animals found that several days of dark exposure increased firing rates in V1, raising the possibility that prolonged dark exposure increases overall V1 excitability (35); however, here we found a small but significant reduction in mean firing rate across the population in freely behaving animals, suggesting that circuit excitability is, if anything, reduced by 60 h of dark exposure. Interestingly, this change in firing rate (FR) does not seem to trigger a homeostatic compensation in the opposite direction. This could be because a slow and gradual shift in neuronal FR set points occurs during darkness, or perhaps, we are simply reintroducing light before homeostatic changes have a chance to influence FRs. Although the circuit mechanism by which dark exposure unmasks

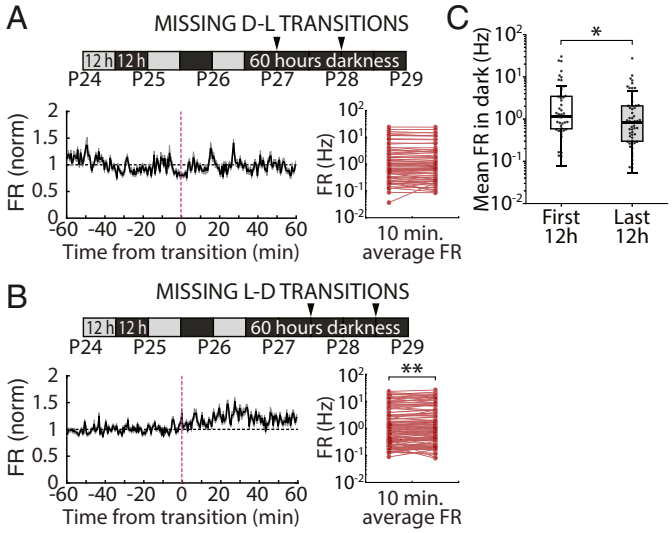


**Fig. 8.** Firing-rate dynamics in V1 during prolonged darkness. (*A*) Mean firingrate dynamics of RSUs recorded during 60 h of continuous dark exposure at expected D-L transitions that did not occur (missing transitions). (*Lower Left*) Firing rate in a 2-h period around the missing transition averaged across all neurons and both transitions. (*Lower Right*) Average FR for all neurons calculated in 10-min bins before and after the transition, showing no change in activity when lights did not come on at the expected D-L transitions (4.56 ± 5.94%, $P = 0.162$; Wilcoxon signed rank test). (*B*) As in *A* but for L-D transitions. We found a significant increase in FR when L-D transitions were expected but did not occur (10.70 ± 2.75%, $^{**}P = 0.0045$; Wilcoxon signed rank test). (*C*) Mean firing rate for all recorded RSUs averaged across the first 12-h period within the 60 h of darkness (first 12 h, mean ± SEM: 4.00 ± 0.97 Hz, median: 1.18 Hz, $n = 47$) and the last 12-h period of darkness before light reexposure (last 12 h, mean ± SEM: 2.27 ± 0.57 Hz, median: 0.85 Hz, $n = 55$). $^{*}P = 0.044$ (Wilcoxon rank sum test).

robust responses to D-L transitions is unclear, these experiments suggest that normal visual experience is necessary to maintain the ability of V1 circuits to stabilize their firing across these transitions.

In contrast to our observations on the stability of firing rates, we found that pairwise correlations in visual cortex were markedly higher in the light phase than in the dark phase (Fig. 5). This is consistent with previous reports that ongoing spontaneous activity in the dark is less correlated than activity elicited by natural scene stimuli (8, 65). Correlations are dependent on the degree of synchrony within neuronal circuits (66, 67) and are higher during anesthesia (68), raising the possibility that this is a simple reflection of time spent in different behavioral states during the light and dark phase. However, we observed the same increased correlation in light when only analyzing periods when animals were awake, ruling out this possibility. Thus, we conclude that, in freely behaving and viewing animals, sensory input can shift visual cortical circuits to more correlated dynamical states, even in a condition of low synchrony when animals are awake. It should be noted that these correlations may simply reflect the timing of shared visual input to neurons with similar tuning and thus, might not provide additional information above that carried by the timing of individual neuronal responses.

Our results add to a growing body of work suggesting that ongoing activity in mammalian V1 plays an important role in modulating sensory responses as well as in integrating other sensory, motor, and motivational signals (5, 8, 10, 11, 48, 58, 59, 62, 69–75). Our results also show that firing rates of most V1 neurons are remarkably stable over both long and short timescales and in the presence and absence of visual information, suggesting that most visual information during natural viewing is not encoded by changes in firing rates. Instead, our data suggest that perturbations in firing primarily occur during unexpected changes in visual input, indicating an effect of entrainment/expectation and the existence of an active mechanism for stabilization of activity. This may be of particular importance given the observation that pairwise correlations are increased when animals are exposed to visual input, as global fluctuations in firing rate can strongly affect the strength of correlations between pairs of neurons (66). Thus, it is possible that stable firing rates enable changes in correlations to reflect differences in sensory input and hence, to promote effective sensory processing.

## Methods

All surgical and experimental procedures were approved by the Animal Care and Use Committee of Brandeis University and complied with the guidelines of the NIH.

**Surgery and In Vivo Electrophysiology Experiments.** The data analyzed in this study were collected in previous electrophysiological experiments (13; $n = 7$ rats) as well as from a new set of animals ($n = 14$ rats; $n = 21$ rats total). All surgical procedures were as described previously (16). Briefly, Long–Evans rats of either sex were bilaterally implanted with custom 16-channel, 33-μm tungsten microelectrode arrays (Tucker–Davis Technologies) into monocular primary visual cortex (V1m) on P21. Location was confirmed post hoc via histological reconstruction. Two EMG wires were implanted deep in the nuchal muscle. Animals were allowed to recover for 2 full days postsurgery in transparent plastic cages with ad libitum access to food and water. Recording began on the third day after surgery. The recording chamber (a 12- × 12-inch Plexiglas cage with walls lined with high contrast square wave gratings with spatial frequency of 0.3 to 1 cycles/cm) was lined with 1.5 inches of bedding and housed 2 rats. Animals had ad libitum food and water and were separated by a clear plastic divider with 1-inch holes to allow for tactile and olfactory interaction while preventing jostling of headcaps and arrays. Electrodes were connected to commutators (TDT) to allow animals to freely behave throughout the recordings. Novel toys were introduced every 24 h to promote activity and exploration and provide additional visual stimulation. Lighting and temperature were kept constant (L-D 12:12, lights on at 7:30 AM, 21 °C, humidity 25 to 55%). Light levels during light and dark were obtained by measuring irradiance at the cage floor using an optical energy meter (ThorLabs PM100D). Irradiance values were 48.0 μW/m$^2$ (light) and less than 0.0001 μW/m$^2$ (dark) for incident light with a wavelength of 510 nm. Data were collected continuously for 9 to 11 d (200 to 240 h). Some animals ($n = 11$ rats) underwent a lid suture and/or eye reopening procedure on the third day of recording; in this study, we only analyzed data collected from the control hemisphere ipsilateral to the manipulated eye. For dark exposure experiments, animals were kept in the dark starting on days 4 and 5 of the recording (i.e., starting at the time of lights off on day 3 from P26 to P28). Lights came on at the regular time (7:30 AM) on day 6 (P29).

**Electrophysiological Recordings.** In vivo electrophysiological recordings were performed as previously described (13). Briefly, data were acquired at 25 kHz, digitized, and streamed to disk for offline processing using a Tucker–Davis Technologies Neurophysiology Workstation and Data Streamer. Spike extraction and sorting were performed using custom MATLAB code. Spikes were detected as threshold crossings (−4 SD from mean signal) and resampled at 3 times the original rate. Each wire's waveforms were then subjected to principal component analysis, and the first 4 principal components were used for clustering using KlustaKwik (76). Clusters were merged or trimmed as described previously (13). Spike sorting was done using custom MATLAB code relying on a random forest classifier trained on a manually scored dataset of 1,200 clusters. For each cluster identified from the output of KlustaKwik, we extracted a set of 19 features, including ISI contamination (percentage of ISIs < 3 ms), similarity to RSU and FS waveform templates, 60-Hz noise contamination, rise and decay times and slope of the mean waveform, waveform amplitude, and width. Cluster quality was also ensured by thresholding of L-ratio and isolation distance (77). The random forest algorithm classified clusters as noise, multiunit, or single unit. Only single units with a clear refractory period were used for additional analysis. Units were classified as RSU or FS based on the time between the negative peak and the first subsequent positive peak of the mean waveform (Fig. 1*B*). Clusters were classified as RSUs if this value was >0.39 ms and as FS otherwise (19), with a lower threshold of 0.19 ms to eliminate noise. We used previously established criteria and methods to select neurons that we could reliably follow over time (13). Briefly, we considered neurons to be continuously recorded if they satisfied the following criteria: (*i*) waveforms constituting an isolatable cluster, (*ii*) presence of absolute refractory period, (*iii*) minimal change in spike shape across recording days assessed by computing the sum of squared errors between daily average waveforms, (*iv*) high signal-to-noise ratio, and (*v*) stability of firing rate (no continuous increase or

decrease). Only neurons that could be recorded for at least 48 consecutive hours were used for analysis of L-D transitions. For extended dark experiments, we analyzed all neurons that were online for at least 1 h preceding and 1 h following the time of light reexposure. For estimates of mean firing during the extended dark phase, we analyzed the activity of all cells that could be recorded in the first and last 12-h periods during the 60 h of darkness.

**Behavioral-State Classification.** The behavioral state of animals was classified using a combination of LFP, EMG, and estimate of locomotion based on video analysis (13). LFPs were extracted from 3 separate recorded channels, resampled at 200 Hz, and averaged. The power spectral density was computed in 10-s bins using a fast Fourier transform method (MATLAB "spectrogram" function) using frequency steps of 0.1 Hz from 0.3 to 15 Hz. Power in the delta (0.3 to 4 Hz) and theta (6 to 9 Hz) bands was computed as a fraction of total power in each time bin. A custom algorithm was used to score each 10-s bin and assign 1 of 4 behavioral codes based on the power in each frequency band as well as EMG and movement activity: active wake (high EMG and movement, low delta and theta), quiet wake (low EMG and movement, low delta and theta), rapid eye movement (REM) sleep (very low EMG, no movement, low delta, high theta), and non-rapid eye movement (NREM) sleep (low EMG and movement, high delta, low theta). For each animal, each hour of data was scored separately. The first 10 h were scored manually and used as an initial training set for a random forest classifier (implemented in Python). The classifier was then used to score each successive hour, with manual corrections performed as needed. The classifier was retrained after every hour scored, with a maximum number of 10,000 bins used for training (using only the most recent 10,000 bins).

**Extended Darkness, Immunostaining, and Image Analysis.** For analysis of c-fos protein after extended darkness, we transferred animals ($n$ = 13 rats) to a custom dark box on P21. A light timer was set up to allow for complete control of the L-D cycle inside the box. When animals were P26, lights were allowed to turn off at the regular time (7:30 PM) and set up to not turn back on. Animals were in complete darkness for 60 h from the night of P26 until the night of P28 (ages matched with electrophysiological recordings). On the morning of P29, lights were allowed to turn back on at 7:30 AM. Animals were allowed to experience 1 h of light before being deeply anesthetized and transcardially perfused. Control animals were either not exposed to darkness but kept on a regular 12-h cycle or anesthetized at 7:30 AM on P29 (before lights on) in the dark using infrared night vision goggles and then immediately perfused. Brains were fixed in 3.7% formaldehyde, and 60-μm coronal slices of V1m were taken on a vibratome (Leica VT1000S). Slices were immersed in a solution of phosphate-buffered saline (PBS) and $NaN_3$ and stored for immunostaining. To ensure consistent results between groups, all conditions were run in parallel. Slices were incubated in a primary antibody solution (1:1,000; rabbit anti–c-fos; Cell Signaling Technologies) at room temperature for 24 h. They were then rinsed and incubated for 2 h with a secondary antibody (anti-rabbit Alexa Fluor 568; 1:400; Thermofisher). Sections were mounted on microscope slides with a DAPI-containing medium (DAPI Fluoromount-G; Southern Biotech), coverslipped, and allowed to dry for 24 h before imaging. Imaging was performed on a confocal microscope (Zeiss Laser Scanning Microscope 880). A 10× objective was used to take $z$ stacks of V1m in the DAPI and c-fos channels. Imaging settings were optimized for each staining/imaging session and kept constant across conditions; all conditions were imaged on a given session. Images were imported into Metamorph software for analysis. A granularity analysis was used to determine locations of cell bodies, and colocalized DAPI- and c-fos–positive granules were counted as c-fos–positive neurons. For each slice, we analyzed the whole field of view, excluding the slice edges, as they displayed DAPI staining artifacts.

**Analysis of Electrophysiological Data.** All electrophysiology data were analyzed using a custom code package written in Python. The precise time of lights on/off was determined by analysis of video recordings or using a light-sensitive resistor. All analyses were performed on the 10 min before and after transitions. Perievent time histograms were obtained by binning data in 0.25-s bins and normalizing data to the pretransition period. Firing rates were estimated by sliding a 1- or 2-min window in 20-s steps. Mean and SEM were estimated by averaging across days. To compare population data across transitions, we calculated the average firing rate in the 10 min before and after the transition without binning. For analysis restricted to a given behavioral state, we only considered transitions during which the animal was in that state for the whole 20 min (10 min before and after the transition). To estimate the number of individual neurons that consistently changed their firing rate in response to L-D and D-L transitions, we used a paired $t$ test to determine whether the neuron's firing followed a consistent pattern of change across multiple transitions. We used a bootstrap method to estimate the number of cells expected to pass our significance threshold by chance; for each iteration of the bootstrap, we chose a random time point within the first 24 h. We then created dummy transition times at 12-h intervals from that starting time point and used these dummy transition points to repeat the above analysis for each cell. This procedure was repeated 100 times (i.e., with 100 different dummy transition points) to obtain 100 values for the percentage of significantly changing cells. We used this dataset to estimate the mean and 95% CIs for this parameter. Only neurons that were followed through at least 4 transitions were used for analysis of circadian transitions. For noncircadian transitions, we analyzed neurons that experienced at least 6 transitions.

**Pairwise Correlations.** Each spike train was binned into spike counts of bin size 100 ms, generating a vector of spike counts for each cell. The spike count correlation coefficient $\rho$ for a pair of neurons was computed in 30-min episodes using a sliding window of 5 min. This produced 139 values for each neuron pair on every single half day (12 h of light and 12 h of darkness). The average of these values then determined the correlation value of each pair for every single half day:

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where $X$ and $Y$ represent the spike count vectors of 2 cells, respectively; $\mu_X$ and $\mu_Y$ are the means of $X$ and $Y$, respectively; $\sigma_X$ and $\sigma_Y$ denote the SDs of $X$ and $Y$, respectively; and $E$ is the expectation. This produced the matrices of pairwise spike count correlations on different half days. To generate the normalized correlation curve, correlations were normalized to the average correlation of each animal at P26 during the light period. Correlations in mixed behavioral states were computed with the above-stated method using the entire 12-h periods of light or dark, while correlations in wake only took into account the wake episodes. Results with bin size of 5 ms followed the same approach.

**Experimental Design.** Long–Evans rats of both sexes were used throughout all experiments. To estimate the effect of D-L and L-D transitions on firing rates of V1 neurons, we pooled data from previous experiments (13; $n$ = 7) as well as newly performed experiments ($n$ = 4 rats). Experimental design and timeline were the same across all of these experiments. This dataset was used to produce Figs. 1, 2, 3, and 4. For ISI analyses, we excluded 2 animals for which the precise transition times were known with uncertainty greater than 0.25 s. To analyze the effect of unexpected transitions ($n$ = 6 rats) as well as for prolonged darkness experiments ($n$ = 4 rats), we used datasets obtained from rats of a similar age (P24 to P35) as those used in analysis of circadian transitions. Electrophysiological data were acquired using the same electrode arrays and recording system in all experiments. For immunohistochemistry experiments, all rats ($n$ = 13 animals) were age matched to electrophysiological recordings, and all staining procedures were conducted in parallel across conditions to minimize variability. Slices from all conditions were imaged in every imaging session.

**Statistical Analyses.** To compare means of 2 populations, we used Wilcoxon rank sum tests. For paired data, both for firing rates and spike count correlations, comparisons were done using a Wilcoxon signed rank test. To compare a population mean to a given value (e.g., 0), we used 1-sample $t$ tests for normally distributed data and Wilcoxon signed rank tests for nonnormal distributions. Normality was tested using D'Agostino's $K^2$ test. To compare cumulative distributions, we used Kolmogorov–Smirnov tests. Data are represented as mean ± SEM. Box plots represent median ± interquartile range, with whiskers extending to the rest of the distribution.

**Code Accessibility.** All code used for analysis is available from the authors on request.


1. E. Zohary, M. N. Shadlen, W. T. Newsome, Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
2. M. N. Shadlen, W. T. Newsome, The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *J. Neurosci.* **18**, 3870–3896 (1998).
3. B. B. Averbeck, P. E. Latham, A. Pouget, Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
4. A. Arieli, D. Shoham, R. Hildesheim, A. Grinvald, Coherent spatiotemporal patterns of ongoing activity revealed by real-time optical imaging coupled with single-unit recording in the cat visual cortex. *J. Neurophysiol.* **73**, 2072–2093 (1995).

5. M. Tsodyks, T. Kenet, A. Grinvald, A. Arieli, Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* **286**, 1943–1946 (1999).
6. Y. H. Ch'ng, R. C. Reid, Cellular imaging of visual cortex reveals the spatial and functional organization of spontaneous activity. *Front. Integr. Nuerosci.* **4**, 20 (2010).
7. T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, A. Arieli, Spontaneously emerging cortical representations of visual attributes. *Nature* **425**, 954–956 (2003).
8. J. Fiser, C. Chiu, M. Weliky, Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature* **431**, 573–578 (2004).
9. J. N. MacLean, B. O. Watson, G. B. Aaron, R. Yuste, Internal dynamics determine the cortical response to thalamic stimulation. *Neuron* **48**, 811–823 (2005).
10. A. Luczak, P. Barthó, K. D. Harris, Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron* **62**, 413–425 (2009).
11. A. Luczak, P. Barthó, K. D. Harris, Gating of sensory input by spontaneous cortical activity. *J. Neurosci.* **33**, 1684–1695 (2013).
12. J. L. Gallant, C. E. Connor, D. C. Van Essen, Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport* **9**, 85–90 (1998).
13. K. B. Hengen, A. Torrado Pacheco, J. N. McGregor, S. D. Van Hooser, G. G. Turrigiano, Neuronal firing rate homeostasis is inhibited by sleep and promoted by wake. *Cell* **165**, 180–191 (2016).
14. K. D. Miller, D. J. MacKay, The role of constraints in Hebbian learning. *Neural Comput.* **6**, 100–126 (1994).
15. G. G. Turrigiano, S. B. Nelson, Homeostatic plasticity in the developing nervous system. *Nat. Rev. Neurosci.* **5**, 97–107 (2004).
16. K. B. Hengen, M. E. Lambo, S. D. Van Hooser, D. B. Katz, G. G. Turrigiano, Firing rate homeostasis in visual cortex of freely behaving rodents. *Neuron* **80**, 335–342 (2013).
17. T. Keck *et al*., Synaptic scaling and homeostatic plasticity in the mouse visual cortex in vivo. *Neuron* **80**, 327–334 (2013).
18. A. K. Dhawale *et al*., Automated long-term recording and analysis of neural activity in behaving animals. *eLife* **6**, e27702 (2017).
19. C. M. Niell, M. P. Stryker, Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28**, 7520–7536 (2008).
20. Y. Kawaguchi, Y. Kubota, Correlation of physiological subgroupings of nonpyramidal cells with parvalbumin- and calbindinD28k-immunoreactive neurons in layer V of rat frontal cortex. *J. Neurophysiol.* **70**, 387–396 (1993).
21. L. G. Nowak, R. Azouz, M. V. Sanchez-Vives, C. M. Gray, D. A. McCormick, Electrophysiological classes of cat primary visual cortical neurons in vivo as revealed by quantitative analyses. *J. Neurophysiol.* **89**, 1541–1566 (2003).
22. A. Kohn, Visual adaptation: Physiology, mechanisms, and functional benefits. *J. Neurophysiol.* **97**, 3155–3164 (2007).
23. S. C. Wissig, A. Kohn, The influence of surround suppression on adaptation effects in primary visual cortex. *J. Neurophysiol.* **107**, 3370–3384 (2012).
24. A. Benucci, A. B. Saleem, M. Carandini, Adaptation maintains population homeostasis in primary visual cortex. *Nat. Neurosci.* **16**, 724–729 (2013).
25. K. M. Rosen, M. A. McCormack, L. Villa-Komaroff, G. D. Mower, Brief visual experience induces immediate early gene expression in the cat visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5437–5441 (1992).
26. G. D. Mower, Differences in the induction of Fos protein in cat visual cortex during and after the critical period. *Brain Res. Mol. Brain Res.* **21**, 47–54 (1994).
27. B. Kaminska, L. Kaczmarek, A. Chaudhuri, Visual stimulation regulates the expression of transcription factors and modulates the composition of AP-1 in visual cortex. *J. Neurosci.* **16**, 3968–3978 (1996).
28. L. Kaczmarek, A. Chaudhuri, Sensory regulation of immediate-early gene expression in mammalian visual cortex: Implications for functional mapping and neural plasticity. *Brain Res. Brain Res. Rev.* **23**, 237–256 (1997).
29. D. P. Bartel, M. Sheng, L. F. Lau, M. E. Greenberg, Growth factors and membrane depolarization activate distinct programs of early response gene expression: Dissociation of fos and jun induction. *Genes Dev.* **3**, 304–313 (1989).
30. M. Sheng, G. McFadden, M. E. Greenberg, Membrane depolarization and calcium induce c-fos transcription via phosphorylation of transcription factor CREB. *Neuron* **4**, 571–582 (1990).
31. M. Sheng, M. E. Greenberg, The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron* **4**, 477–485 (1990).
32. I. V. Kaplan, Y. Guo, G. D. Mower, Immediate early gene expression in cat visual cortex during and after the critical period: Differences between EGR-1 and Fos proteins. *Brain Res. Mol. Brain Res.* **36**, 12–22 (1996).
33. Y. Yamada *et al*., Differential expression of immediate-early genes, c-fos and zif268, in the visual cortex of young rats: Effects of a noradrenergic neurotoxin on their expression. *Neuroscience* **92**, 473–484 (1999).
34. G. D. Mower, I. V. Kaplan, Immediate early gene expression in the visual cortex of normal and dark reared cats: Differences between fos and egr-1. *Brain Res. Mol. Brain Res.* **105**, 157–160 (2002).
35. M. C. D. Bridi *et al*., Two distinct mechanisms for experience-dependent homeostasis. *Nat. Neurosci.* **21**, 843–850 (2018).
36. D. H. Hubel, T. N. Wiesel, Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
37. D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
38. F. W. Campbell, B. G. Cleland, G. F. Cooper, C. Enroth-Cugell, The angular selectivity of visual cortical cells to moving gratings. *J. Physiol.* **198**, 237–250 (1968).
39. J. D. Pettigrew, The effect of visual experience on the development of stimulus specificity by kitten cortical neurones. *J. Physiol.* **237**, 49–74 (1974).
40. G. H. Henry, B. Dreher, P. O. Bishop, Orientation specificity of cells in cat striate cortex. *J. Neurophysiol.* **37**, 1394–1409 (1974).
41. J. A. Movshon, The velocity tuning of single units in cat striate cortex. *J. Physiol.* **249**, 445–468 (1975).
42. R. L. De Valois, E. W. Yund, N. Hepler, The orientation and direction selectivity of cells in macaque visual cortex. *Vision Res.* **22**, 531–544 (1982).
43. M. S. Gizzi, E. Katz, R. A. Schumer, J. A. Movshon, Selectivity for orientation and direction of motion of single neurons in cat striate and extrastriate visual cortex. *J. Neurophysiol.* **63**, 1529–1543 (1990).
44. M. Carandini, D. Ferster, Membrane potential and firing rate in cat primary visual cortex. *J. Neurosci.* **20**, 470–484 (2000).
45. W. E. Vinje, J. L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
46. B. Haider *et al*., Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron* **65**, 107–121 (2010).
47. R. Herikstad, J. Baker, J. P. Lachaux, C. M. Gray, S. C. Yen, Natural movies evoke spike trains with low spike time variability in cat primary visual cortex. *J. Neurosci.* **31**, 15844–15860 (2011).
48. M. Vinck, R. Batista-Brito, U. Knoblich, J. A. Cardin, Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron* **86**, 740–754 (2015).
49. M. Dipoppa *et al*., Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron* **98**, 602–615.e8 (2018).
50. D. M. Berson, F. A. Dunn, M. Takao, Phototransduction by retinal ganglion cells that set the circadian clock. *Science* **295**, 1070–1073 (2002).
51. D. M. Dacey *et al*., Melanopsin-expressing ganglion cells in primate retina signal colour and irradiance and project to the LGN. *Nature* **433**, 749–754 (2005).
52. T. M. Schmidt, P. Kofuji, Functional and morphological differences among intrinsically photosensitive retinal ganglion cells. *J. Neurosci.* **29**, 476–482 (2009).
53. T. M. Brown *et al*., Melanopsin contributions to irradiance coding in the thalamocortical visual system. *PLoS Biol.* **8**, e1000558 (2010).
54. K. E. Davis, C. G. Eleftheriou, A. E. Allen, C. A. Procyk, R. J. Lucas, Melanopsin-derived visual responses under light adapted conditions in the mouse dLGN. *PLoS One* **10**, e0123424 (2015).
55. A. E. Allen, R. Storchi, F. P. Martial, R. A. Bedford, R. J. Lucas, Melanopsin contributions to the representation of images in the early visual system. *Curr. Biol.* **27**, 1623–1632.e4 (2017).
56. A. J. Emanuel, M. T. H. Do, Melanopsin tristability for sustained and broadband phototransduction. *Neuron* **85**, 1043–1055 (2015).
57. A. E. Allen *et al*., Melanopsin-driven light adaptation in mouse vision. *Curr. Biol.* **24**, 2481–2490 (2014).
58. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
59. T. Egner, J. M. Monti, C. Summerfield, Expectation and surprise determine neural population responses in the ventral visual stream. *J. Neurosci.* **30**, 16601–16608 (2010).
60. M. G. Shuler, M. F. Bear, Reward timing in the primary visual cortex. *Science* **311**, 1606–1609 (2006).
61. L. Stănişor, C. van der Togt, C. M. Pennartz, P. R. Roelfsema, A unified selection signal for attention and reward in primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 9136–9141 (2013).
62. C. M. Niell, M. P. Stryker, Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65**, 472–479 (2010).
63. A. Goel, H. K. Lee, Persistence of experience-induced homeostatic synaptic plasticity through adulthood in superficial layers of mouse visual cortex. *J. Neurosci.* **27**, 6692–6700 (2007).
64. M. P. Blackman, B. Djukic, S. B. Nelson, G. G. Turrigiano, A critical and cell-autonomous role for MeCP2 in synaptic scaling up. *J. Neurosci.* **32**, 13529–13536 (2012).
65. Y. Karimipanah, Z. Ma, J. K. Miller, R. Yuste, R. Wessel, Neocortical activity is stimulus- and scale-invariant. *PLoS One* **12**, e0177396 (2017).
66. K. D. Harris, A. Thiele, Cortical state and attention. *Nat. Rev. Neurosci.* **12**, 509–523 (2011).
67. M. L. Schölvinck, A. B. Saleem, A. Benucci, K. D. Harris, M. Carandini, Cortical state determines global variability and correlations in visual cortex. *J. Neurosci.* **35**, 170–178 (2015).
68. D. S. Greenberg, A. R. Houweling, J. N. Kerr, Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nat. Neurosci.* **11**, 749–751 (2008).
69. S. Treue, Neural correlates of attention in primate visual cortex. *Trends Neurosci.* **24**, 295–300 (2001).
70. M. Goard, Y. Dan, Basal forebrain activation enhances cortical coding of natural scenes. *Nat. Neurosci.* **12**, 1444–1449 (2009).
71. D. L. Ringach, Spontaneous and driven cortical activity: Implications for computation. *Curr. Opin. Neurobiol.* **19**, 439–444 (2009).
72. A. Destexhe, Intracellular and computational evidence for a dominant role of internal network activity in cortical computations. *Curr. Opin. Neurobiol.* **21**, 717–725 (2011).
73. G. B. Keller, T. Bonhoeffer, M. Hübener, Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* **74**, 809–815 (2012).
74. A. Ayaz, A. B. Saleem, M. L. Schölvinck, M. Carandini, Locomotion controls spatial integration in mouse visual cortex. *Curr. Biol.* **23**, 890–894 (2013).
75. A. B. Saleem, A. Ayaz, K. J. Jeffery, K. D. Harris, M. Carandini, Integration of visual motion and locomotion in mouse visual cortex. *Nat. Neurosci.* **16**, 1864–1869 (2013).
76. K. D. Harris, D. A. Henze, J. Csicsvari, H. Hirase, G. Buzsáki, Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* **84**, 401–414 (2000).
77. N. Schmitzer-Torbert, J. Jackson, D. Henze, K. Harris, A. D. Redish, Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131**, 1–11 (2005).

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

Waitzmann, F.*, Wu, Y. K.* & Gjorgjieva, J. Top-down modulation in canonical cortical circuits with inhibitory short-term plasticity. *bioRxiv* (2023). https://doi.org/10.1101/2023.06.13.544791

# Top-down modulation in canonical cortical circuits with 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 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 show how response reversal relates to inhibition stabilization and the paradoxical effect in the presence of several short-term plasticity mechanisms demonstrating that response reversal coincides with a change in the indispensability of SST 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), 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 the 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).

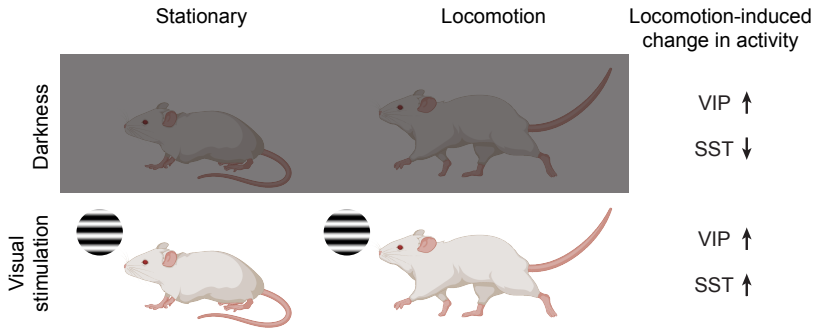


**Fig. 1.** Schematic diagrams illustrating that under different stimulus conditions, locomotion-induced modulatory input via VIP affects SST response oppositely. Top: in darkness, locomotion-induced top-down modulation increases VIP activity but decreases SST activity (Fu et al., 2014). Bottom: in the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

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 other factors, such as dynamically changing synapses, may contribute to nonlinear population responses like response reversal. 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, and synapses from different interneuron subtypes can undergo different short-term plastic changes (Campagnola et al., 2022). However, little is known about how these experimentally identified short-term plasticity mechanisms shape network dy-

namics 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. Increasing 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 interneuron-specific 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 SST is required for network stabilization. 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 sufficient biological detail and mathematical analysis and has 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), 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 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 re-

sponses can be approximated by a rectified linear function of the input (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-VIP STD, and SST-to-VIP short-term facilitation (STF) (Fig. 2A; Fig. S1). 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 $i \in \{E, P, S, V\}$ and $[\cdot]_+$ denotes linear rectification. $\tau_i$ 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 $\alpha$, 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-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. $\tau_x$ 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. $\tau_u$ 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 $\alpha$ to E and PV. Increasing $\alpha$ leads to a supralinear increase in the 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, for a low $\alpha$, 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 $\alpha$ 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 (Fig. 2D), 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. 2E, F). In contrast to networks with iSTP, the change of SST activity is largely unaffected for different values of $\alpha$ when iSTP is disabled (Fig. 2G). Interestingly, in our model, for a high $\alpha$ during the stimulation period, despite the increased activity of all inhibitory populations, the steady state of excitatory activity also increases (Fig. 2C). This observation appears to differ from what would be predicted by a classical disinhibition mechanism in which reducing inhibition increases excitatory activity. We confirm this by plotting the amount of recurrent excitation, recurrent inhibition, and the sum of recurrent excitation and inhibition that the excitatory population receives during the simulation(Fig. 2H, I). Surprisingly, even for a low $\alpha$, despite decreased SST activity, top-down modulation via VIP increases the total inhibition to the excitatory population at the steady state (Fig. 2H). Enhanced inhibition to the excitatory population at the steady state during top-down modulation is also observed for a high $\alpha$ (Fig. 2I). We systematically investigated the change in the input to the excitatory population due to top-down modulation at different levels of bottom-up input $\alpha$. We found that top-down modulation always increases the amount of inhibition to the excitatory population irrespective of whether it increases or decreases the activity of the SST population as $\alpha$ changes (Fig. 2J). These results suggest that rather than the decrease in the total inhibition, the increase in the recurrent excitation contributes to the elevated excitatory activity (Fig. 2H-J). Importantly, a joint 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 exhibit an increase in the total inhibitory inputs to the excitatory population (Fig. S3).

Taken together, our numerical simulation results reveal that experimentally identified forms of iSTP enable 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 $\alpha$. 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. S4A), $\delta g_V$ 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 given 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 derivative of the short-term plasticity variable with respect to the activity of population $j$, evaluated at the steady state (see SI Text).

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

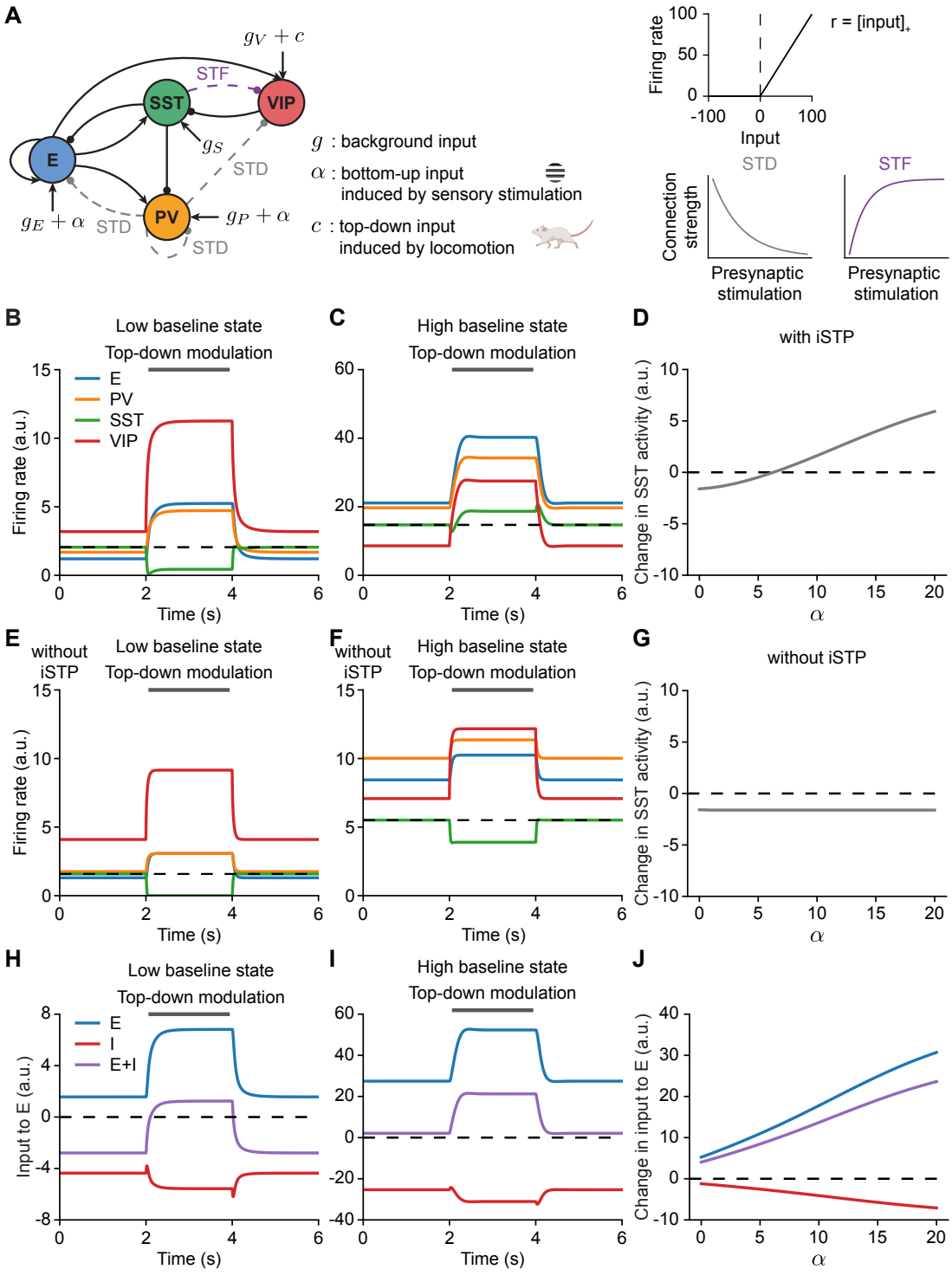

**Fig. 2.** Inhibitory short-term plasticity enables response reversal of SST induced by top-down modulation via VIP. (**A**) Schematic of network model with one excitatory (E) population and three distinct inhibitory populations, including PV, SST, and VIP. Short-term depressing (STD) and short-term facilitating (STF) connections are indicated by the dashed lines. Each population receives a background input $g$. E and PV receive bottom-up input $\alpha$ depending on sensory stimulation, and VIP receives top-down input $c$ during locomotion. Top right: rectified linear input-output function; Bottom right: cartoons showing how inhibitory connection strength changes with presynaptic stimulation under STD and STF. (**B**) Network responses to top-down modulation without any bottom-up input ($\alpha$ = 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 $\alpha$ = 15 corresponding to sensory stimulation. (**D**) Change in SST response induced by top-down modulation to VIP as a function of bottom-up input $\alpha$ in networks with iSTP. (**E**) Same as B but for networks without iSTP. (**F**) Same as C but for networks without iSTP. (**G**) Same as D but for networks without iSTP. (**H**) Input to the E population at $\alpha$ = 0. Different colors indicate different sources: input from the E population, input from the I populations, and the sum of the inputs from the E and I populations. (**I**) Same as E but at $\alpha$ = 15. (**J**) Change in different sources of recurrent inputs to the E population measured between baseline and at steady state during top-down modulation as a function of bottom-up input $\alpha$.

In agreement with our simulation results (Fig. 2B, C), we observed that $\mathbf{R}_{SV}$ changes its sign when calculated at different values of $\alpha$ (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 perturbations (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 $\alpha$. Note that while here we are interested in how neural activity changes in response to a given perturbation, several other studies have investigated the contributions of higher-order motifs to the perturbation-induced change of neural activity in excitatory and inhibitory networks (Pernice et al., 2011; Sadeh and Clopath, 2020).

As shown in Eq. 9, since $D$ and $\delta g_V$ 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:

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

where

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

and

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

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

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

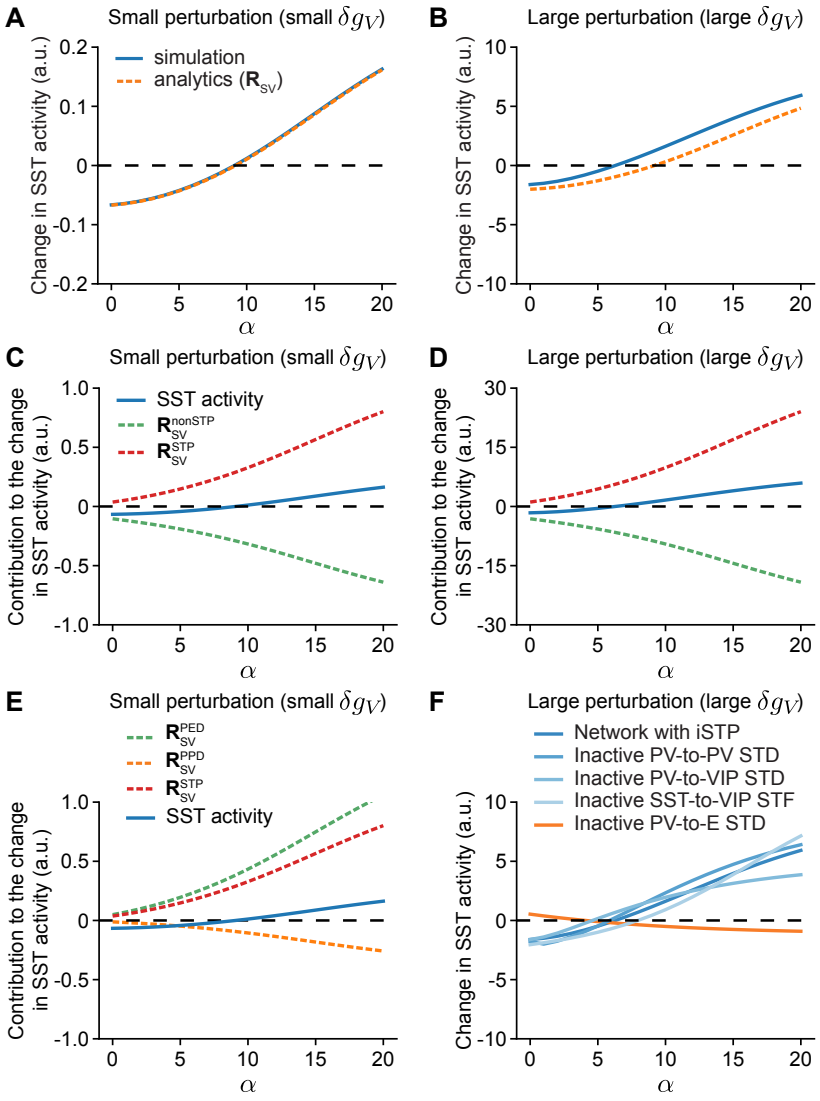


**Fig. 3.** Comparison between analytical predictions and numerical simulations on the change in SST activity and identification of PV-to-E STD as the crucial STP mechanism for the generation of response reversal. (**A**) Analytical prediction of the change in SST population response induced by the perturbation to VIP ($\mathbf{R}_{SV}$) matches closely with numerical simulation for a small perturbation. (**B**) Same as A but with a large perturbation. (**C**) Analytical contributions of the STP-dependent term $\mathbf{R}_{SV}^{STP}$ and the STP-independent term $\mathbf{R}_{SV}^{nonSTP}$ to the change in SST activity as a function of bottom-up input $\alpha$ for a small perturbation. (**D**) Same as C but with a large perturbation. (**E**) Analytical contributions of the PV-to-E STD-dependent term $\mathbf{R}_{SV}^{PED}$, the PV-to-PV STD-dependent term $\mathbf{R}_{SV}^{PPD}$, and the overall STD-dependent term $\mathbf{R}_{SV}^{STP}$ to the change in SST activity as a function of bottom-up input $\alpha$ for a small perturbation. (**F**) Change in SST response induced by top-down modulation to VIP as a function of bottom-up input $\alpha$ with a large perturbation for different network configurations marked with different colors. Here, for small perturbations $\delta g_V = 0.1$ and for large perturbations $\delta g_V = 3$.

As the short-term plasticity-independent term $\mathbf{K}_{SV}^{\text{nonSTP}}$ is governed by the static network weights, it is constant over the entire range of change in bottom-up input, i.e., $\mathbf{K}_{SV}^{\text{nonSTP}}$ does not change with $\alpha$. Note that because of Eq. 9 and since $D$ is always positive but subject to change in magnitude (Fig. S4A), $\mathbf{R}_{SV}^{\text{nonSTP}}$ changes in magnitude as well (Fig. 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. 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. S4B), 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 $\alpha$, whereas $\mathbf{K}_{SV}^{\text{PED}}$ increases with increasing bottom-up input $\alpha$ (see SI Text, Fig. S4C).

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. 3E). Therefore, when bottom-up input $\alpha$ 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 (Fig. 3E). As no terms directly associated with PV-to-VIP 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 while inactivating specific iSTP mechanisms to confirm our analysis. 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-VIP STD, or SST-to-VIP STF. Such networks show similar patterns to networks with intact iSTP (Fig. 3F). 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. 3F). Furthermore, we found that PV-to-E STD is crucial for generating the effective supralinear input-output relation observed in the baseline state for varying bottom-up input $\alpha$ (Fig. S2). Inactivating PV-to-E STD completely diminished the supralinearity of the effective input-output relations in contrast to inactivating other iSTP mechanisms (Fig. S5A). In addition, as bottom-up input increases, the resulting inhibitory current from PV to E is suppressed by PV-to-E STD (Fig. S5B). This suppression is greater for stronger bottom-up inputs leading to a sublinear increase in PV current, which is important for the generation of the effective supralinear input-output relation (Fig. S5B).

Taken together, our analysis and numerical simulations reveal 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 at a high baseline activity state (high $\alpha$), additional VIP inhibition induced by top-down modulation increases the activity of SST. This phenomenon 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

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

and $\delta g_S$ represents the change of input to SST. Furthermore,

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

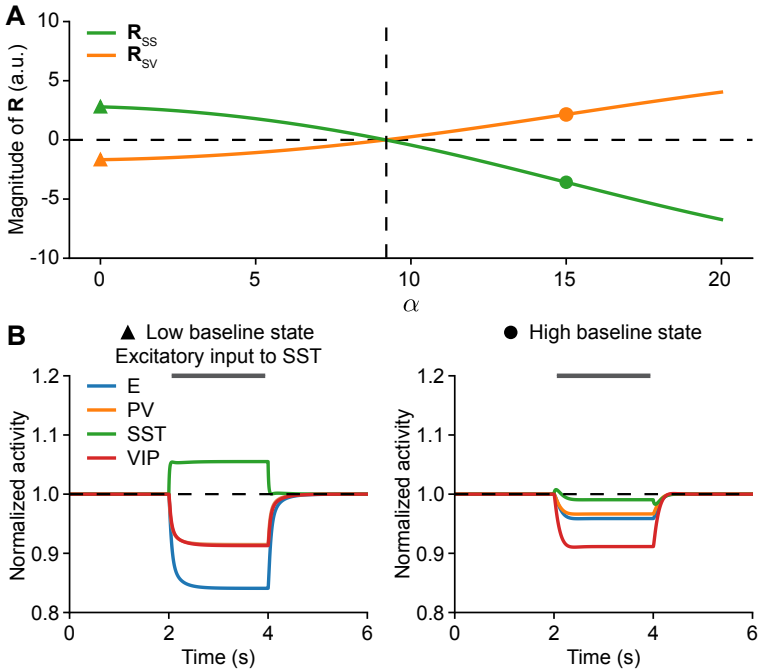


**Fig. 4.** Relationship between response reversal and paradoxical response of SST. (**A**) Analytical predictions of the change in SST response induced by an excitatory perturbation ($\delta g_S$) to SST, $\mathbf{R}_{SS}$, and change in SST response induced by an excitatory perturbation ($\delta g_V$) to VIP, $\mathbf{R}_{SV}$, as a function of bottom-up input $\alpha$. Here, $\delta g_V = \delta g_S = 3$. (**B**) Left: Normalized activity when injecting additional excitatory current into SST 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 at a high baseline state corresponding to $\alpha = 15$ marked with a dot in A. SST shows a paradoxical response.

When $\delta g_S$ 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 $\alpha$ 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. 4A, B left). In contrast, for high $\alpha$ 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. 4A, B right).

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 type of interneuron. Beyond identifying inhibition stabilization, we particularly aimed to ascertain the specific interneuron subtype that stabilizes the model networks in 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 of the network dynamics 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 in the model (defined as the network without any interneurons) is positive for all values of $\alpha$, suggesting that the E subnetwork is unstable and the network is inhibition-stabilized for all stimulation conditions (Fig. S6). Furthermore, we found that the E subnetwork being unstable (i.e. $J_{EE} > 1$) at the high bottom-up input is a necessary condition to observe response reversal (see SI Text). VIP does not stabilize the network, as the leading eigenvalue of the Jacobian of the E-VIP subnetwork (the network without PV and SST interneurons) is always positive (Fig. S6). By computing the leading eigenvalue of the Jacobian of the E-PV-VIP subnetwork (the network without SST interneurons), we found that the eigenvalue switches from negative to positive when the response of SST to top-down modulation is reversed (Fig. 5A), indicating that SST is required for network stabilization when top-down modulation via VIP increases SST activity. Furthermore, the leading negative eigenvalue of the Jacobian of the E-PV-VIP subnetwork for low $\alpha$ suggests that in the regime in which top-down modulation via VIP decreases SST activity, the network does not require SST for stabilization and can be stabilized by PV. To determine whether PV could be the only interneuron subtype stabilizing the network in that regime, we calculated the leading eigenvalue of the Jacobian of the E-SST-VIP subnetwork (the network without PV interneurons). We found that this eigenvalue is always negative in the current model (Fig. 5A), 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 when clamping PV activity, the fixed point in the given network is stable to perturbations over the entire range of $\alpha$ and reaches the same fixed point after the transient perturbation (Fig. 5B). In contrast, when clamping SST activity, while the fixed point at low $\alpha$ is stable to perturbations, a transient perturbation at high $\alpha$ leads to unstable dynamics (Fig. 5C).

Consistent with the change in the requirement of SST for network stabilization, we observed a

transition in the prevalence of inhibition received by the excitatory population from PV to SST with increasing $\alpha$ (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 in contrast to non-inhibition-stabilization networks (Fig. S3; Litwin-Kumar et al., 2016; Miller and Palmigiano, 2020; Wu and Gjorgjieva, 2023). In inhibition-stabilized networks with iSTP, 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.

To systematically investigate how response reversal and paradoxical effects of SST relate to interneuron-specific stabilization, we conducted mathematical analyses and found that $\mathbf{K}_{SV}$ and $\mathbf{K}_{SS}$ are linked to the determinant of the Jacobian of the E-PV-VIP network, det($\mathbf{M}_{\text{E-PV-VIP}}$) (see SI Text). In the network we considered here, because of the short-term plasticity mechanisms, the Jacobian of the E-PV-VIP subnetwork is a 6-by-6 matrix. When $\mathbf{K}_{SV}$ is positive (i.e., top-down modulation increases SST activity), $\mathbf{K}_{SS}$ is negative (i.e., the network exhibits paradoxical effects of SST), and det($\mathbf{M}_{\text{E-PV-VIP}}$) is negative. Note that in high-dimensional systems, the determinant of the Jacobian matrix alone is not sufficient to determine network stability. For a six-dimensional system, a negative det($\mathbf{M}_{\text{E-PV-VIP}}$) implies an odd number of positive eigenvalues corresponding to unstable eigenvectors/modes and thus the necessity for SST stabilization. However, the network can also require SST for stabilization in the presence of a positive det($\mathbf{M}_{\text{E-PV-VIP}}$), for instance, when the Jacobian of the E-PV-VIP subnetwork has an even number of unstable modes. To confirm the change in the number of unstable modes, we examined the second-largest eigenvalue of the E-PV-VIP subnetwork and found that the second-largest eigenvalue is always negative in the given network (Fig. 5D). As a result, the number of unstable modes changes from even to odd (Fig. 5E) when the SST response reverses from suppression to enhancement, and the network exhibits the paradoxical effect in the response of SST. Note that we did not find a direct mathematical relationship between PV stabilization and response reversal of SST (see SI Text). In other words, response reversal does not imply a change in PV stabilization. Consequently, PV stabilization and how it changes with bottom-up inputs, as presented in our study, are contingent on specific parameters.

Taken together, these results suggest that with increasing bottom-up input, representing a change in stimulation condition, the impact of top-down modulation on SST activity transitions from suppression to enhancement, the network exhibits a paradoxical response of SST, requires SST for

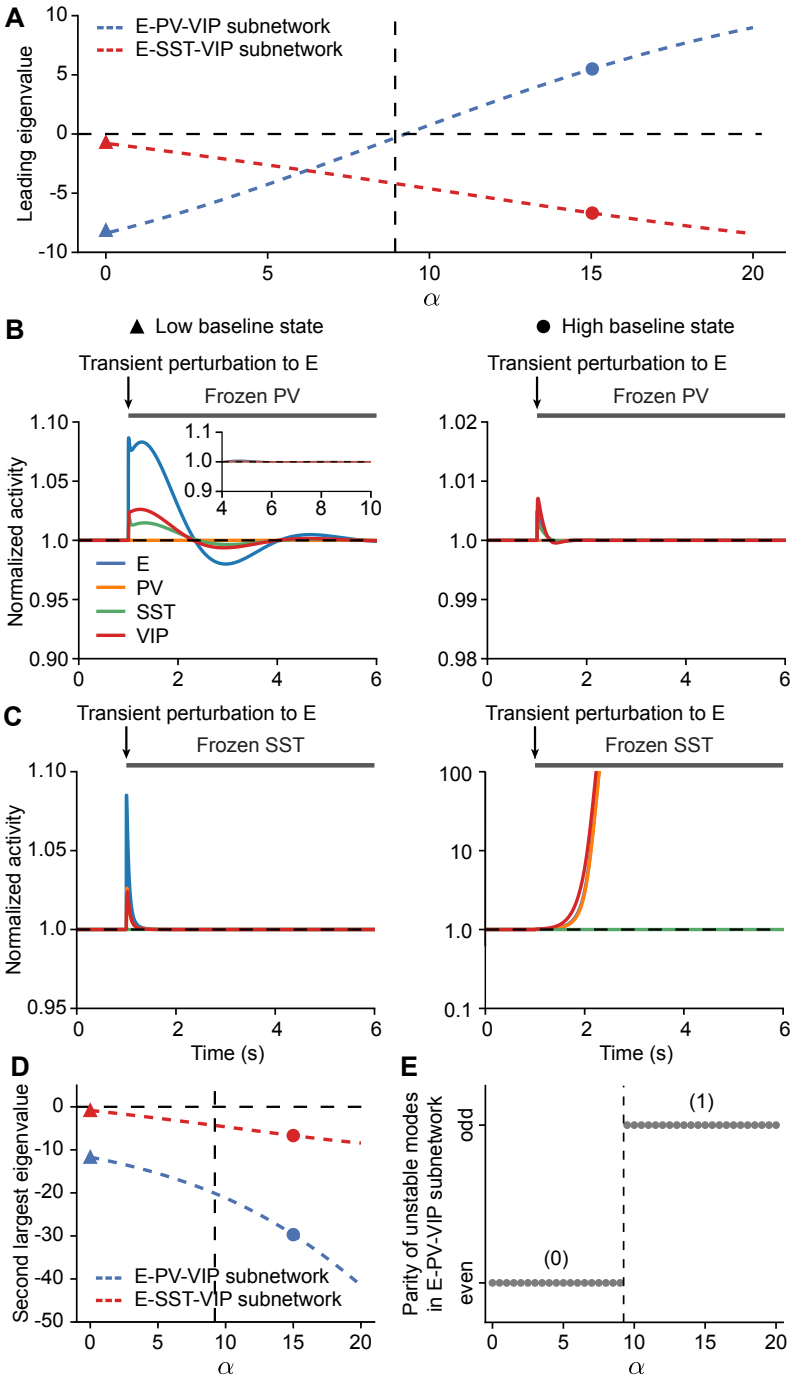


**Fig. 5.** The network undergoes a change in the indispensability of SST for network stabilization with increasing bottom-up input. (**A**) Leading eigenvalues of the E-PV-VIP subnetwork and the E-SST-VIP subnetwork as a function of bottom-up input $\alpha$. The response reversal boundary extracted from analytical calculations ($\mathbf{R}_{SV}$ = 0) is indicated by the vertical dashed line. (**B**) Left: Normalized activity when injecting an additional transient excitatory current into E while freezing PV for networks at a low baseline state corresponding to $\alpha$ = 0 marked with a triangle in A. The small transient excitatory input 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 $\alpha$ = 15 marked with a dot in A. (**C**) Similar to B but with frozen SST. (**D**) Same as A but for the second largest eigenvalue. (**E**) Parity of the number of unstable modes in the E-PV-VIP subnetwork as a function of bottom-up input $\alpha$. Numbers indicate the amount of unstable modes.

stabilization, and the E-PV-VIP subnetwork has an odd number of unstable modes (Figs. 4A, 5A, 6).

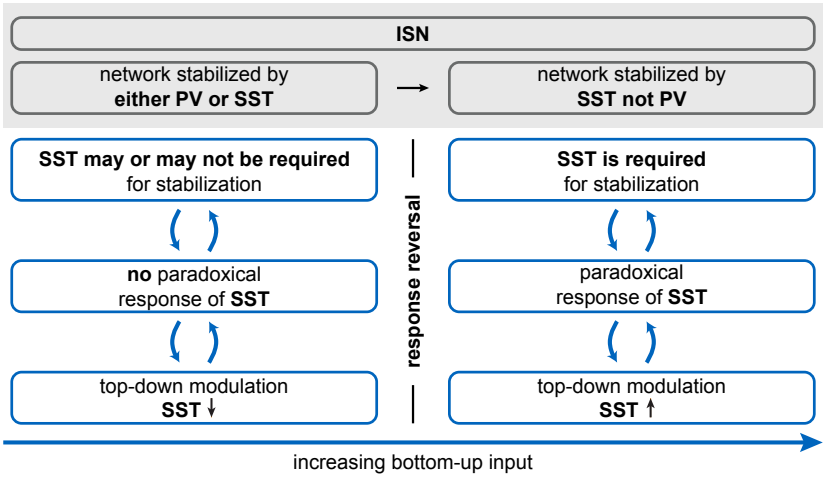


**Fig. 6.** The relationship between response reversal, paradoxical effect, and inhibition stabilization. At low bottom-up input, top-down modulation decreases SST activity, and the network does not exhibit a paradoxical response of SST such that SST may not be required for stabilization, or SST may be required for stabilization, but the E-PV-VIP subnetwork has an even number of unstable modes. As demonstrated in Fig. 5, in this regime, the network is inhibition stabilized and stabilized by either PV or SST. With increasing bottom-up input, the response of SST induced by top-down modulation is reversed from suppression to enhancement, the network exhibits a paradoxical response of SST and requires SST for stabilization with an odd number of unstable modes in the E-PV-VIP subnetwork. As demonstrated in Fig. 5, in this regime, the network is inhibition stabilized and stabilized by SST but not PV. Note that while the relationship between response reversal, paradoxical effects, and inhibition stabilization marked in blue boxes does not depend on the choice of parameters, the possible interneuron-specific stabilization regimes shaded in gray are contingent on specific parameters (see SI Text).

## Modeling results are robust to variations in short-term plasticity mechanisms, inputs, and network connectivity

To demonstrate that our results are valid for a variety of perturbations, we performed different sensitivity analyses on short-term plasticity mechanisms, inputs, and network connectivity.

We first investigated if additional short-term plasticity mechanisms affect our results. In this study, we used a rate-based population model, ignoring the large number of connections between individual neurons on a microscopic level. Given the dominant number of excitatory neurons in the cortex, we might have underestimated the effective depression of the E-to-E connection and facilitation of the E-to-SST connection compared to real circuits (Campagnola et al., 2022). We therefore sought to examine their influence on our results by analyzing how they might affect the analytical expression of $\mathbf{K}_{SV}$ and network simulations. We found that the response reversal of SST from suppression to enhancement with increasing bottom-up input, as reported experimentally, is preserved in the presence of E-to-E STD (Fig. 7A). However, the change in SST activity

evolves non-monotonically with increasing bottom-up input, starting to decrease and eventually being reversed from enhancement to suppression at high $\alpha$ (Fig. 7A). Due to E-to-E STD, the effective excitatory-to-excitatory coupling decreases, resulting in a stable E subnetwork, and the network eventually becomes a non-inhibition-stabilized network (non-ISN) as demonstrated by the leading eigenvalues of the E subnetwork and E-VIP subnetwork switching from positive to negative with increasing bottom-up input (Fig. 7B). Interestingly, the response reversal of SST from enhancement to suppression does not occur at the same time as the network transitions from ISN to non-ISN. We proved that being an ISN is a necessary but not a sufficient condition to generate enhanced SST activity induced by top-down modulation, and non-ISNs cannot generate enhanced SST activity induced by top-down modulation (see SI Text). Consistent with our previous results, in the presence of E-to-E STD, the paradoxical response of SST is also linked to the change in SST activity induced by top-down modulation (Fig. 7C). More specifically, the network exhibits (no) paradoxical response of SST when top-down modulation increases (decreases) SST activity (Fig. 7C).

Different from networks without E-to-E STD, by examining the leading eigenvalue of the E-PV-VIP and E-SST-VIP subnetwork (Fig. 7D), we observed a repertoire of interneuron-specific stabilization regimes and some novel regime transitions (Fig. 8). For instance, we observed a transition from being stabilized by PV but not SST (as reflected by a leading positive eigenvalue of the E-SST-VIP subnetwork and a leading negative eigenvalue of the E-PV-VIP subnetwork) to being stabilized by both PV and SST (as reflected by leading positive eigenvalues of the E-PV-VIP and E-SST-VIP subnetwork). We also observed a transition from being stabilized by SST but not PV (as reflected by a leading positive eigenvalue of the E-PV-VIP subnetwork and a leading negative eigenvalue of the E-SST-VIP subnetwork) to being stabilized by either PV or SST (as reflected by leading negative eigenvalues of the E-PV-VIP and E-SST-VIP subnetwork) (Fig. 7D, 8). We further confirmed these distinct regimes by injecting a transient excitatory perturbation into the excitatory population while clamping the activity of PV, or SST, or both PV and SST (Fig. S8). Despite novel regimes observed in the presence of E-to-E STD, the link between response reversal, paradoxical effects of SST, and the parity of unstable modes in the E-PV-VIP subnetwork remains unchanged (Fig. 7C-F). When top-down modulation decreases SST activity, the network exhibits no paradoxical response of SST, and the E-PV-VIP subnetwork has an even number of unstable modes. When top-down modulation increases SST activity, the network exhibits a paradoxical response of SST, and the E-PV-VIP subnetwork has an odd number of unstable modes implying that SST is required for network stabilization (Fig. 8).

In the presence of E-to-SST STF, the analytical expression of $\mathbf{K}_{\mathrm{SV}}$ (Eq. 10), dictating the sign of the change of SST induced by top-down modulation, remains unchanged. It does not contain any E-to-SST dependent terms, suggesting that the emergence of response reversal is unaffected by E-to-SST STF (Fig. S9A). Consistent with the analysis, our simulation results show that adding

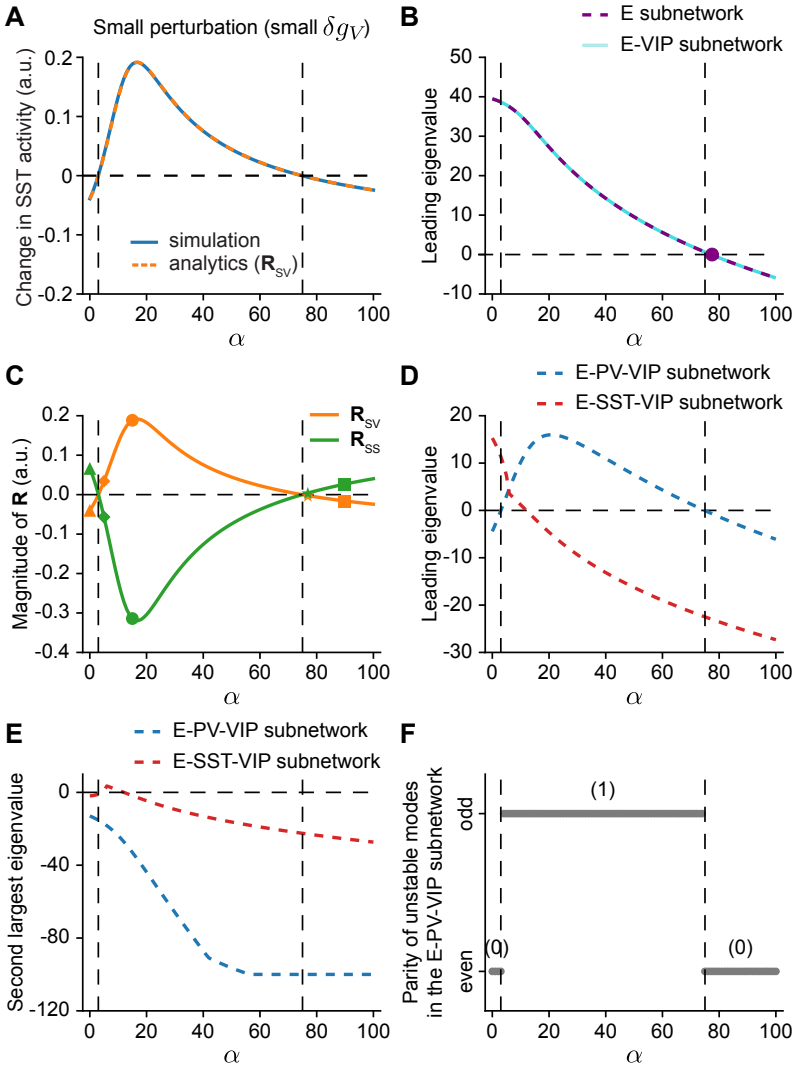


**Fig. 7.** Modeling results are robust in the presence of E-to-E STD. (**A**) Change in SST activity as a function of bottom-up input $\alpha$ for networks also including E-to-E STD, showing numerical results and analytical predictions. The response reversal boundaries extracted from analytical calculations ($\mathbf{R}_{SV}$ = 0) are indicated by the vertical dashed lines. Here, $\delta g_V$ = 0.1. (**B**) Leading eigenvalue of the E subnetwork and E-VIP subnetwork as a function of bottom-up input $\alpha$. The leading eigenvalue eventually turns negative, indicating that the network becomes non-inhibition stabilized. The dot represents the $\alpha$ level at which the leading eigenvalues are zero. (**C**) Relationship between response reversal and the paradoxical response of SST. Analytical predictions of the change in SST response induced by an excitatory perturbation ($\delta g_S$) to SST, $\mathbf{R}_{SS}$, and change in SST response induced by an excitatory perturbation ($\delta g_V$) to VIP, $\mathbf{R}_{SV}$, as a function of bottom-up input $\alpha$. Here, $\delta g_V = \delta g_S$ = 0.1. (**D**) Similar to B but for the E-PV-VIP subnetwork and the E-SST-VIP subnetwork. (**E**) Similar to D but for the second largest eigenvalues. (**F**) Parity of the number of unstable modes in the E-PV-VIP subnetwork as a function of bottom-up input $\alpha$. Numbers in the brackets indicate the amount of unstable modes.

E-to-SST STF does not alter the dynamics and the generation of response reversal (Fig. S9B, C). Given the omnipresence of short-term plasticity mechanisms in the mouse visual cortex amongst various populations and the centrality of SST to response reversal, we further incorporated short-term plasticity mechanisms in all connections considered in our model. These simulations show that response reversal can still be observed (Fig. S10). In addition, we examined how different inputs and network connectivity affect our results and found that response reversal is preserved in networks with varying inputs and connectivity strengths (Figs. S11, S12).

In conclusion, through multiple sensitivity analyses, we demonstrated the robustness of our findings to variations in short-term plasticity mechanisms, inputs, and network connectivity.

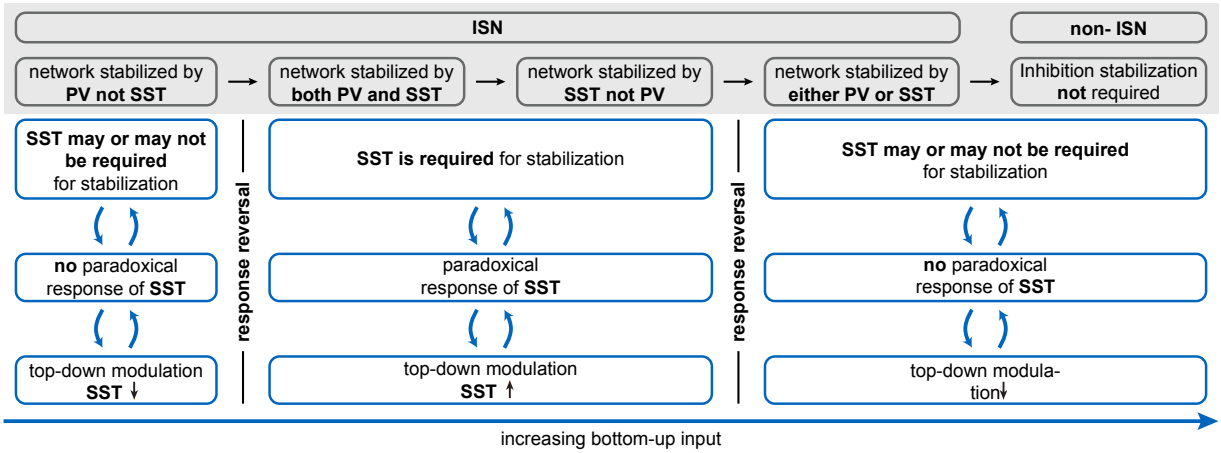


**Fig. 8.** The relationship between response reversal, inhibition stabilization, and paradoxical response in networks also including E-to-E STD. At the low bottom-up input, top-down modulation decreases SST activity, and the network does not exhibit a paradoxical response of SST; thus, SST may not be required for stabilization, or SST may be required for stabilization, but the E-PV-VIP subnetwork has an even number of unstable modes. As demonstrated in Fig. 7, in this regime, the network is inhibition stabilized and stabilized by PV but not SST. With increasing bottom-up input, the response of SST induced by top-down modulation is reversed from suppression to enhancement. The network further exhibits the paradoxical effect in the response of SST and requires SST for stabilization with an odd number of unstable modes in the E-PV-VIP subnetwork. As demonstrated in Fig. 7, in this regime, the network is inhibition stabilized and stabilized by both PV of SST and then transitions into being stabilized by SST but not PV. Further increasing bottom-up input, the response of SST induced by top-down modulation is reversed from enhancement to suppression, and the network does not exhibit a paradoxical response of SST: thus, SST may not be required for stabilization, or SST may be required for stabilization, but the E-PV-VIP subnetwork has an even number of unstable modes. As demonstrated in Fig. 7, in this regime, the network is inhibition stabilized and stabilized by either PV or SST and finally transitions into a non-ISN. Note that while the relationship between response reversal, paradoxical effects, and inhibition stabilization marked in blue boxes does not depend on the choice of parameters, the possible interneuron-specific stabilization regimes shaded in gray are contingent on specific parameters.

## 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.

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).

Concerning the modeled short-term plasticity mechanisms, our analysis primarily focused on PV-to-E STD, PV-to-PV STD, PV-to-VIP STD, and SST-to-VIP STF. Additional simulations of networks also including E-to-E STD, or E-to-SST STF, or STP mechanisms in all existing synapses demonstrated the robust occurrence of response reversal in SST. In addition to the incorporated 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). 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).

Finally, 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 can be 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 of 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 SST activity induced by top-down modulation gets elevated, 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 indispensability of SST for network stabilization. In darkness, when top-down modulation decreases SST activity, SST may not be required for network stabilization (i.e., solely PV can stabilize the network). In contrast, in the presence of visual stimuli, top-down modulation increases SST activity, and the network stabilization requires SST, and the E-PV-VIP subnetwork has an odd number of unstable modes. It is worth noting that when the network requires SST for stabilization, the network can require only SST but not PV (Fig. 5, S8) or both PV and SST for stabilization (Fig. S8). The novel observation that the network requires both PV and SST for stabilization is interesting and will require further investigation. 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 suggests that the specific inhibitory cell type stabilizing the network can change dynamically depending on the stimulation condition. As SST primarily targets dendrites of excitatory neurons (Kubota, 2014; Tremblay et al., 2016), stabilization through SST can be mechanistically realized via establishing a spatially precise E/I balance within individual dendritic segments. Recent experimental observations support the existence of such localized E/I balance at the dendritic segment level (Iascone et al., 2020), and SST is ideal for establishing the dendritic E/I balance and thus can provide an important source of stabilization. Furthermore, our results suggest a shift in inhibition source from PV to SST, typically accompanied by the occurrence of response reversal. Exploring the computational implications of this interneuron-specific inhibition shift and response reversal raises intriguing questions. Since PV neurons preferentially target perisomatic regions of excitatory neurons, whereas SST neurons target distal dendritic regions of excitatory neurons, the switch of dominant inhibition from soma to dendrite might prioritize inputs to perisomatic regions over inputs to distal dendritic regions and thus could be important for gating information (Udakis et al., 2020). Furthermore, inhibition also plays an important role in controlling plasticity (Letzkus et al., 2015). The redistribution of inhibition sources might imply different abilities of distinct interneurons to control plasticity in different regimes.

Last, for the network connectivity and the set of short-term plasticity mechanisms considered here, our results show that when SST is required for network stabilization and the E-PV-VIP subnetwork has an odd number of unstable modes, 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; 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). How paradoxical effects of a given cell type relate to the number of unstable modes in subnetworks excluding that cell type has been studied in networks without short-term plasticity in a recent theoretical study (Miller and Palmigiano, 2020). 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 (Eqs. 1 to 4) in matrix form while replacing the short-term plasticity variables with their steady-state values, 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 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. The steady states of the short-term plasticity variables are obtained by setting the Eqs. 5 to 8 to 0. 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. If short-term plasticity is not present on the synapses from $j$ to $i$, the corresponding element $\mathbf{P}_{ij}$ is 1 (for further details, see SI Text). By linearizing about the fixed point and ignoring higher-order terms, we obtain the following equation:

$$\mathbf{T}\frac{d}{dt}\delta\mathbf{r} = -\delta\mathbf{r} + \mathbf{F}(\mathbf{P} \circ \mathbf{J})\delta\mathbf{r} + \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\delta\mathbf{r} + \mathbf{F}\delta\mathbf{g}. \tag{21}$$

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 evaluated at the fixed point. $\mathbf{P}'$ 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.

The fixed point solution of Eq. 21 quantifies the change in population rates $\delta\mathbf{r}$ to an input perturbation $\delta\mathbf{g}$:

$$\begin{aligned}\delta\mathbf{r} &= \left(\mathbf{1} - \mathbf{F}\left(\mathbf{P} \circ \mathbf{J}\right) - \mathbf{F}\left(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r})\right)\right)^{-1}\mathbf{F}\,\delta\mathbf{g} \\ &= \frac{1}{\det(\mathbf{L})}\mathrm{adj}(\mathbf{L})\,\mathbf{F}\,\delta\mathbf{g},\end{aligned} \tag{22}$$

with

$$\mathbf{L} = \mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}\left(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r})\right), \tag{23}$$

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

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

$$\mathbf{R} = \frac{1}{\det \mathbf{L}} \mathrm{adj}(\mathbf{L})\, \mathbf{F}\, \delta\mathbf{G}. \tag{24}$$

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:

$$D = \frac{1}{\det(\mathbf{L})} = \frac{1}{\det(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r})))} \tag{25}$$

and

$$\mathbf{K} = \mathrm{adj}\,(\mathbf{L}) = \mathrm{adj}\,(\mathbf{1} - \mathbf{F}(\mathbf{P} \circ \mathbf{J}) - \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r})))\mathbf{F}. \tag{26}$$

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

$$\mathbf{R} = D\,\mathbf{K}\,\delta\mathbf{G}. \tag{27}$$

Note that if the network is stable, all eigenvalues of the Jacobian $\mathbf{T}^{-1}(-\mathbf{1}+\mathbf{F}(\mathbf{P} \circ \mathbf{J})+\mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r})))$ have negative real parts. Therefore, all eigenvalues of $\mathbf{L}$ have positive real parts, and $D$ is always positive (Fig. S4A).

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. Since the derivatives of the rectified linear input-output functions are 1 in regimes where all cell populations have positive firing rates, we have

$$\mathbf{R}_{SV} = D\,\mathbf{K}_{SV}\,\delta g_V \tag{28}$$

with

$$\begin{aligned}\mathbf{K}_{SV} = {} & (x^*_{PP} + x^{*'}_{PP} r_P)(J_{EE} - 1) J_{PP} J_{SV} \\ & - (x^*_{EP} + x^{*'}_{EP} r_P) J_{SV} J_{PE} J_{EP} + (J_{EE} - 1) J_{SV}.\end{aligned} \tag{29}$$

## 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. Top-down modulation was applied in the interval of 5 to 7 seconds. Networks were initialized using the parameters in the Supplementary Tables. Short-term plasticity variables were initially set to 1 and reached their steady-state values within the first second. Figures depict 6 seconds of network activity following 3 seconds of relaxation after initialization. Bottom-up input $\alpha$ was modeled in the interval $[0, 20]$ with a step size of 0.5 unless stated otherwise. All simulation parameters are listed in the Supplementary Tables.

### Data Availability

The code used for model simulations is available on GitHub https://github.com/comp-neural-circuits/top-down-modulation-with-iSTP.

## Contributions

F.W., Y.K.W., and J.G. designed research; F.W. and Y.K.W. performed research; F.W. and Y.K.W. contributed new reagents/analytic tools; F.W. and Y.K.W. analyzed data; F.W., Y.K.W., and J.G. wrote the paper.

## Acknowledgments

We thank JaeAnn Dwulet, 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.

## 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.

Iascone DM, Li Y, Sümbül U, Doron M, Chen H, Andreu V, et al. Whole-neuron synaptic mapping reveals spatially precise excitatory/inhibitory balance limiting dendritic and somatic spiking. Neuron 2020;106(4):566–578. doi: https://doi.org/10.1016/j.neuron.2020.02.015.

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.

Kubota Y. Untangling GABAergic wiring in the cortical microcircuit. Current Opinion in Neurobiology 2014;26:7–14. doi: https://doi.org/10.1016/j.conb.2013.10.003.

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.

Letzkus JJ, Wolff SB, Lüthi A. Disinhibition, a circuit mechanism for associative learning and memory. Neuron 2015;88(2):264–276. doi: https://doi.org/10.1016/j.neuron.2015.09.024.

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.

Pernice V, Staude B, Cardanobile S, Rotter S. How Structure Determines Correlations in Neuronal Networks. PLOS Computational Biology 2011 05;7(5):1–14. doi: https://doi.org/10.1371/journal.pcbi.1002059.

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. Theory of neuronal perturbome in cortical networks. Proceedings of the National Academy of Sciences 2020 Oct;117(43):26966–26976. doi: https://doi.org/10.1073/pnas.2004568117.

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.

Udakis M, Pedrosa V, Chamberlain SEL, Clopath C, Mellor JR. Interneuron-specific plasticity at parvalbumin and somatostatin inhibitory synapses onto CA1 pyramidal neurons shapes hippocampal output. Nature Communications 2020;11(1):4395. doi: https://doi.org/10.1038/s41467-020-18074-8.

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. Physical Review Research 2023;5(3):033023. doi: https://doi.org/10.1103/PhysRevResearch.5.033023.

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.

# Supporting Information

## Supporting Information Text

### 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 (Eqs. 1 to 4) in matrix form while replacing the short-term plasticity variables with their steady-state values, as follows:

$$\mathbf{T}\frac{d}{dt}\mathbf{r} = -\mathbf{r} + \mathbf{f}(\mathbf{P} \circ \mathbf{J}\mathbf{r} + \mathbf{g}), \tag{S1}$$

where $\mathbf{T}$ is a diagonal matrix of time constants of the firing rate dynamics, $\mathbf{r}$ a vector of firing rates of 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:

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

$$\mathbf{J} = \begin{bmatrix} J_{EE} & -J_{EP} & -J_{ES} & 0 \\ J_{PE} & -J_{PP} & -J_{PS} & 0 \\ J_{SE} & 0 & 0 & -J_{SV} \\ J_{VE} & -J_{VP} & -J_{VS} & 0 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_E + \alpha \\ g_P + \alpha \\ g_S \\ g_V \end{bmatrix}. \tag{S2}$$

The steady states of the short-term plasticity variables are obtained by setting the Eqs. 5 to 8 to 0 and given by:

$$x^*_{EP} = x^*_{PP} = x^*_{VP} = \frac{1}{1 + U_d r_P \tau_x}, \tag{S3}$$

$$u^*_{VS} = \frac{1 + U_f U_{\max} r_S \tau_u}{1 + U_f r_S \tau_u}. \tag{S4}$$

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. 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 and ignoring higher-order terms, we obtain the following equation:

$$\mathbf{T}\frac{d}{dt}\delta\mathbf{r} = -\delta\mathbf{r} + \mathbf{F}(\mathbf{P} \circ \mathbf{J})\delta\mathbf{r} + \mathbf{F}(\mathbf{P}' \circ \mathbf{J}\,\mathrm{diag}(\mathbf{r}))\delta\mathbf{r} + \mathbf{F}\delta\mathbf{g}. \tag{S5}$$

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 evaluated at the fixed point. $\mathbf{P}'$ 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. 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:

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

$$\text{diag}(\mathbf{r}) = \begin{bmatrix} r_E & 0 & 0 & 0 \\ 0 & r_P & 0 & 0 \\ 0 & 0 & r_S & 0 \\ 0 & 0 & 0 & r_V \end{bmatrix}, \ \delta\mathbf{g} = \begin{bmatrix} \delta g_E \\ \delta g_P \\ \delta g_S \\ \delta g_V \end{bmatrix}. \tag{S6}$$

The fixed point solution of Eq. S5 quantifies the change in population rates $\delta\mathbf{r}$ to an input perturbation $\delta\mathbf{g}$:

$$\begin{aligned} \delta\mathbf{r} &= \Big(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r}))\Big)^{-1}\mathbf{F}\,\delta\mathbf{g} \\ &= \frac{1}{\det\left(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r}))\right)}\text{adj}\left(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r}))\right)\mathbf{F}\,\delta\mathbf{g}, \end{aligned} \tag{S7}$$

where $\mathbf{1}$ denotes the identity matrix, and 'det' and 'adj' represent the matrix's determinant and adjugate, respectively. By replacing $\delta\mathbf{g}$ with a diagonal matrix $\delta\mathbf{G}$ whose diagonal elements are $\delta\mathbf{g}$, we can obtain a response matrix $\mathbf{R}$ as follows:

$$\mathbf{R} = \frac{1}{\det\left(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r}))\right)}\text{adj}\left(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r}))\right)\mathbf{F}\,\delta\mathbf{G} \tag{S8}$$

with

$$\delta\mathbf{G} = \begin{bmatrix} \delta g_E & 0 & 0 & 0 \\ 0 & \delta g_P & 0 & 0 \\ 0 & 0 & \delta g_S & 0 \\ 0 & 0 & 0 & \delta g_V \end{bmatrix}. \tag{S9}$$

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:

$$D = \frac{1}{\det(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r})))} \tag{S10}$$

and

$$\mathbf{K} = \text{adj}\left(\mathbf{1} - \mathbf{F}(\mathbf{P}\circ\mathbf{J}) - \mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\text{diag}(\mathbf{r}))\right)\mathbf{F}. \tag{S11}$$

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

$$\mathbf{R} = D\,\mathbf{K}\,\delta\mathbf{G}. \tag{S12}$$

Note that if the network is stable, all eigenvalues of the Jacobian $\mathbf{T}^{-1}(-\mathbf{1}+\mathbf{F}(\mathbf{P}\circ\mathbf{J})+\mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\mathrm{diag}(\mathbf{r})))$ have negative real parts. Therefore, all eigenvalues of $\mathbf{1}-\mathbf{F}(\mathbf{P}\circ\mathbf{J})-\mathbf{F}(\mathbf{P}'\circ\mathbf{J}\,\mathrm{diag}(\mathbf{r}))$ have positive real parts and $D$ is always positive (Fig. S4A).

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:

$$\mathbf{R}_{SV} = D\,\mathbf{K}_{SV}\,\delta g_V \tag{S13}$$

with

$$\begin{aligned}\mathbf{K}_{SV} = &\left((x^*_{PP}+x^{*'}_{PP}r_P)J_{EE}J_{PP}J_{SV}-(x^*_{EP}+x^{*'}_{EP}r_P)J_{SV}J_{PE}J_{EP}\right)f'_E f'_P f'_S f'_V\\ &-(x^*_{PP}+x^{*'}_{PP}r_P)J_{PP}J_{SV}f'_P f'_S f'_V+J_{EE}J_{SV}f'_E f'_S f'_V-J_{SV}f'_S f'_V.\end{aligned} \tag{S14}$$

To have response reversal of SST from suppression to enhancement, 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 $\delta g_V$ 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:

$$\begin{aligned}\mathbf{K}_{SV} = &(x^*_{PP}+x^{*'}_{PP}r_P)J_{EE}J_{PP}J_{SV}-(x^*_{EP}+x^{*'}_{EP}r_P)J_{SV}J_{PE}J_{EP}\\ &-(x^*_{PP}+x^{*'}_{PP}r_P)J_{PP}J_{SV}+J_{EE}J_{SV}-J_{SV}.\end{aligned} \tag{S15}$$

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

$$\mathbf{K}_{SV} = \mathbf{K}^{\mathrm{STP}}_{SV} + \mathbf{K}^{\mathrm{nonSTP}}_{SV} \tag{S16}$$

with

$$\mathbf{K}^{\mathrm{STP}}_{SV} = (x^*_{PP}+x^{*'}_{PP}r_P-1)(J_{EE}-1)J_{PP}J_{SV}-(x^*_{EP}+x^{*'}_{EP}r_P-1)J_{SV}J_{PE}J_{EP}, \tag{S17}$$

and

$$\mathbf{K}^{\mathrm{nonSTP}}_{SV} = (J_{EE}-1)(J_{PP}J_{SV}+J_{SV})-J_{SV}J_{PE}J_{EP}. \tag{S18}$$

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

$$\mathbf{K}^{\mathrm{STP}}_{SV} = \mathbf{K}^{\mathrm{PPD}}_{SV} + \mathbf{K}^{\mathrm{PED}}_{SV} \tag{S19}$$

with

$$\mathbf{K}_{SV}^{\mathrm{PPD}} = -(1 - x_{PP}^{*} - x_{PP}^{*'} r_P)(J_{EE} - 1) J_{PP} J_{SV} \tag{S20}$$

and

$$\mathbf{K}_{SV}^{\mathrm{PED}} = (1 - x_{EP}^{*} - x_{EP}^{*'} r_P) J_{SV} J_{PE} J_{EP}. \tag{S21}$$

Furthermore,

$$\begin{aligned} 1 - x_{EP}^{*} - x_{EP}^{*'} r_P &= 1 - x_{PP}^{*} - x_{PP}^{*'} r_P \\ &= 1 - x_{PP}^{*} + \frac{U_d \tau_x}{(1 + U_d \tau_x r_P)^2} r_P \\ &= 1 - \frac{1 + U_d \tau_x r_P}{(1 + U_d \tau_x r_P)^2} + \frac{U_d \tau_x r_P}{(1 + U_d \tau_x r_P)^2} \\ &= 1 - \frac{1}{(1 + U_d \tau_x r_P)^2} \\ &> 0. \end{aligned} \tag{S22}$$

The PV-to-E STD-dependent term $\mathbf{K}_{SV}^{\mathrm{PED}}$, therefore, 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. Taking the derivative of Eq. S22 with respect to $r_P$, we get:

$$\begin{aligned} \frac{d}{dr_P}(1 - x_{EP}^{*} - x_{EP}^{*'} r_P) &= \frac{d}{dr_P}(1 - x_{PP}^{*} - x_{PP}^{*'} r_P) \\ &= \frac{d}{dr_P}\left(1 - \frac{1}{(1 + U_d \tau_x r_P)^2}\right) \\ &= \frac{2 U_d \tau_x}{(1 + U_d \tau_x r_P)^3} \\ &> 0, \end{aligned} \tag{S23}$$

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 $\alpha$ (Fig. S2), $\mathbf{K}_{SV}^{\mathrm{PED}}$ increases (decreases) while $\mathbf{K}_{SV}^{\mathrm{PPD}}$ decreases (increases) (Fig. S4C). For response reversal of SST from negative to positive with increasing $\alpha$, $\mathbf{K}_{SV}$ needs to switch from negative to positive. Therefore, $\mathbf{K}_{SV}^{\mathrm{PED}}$, rather than $\mathbf{K}_{SV}^{\mathrm{PPD}}$, is the critical short-term plasticity term for response reversal from suppression to enhancement.

### Interneuron-specific stabilization property

Network stabilization can be determined by the leading eigenvalue of the Jacobian of the system. The Jacobian of our network with iSTP is given by:

$$\mathbf{M} = \begin{bmatrix} \frac{J_{EE}-1}{\tau_E} & -\frac{x^*_{EP}J_{EP}}{\tau_E} & -\frac{J_{ES}}{\tau_E} & -\frac{J_{EV}}{\tau_E} & -\frac{J_{EP}r_P}{\tau_E} & 0 & 0 & 0 \\ \frac{J_{PE}}{\tau_P} & \frac{-1-x^*_{PP}J_{PP}}{\tau_P} & -\frac{J_{PS}}{\tau_P} & -\frac{J_{PV}}{\tau_P} & 0 & -\frac{J_{PP}r_P}{\tau_P} & 0 & 0 \\ \frac{J_{SE}}{\tau_S} & -\frac{J_{SP}}{\tau_S} & \frac{-J_{SS}-1}{\tau_S} & -\frac{J_{SV}}{\tau_S} & 0 & 0 & 0 & 0 \\ \frac{J_{VE}}{\tau_V} & -\frac{x^*_{VP}J_{VP}}{\tau_V} & -\frac{u^*_{VS}J_{VS}}{\tau_V} & \frac{-J_{VV}-1}{\tau_V} & 0 & 0 & -\frac{J_{VP}r_P}{\tau_V} & \frac{-J_{VS}r_S}{\tau_V} \\ 0 & -U_d x^*_{EP} & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 & 0 & 0 \\ 0 & -U_d x^*_{PP} & 0 & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 & 0 \\ 0 & -U_d x^*_{VP} & 0 & 0 & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 \\ 0 & 0 & U_f(U_{\max} - u^*_{VS}) & 0 & 0 & 0 & 0 & -\frac{1}{\tau_u} - U_f r_S \end{bmatrix}. \tag{S24}$$

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

To identify if the network is inhibition stabilized, we computed the leading eigenvalue of the Jacobian of the E subnetwork. In this case, all inhibitory populations and their related STP variables 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, i.e. in the ISN regime. 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.

Since

$$\begin{aligned} x^*_{EP} + x^{*'}_{EP} r_P &= x^*_{PP} + x^{*'}_{PP} r_P \\ &= \frac{1 + U_d \tau_x r_P}{(1 + U_d \tau_x r_P)^2} - \frac{U_d \tau_x r_P}{(1 + U_d \tau_x r_P)^2} \\ &= \frac{1}{(1 + U_d \tau_x r_P)^2} \\ &> 0, \end{aligned} \tag{S25}$$

it is easy to see that $\mathbf{K}_{SV}$ (Eq. 29) is negative when $J_{EE} - 1 < 0$. This implies that for any network whose E subnetwork is stable (i.e., the entire network is not inhibition stabilized), top-down modulation via VIP decreases SST activity. The network, therefore, needs to be inhibition stabilized, i.e. the E subnetwork has to be unstable, to obtain an enhanced effect of SST response induced by top-down modulation.

To identify if the network is inhibition stabilized by a specific interneuron subtype, we compute the leading eigenvalue of the subnetwork in which a specific interneuron subtype 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 required (not required) for stabilization. This interneuron-specific stabilization property can also be probed by, at the same time, freezing the respective inhibitory population, transiently perturbing the excitatory population and observing potential changes in the fixed point dynamics.

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

$$\mathbf{M}_{\text{E-PV-VIP}} = \begin{bmatrix} \frac{J_{EE}-1}{\tau_E} & -\frac{x^*_{EP}J_{EP}}{\tau_E} & -\frac{J_{EV}}{\tau_E} & -\frac{J_{EP}r_P}{\tau_E} & 0 & 0 \\ \frac{J_{PE}}{\tau_P} & \frac{-1-x^*_{PP}J_{PP}}{\tau_P} & -\frac{J_{PV}}{\tau_P} & 0 & -\frac{J_{PP}r_P}{\tau_P} & 0 \\ \frac{J_{VE}}{\tau_V} & -\frac{x^*_{VP}J_{VP}}{\tau_V} & \frac{-J_{VV}-1}{\tau_V} & 0 & 0 & -\frac{J_{VP}r_P}{\tau_V} \\ 0 & -U_d x^*_{EP} & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 & 0 \\ 0 & -U_d x^*_{PP} & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P & 0 \\ 0 & -U_d x^*_{VP} & 0 & 0 & 0 & -\frac{1}{\tau_x} - U_d r_P \end{bmatrix}. \tag{S26}$$

Therefore, the determinant of the Jacobian is given by

$$\begin{aligned} \det(\mathbf{M}_{\text{E-PV-VIP}}) = &\left(\frac{J_{EE}-1}{\tau_E}\right)\left(\frac{-1-x^*_{PP}J_{PP}}{\tau_P}\right)\left(\frac{-1}{\tau_V}\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-\frac{1}{\tau_x}-U_d r_P\right) \\ &-\left(\frac{J_{EE}-1}{\tau_E}\right)\left(-\frac{J_{PP}r_P}{\tau_P}\right)\left(\frac{-1}{\tau_V}\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-U_d x^*_{PP}\right)\left(-\frac{1}{\tau_x}-U_d r_P\right) \\ &-\left(-\frac{x^*_{EP}J_{EP}}{\tau_E}\right)\left(\frac{J_{PE}}{\tau_P}\right)\left(\frac{-1}{\tau_V}\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-\frac{1}{\tau_x}-U_d r_P\right) \\ &+\left(-\frac{J_{EP}r_P}{\tau_E}\right)\left(\frac{J_{PE}}{\tau_P}\right)\left(\frac{-1}{\tau_V}\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-\frac{1}{\tau_x}-U_d r_P\right)\left(-\frac{1}{\tau_x}-U_d r_P\right), \end{aligned} \tag{S27}$$

which becomes:

$$\begin{aligned} \det(\mathbf{M}_{\text{E-PV-VIP}}) &= \frac{1}{\tau_E\tau_P\tau_V}\left(-\frac{1}{\tau_x x^*_{PP}}\right)^3\Big[(x^*_{PP}+x^{*'}_{PP}r_P-1)(J_{EE}-1)J_{PP}-(x^*_{EP}+x^{*'}_{EP}r_P-1)J_{EP}J_{PE} \\ &\quad +(J_{EE}-1)(J_{PP}+1)-J_{PE}J_{EP}\Big] \\ &= -\frac{1}{\tau_E\tau_P\tau_V}\left(-\frac{1}{\tau_x x^*_{PP}}\right)^3\mathbf{K}_{SS} \\ &= \frac{1}{\tau_E\tau_P\tau_V}\left(-\frac{1}{\tau_x x^*_{PP}}\right)^3\frac{\mathbf{K}_{SV}}{J_{SV}}. \end{aligned} \tag{S28}$$

In the network we considered here, because of the short-term plasticity mechanisms, the Jacobian of the E-PV-VIP subnetwork is a 6-by-6 matrix. Since det($\mathbf{M}_{\text{E-PV-VIP}}$) is the product of the eigenvalues of the Jacobian for the E-PV-VIP subnetwork, a negative det($\mathbf{M}_{\text{E-PV-VIP}}$) implies an odd number of positive eigenvalues and consequently an odd number of unstable eigenvectors/modes in the E-PV-VIP subnetwork, suggesting that SST is required for network stabilization. However, the network may also require SST for stabilization if det($\mathbf{M}_{\text{E-PV-VIP}}$) is positive, for instance, when the Jacobian of the E-PV-VIP subnetwork has an even number of unstable modes. At the same time, an odd number of unstable modes in the E-PV-VIP subnetwork 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. In contrast, an even number of unstable modes in the E-PV-VIP subnetwork implies a negative $\mathbf{K}_{SS}$ and a positive $\mathbf{K}_{SV}$.

Further, to investigate the relationship between PV stabilization and response reversal of SST, we compute the Jacobian for the E-SST-VIP subnetwork, which is given by:

$$\mathbf{M}_{\text{E-SST-VIP}} = \begin{bmatrix} \frac{J_{EE}-1}{\tau_E} & -\frac{J_{ES}}{\tau_E} & -\frac{J_{EV}}{\tau_E} & 0 \\ \frac{J_{SE}}{\tau_S} & -\frac{J_{SS}-1}{\tau_S} & -\frac{J_{SV}}{\tau_S} & 0 \\ \frac{J_{VE}}{\tau_V} & -\frac{u^*_{VS}J_{VS}}{\tau_V} & \frac{-J_{VV}-1}{\tau_V} & 0 \\ 0 & -U_f(U_{\max}-u^*_{VS}) & 0 & -\frac{1}{\tau_u}-U_f r_S \end{bmatrix}. \tag{S29}$$

Its determinant is then given by

$$\begin{aligned} \det(\mathbf{M}_{\text{E-SST-VIP}}) = -\frac{1}{\tau_E\tau_P\tau_V}\left(\frac{1}{\tau_u}-U_f r_s\right)\Big[&(J_{EE}-1)(1-J_{SV}u^*_{VS}J_{VS})+J_{EV}J_{SE}u^*_{VS}J_{VS} \\ &+J_{ES}J_{SE}+J_{ES}J_{VE}-J_{EV}J_{VE}\Big]. \end{aligned} \tag{S30}$$

We found no relation between the determinant of the E-SST-VIP subnetwork det($\mathbf{M}_{\text{E-SST-VIP}}$) and the response reversal condition $\mathbf{K}_{SV}$. The determinant can switch its sign independent of the response of SST and vice versa. This implies that the requirement of PV for network stabilization cannot be linked to the response reversal condition of SST. As a result, the PV stabilization property can change independently with different bottom-up inputs and is thus parameter-dependent.

### Networks also including E-to-E STD

In the presence of E-to-E STD, we have

$$\begin{aligned} \mathbf{K}_{SV} = &(x^*_{PP}+x^{*'}_{PP}r_P-1)J_{PP}\left[(x^*_{EE}+x^{*'}_{EE}r_E)J_{EE}-1\right]J_{SV}-(x^*_{EP}+x^{*'}_{EP}r_P-1)J_{SV}J_{PE}J_{EP} \\ &+\left[(x^*_{EE}+x^{*'}_{EE}r_E)J_{EE}-1\right](J_{PP}J_{SV}+J_{SV})-J_{SV}J_{PE}J_{EP}. \end{aligned} \tag{S31}$$

with

$$x^*_{EE} = \frac{1}{1+\tau_x U_d r_E} \tag{S32}$$

and

$$x^{*'}_{EE} = -\frac{U_d \tau_x}{(1 + U_d \tau_x r_E)^2}. \tag{S33}$$

Furthermore, we have

$$\begin{aligned}\mathbf{K}_{SS} = &-\Big[(x^*_{PP} + x^{*'}_{PP} r_P - 1) J_{PP} \big[(x^*_{EE} + x^{*'}_{EE} r_E) J_{EE} - 1\big] - (x^*_{EP} + x^{*'}_{EP} r_P - 1) J_{PE} J_{EP} \\ &+ \big[(x^*_{EE} + x^{*'}_{EE} r_E) J_{EE} - 1\big](J_{PP} + 1) - J_{PE} J_{EP}\Big] \\ = &-\mathbf{K}_{SV}/J_{SV}.\end{aligned} \tag{S34}$$

Note that in the presence of E-to-E STD, we still have

$$\mathbf{R}_{SV} = D\,\mathbf{K}_{SV}\,\delta g_V, \tag{S35}$$

$$\mathbf{R}_{SS} = D\,\mathbf{K}_{SS}\,\delta g_S, \tag{S36}$$

$$\mathbf{R}_{SS} = -\frac{\delta g_S}{J_{SV}\delta g_V}\mathbf{R}_{SV}. \tag{S37}$$

Despite the fact that E-to-E STD affects the amplitude of $D$, the relationship between $\mathbf{R}_{SV}$ and $\mathbf{R}_{SS}$ remains unchanged.

To investigate how $\mathbf{K}_{SV}$ relates to inhibition stabilization, we analyzed the Jacobian of the E-to-E subnetwork now including E-to-E STD, $\mathbf{M}_{EE}$, given by:

$$\mathbf{M}_{EE} = \begin{bmatrix} xJ_{EE} - 1 & J_{EE} r_E \\ -U_d x & -\frac{1+\tau_x U_d r_E}{\tau_x} \end{bmatrix}. \tag{S38}$$

If the E subnetwork is stable (i.e., the entire network is not inhibition stabilized), the determinant of the Jacobian $\det(\mathbf{M}_{EE})$ is positive, namely,

$$\det(\mathbf{M}_{EE}) = (xJ_{EE} - 1)(-\frac{1 + \tau_x U_d r_E}{\tau_x}) - J_{EE} r_E(-U_d x) > 0. \tag{S39}$$

At the steady state, the above equation can be written as follows

$$(x^*_{EE} + x^{*'}_{EE} r_E) J_{EE} - 1 < 0. \tag{S40}$$

Since $x^*_{PP} + x^{*'}_{PP} r_P - 1$ and $x^*_{EP} + x^{*'}_{EP} r_P - 1$ are positive (Eq. S22), it is easy to see that when $(x^*_{EE} + x^{*'}_{EE} r_E) J_{EE} - 1 < 0$ (for instance, when the network is not inhibition stabilized), $\mathbf{K}_{SV}$ (Eq. S31) is always negative. Therefore, to have increased SST activity induced by top-down modulation (i.e., $\mathbf{K}_{SV} > 0$), the network has to be inhibition stabilized.

Furthermore, as bottom-up input increases, $r_E$ also increases. Since

$$(x^*_{EE} + x^{*'}_{EE} r_E) J_{EE} - 1 = \frac{J_{EE}}{(1 + \tau_x U_d r_E)^2} - 1, \tag{S41}$$

to generate elevated SST activity by top-down modulation at a high level of bottom-up input, $\frac{J_{EE}}{(1+\tau_x U_d r_E)^2} - 1$ needs to be positive. As a result, $J_{EE}$ has to be large and/or $\tau_x U_d$ has to be small. Note that as shown in Eq. (S32), greater $\tau_x U_d$ implies a stronger depression of E-to-E connection strength.

### The Response Reversal Index

In the exploration of the robustness of our findings under various perturbations, we conducted sensitivity analyses encompassing inputs, network connectivity, and short-term plasticity mechanisms. For a concise evaluation of the transition in SST response from suppression to enhancement induced by top-down modulation and the consequent emergence of response reversal, we introduce the Response Reversal Index (RRI) as follows:

$$\text{RRI} = \begin{cases} \Delta r_{SST}^{\alpha=20} - \Delta r_{SST}^{\alpha=0}, \text{ if } (\Delta r_{SST}^{\alpha=0} \cdot \Delta r_{SST}^{\alpha=20} < 0) \& (\Delta r_{SST}^{\alpha=20} - \Delta r_{SST}^{\alpha=0} > 0), \text{response reversal} \\ 0, \text{ no response reversal} \\ -2, \text{ if at least one population is silent} \\ -4, \text{ if network is unstable} \end{cases}, \tag{S42}$$

where $\Delta r_{SST}^{\alpha=i}$ represents the change in SST response induced by top-down modulation at the bottom-up input $i$. Note that a positive RRI indicates a response reversal of SST from negative to positive.

### Sensitivity analysis to background inputs and connectivity matrix

We tested whether our results were sensitive to the choices of background inputs $g$ and the ratio of $g$ and top-down modulatory input $c$. We varied the mean background input $\bar{g}$ (i.e., the average background input to four different populations) by randomly and independently sampling the corresponding background inputs to individual populations from a uniform distribution (see Methods). We found that response reversal is observed at different levels of background input (Fig. S11A). Furthermore, we covaried the top-down modulatory input $c$ and found that our results are robust for a wide range of ratios between background input and top-down modulatory input (Fig. S11B). These results suggest that our results remain robust across a wide range of bottom-up and top-down inputs.

Although our initial network connectivity is constrained by previous experimental studies (Pfeffer et al., 2013), strengths of individual synapses have huge variability. We therefore examined whether our results hold for a variety of networks with different connectivity strengths. We varied the excitatory to excitatory connectivity strength $J_{EE}$ and found that non-inhibition stabilized networks (i.e., the E subnetwork is stable) are unable to generate response reversal (Fig. S12A).

We mathematically proved that the network needs to be inhibition stabilized (i.e. the E subnetwork has to be unstable) to obtain an enhanced effect on SST response induced by top-down modulation (see Methods). For various $J_{EE}$, the inhibition-stabilized network can exhibit response reversal (Fig. S12A). Further increasing $J_{EE}$ eventually leads to unstable networks, which can be prevented by increasing inhibitory weights. This also rescues response reversal (Fig. S12B). As our model suggests a shift in the primary source of inhibition from PV to SST when response reversal is observed, we examined how the ratio between $J_{EP}$ and $J_{ES}$ affects our conclusions. We found that a large fraction of the sampled inhibition-stabilized networks with a wide range of ratios between $J_{EP}$ and $J_{ES}$ are capable of generating response reversal (Fig. S12C, D). As the PV-to-E connection is subject to short-term depression, we then examined how the effective ratio between $J_{EP}$ and $J_{ES}$ affects response reversal. We, therefore, multiplied $J_{EP}$ with the corresponding short-term plasticity variable $x_{EP}$. Response reversal is observed for effective ratios larger than 1 as well as smaller than 1 (Fig. S12E). This is also the case for a wide range of connection strengths between E and PV (Fig. S12F), between E and SST (Fig. S12G), as well as $J_{SE}$ and $J_{PE}$ (Fig. S12H). These findings indicate the robustness of our results across inhibition-stabilized networks with varying connectivity strengths.

## Supplementary Figures

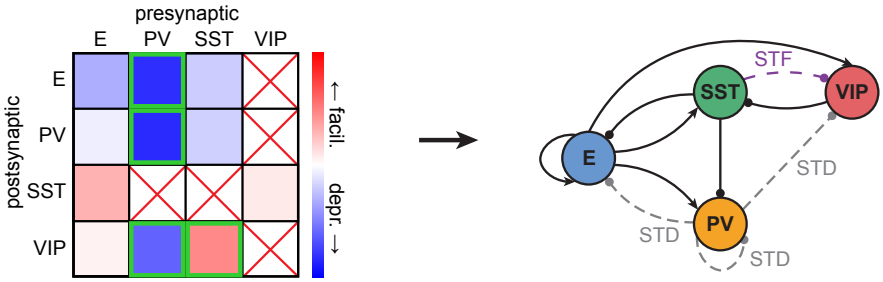


**Fig. S1.** Short-term plasticity mechanisms included in the network model. Left: Different degrees of short-term facilitation (STF) and depression (STD) at different synapses measured by the Allen Institute (Campagnola et al., 2022). Green boxes indicate the four most pronounced mechanisms, the only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

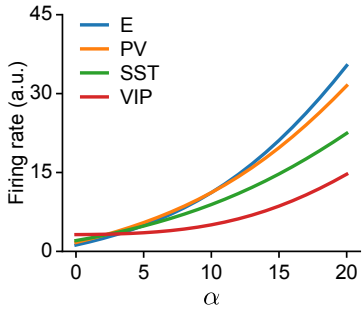


**Fig. S2.** Network activity as a function of bottom-up input $\alpha$ to E and PV.

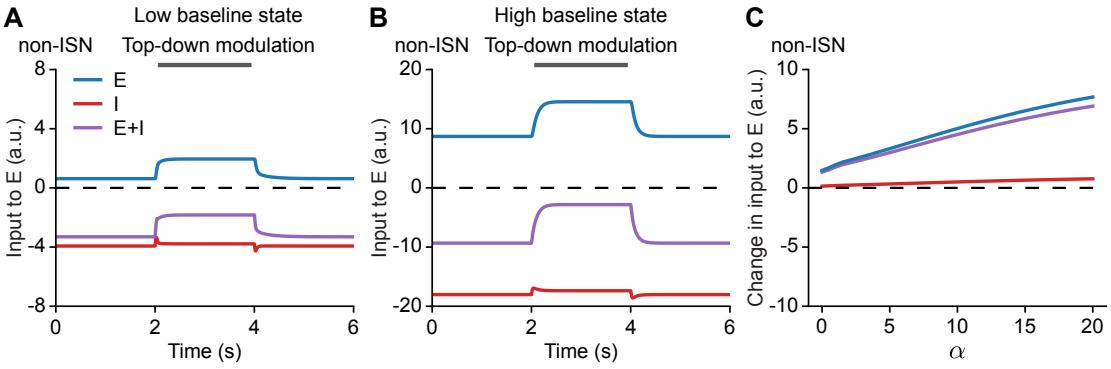


**Fig. S3.** Non-inhibition-stabilized networks (non-ISNs) do not show elevated total inhibitory inputs to the excitatory population. (**A**) Input to the E population at $\alpha$ = 0 corresponding to darkness, i.e., no sensory stimulation. Top-down modulation via VIP is applied during the interval from 2 to 4 s (gray bar). Different colors indicate different sources: input from the E population, input from the I populations, and the sum of the inputs from the E and I populations. (**B**) Same as A but at $\alpha$ = 15 corresponding to sensory stimulation. (**C**) Change in different sources of recurrent inputs to the E population measured between baseline and at steady state during top-down modulation as a function of bottom-up input $\alpha$.

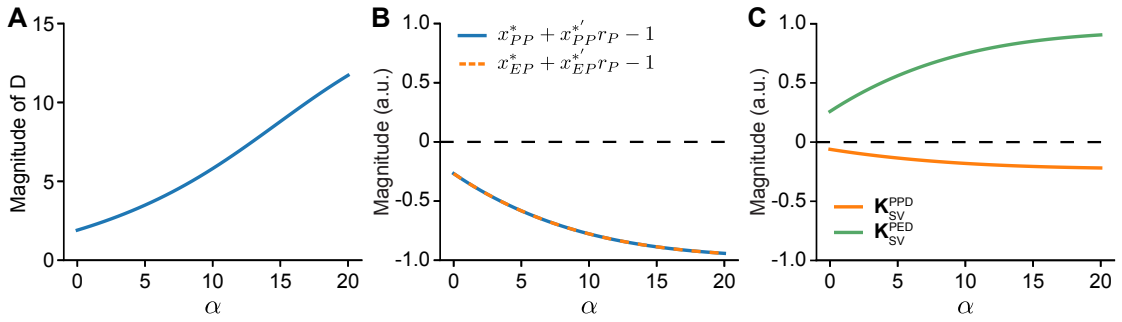


**Fig. S4.** Bottom-up input changes the magnitude of individual components of $\mathbf{R}_{SV}$. (**A**) Magnitude of $D$ in Eq. (S10) as a function of bottom-up input $\alpha$. (**B**) Magnitude of the PV-to-PV STD-related term and the PV-to-E STD-related term (see Eqs. (S19)–(S21)) as a function of bottom-up input $\alpha$. (**C**) PV-to-PV STD-dependent term $\mathbf{K}^{PPD}_{SV}$ and the PV-to-E STD-dependent term $\mathbf{K}^{PED}_{SV}$ (see Eqs. (S19)–(S21)) as a function of bottom-up input $\alpha$.

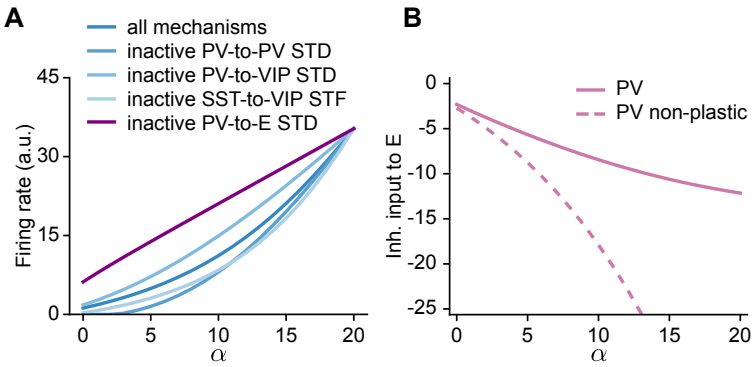


**Fig. S5.** PV-to-E STD generates an effective supralinear input-output relation. (**A**) Excitatory activity as a function of bottom-up input $\alpha$ for different network configurations marked with different colors. (**B**) Inhibitory input from PV to E as a function of bottom-up input $\alpha$. The solid line represents the input when PV-to-E STD is active, whereas the dashed line represents the input when PV-to-E STD is inactive.

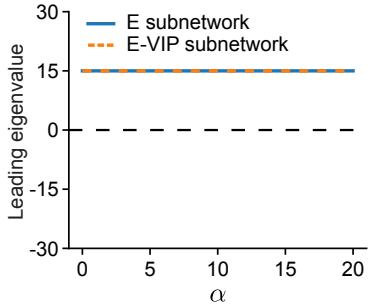


**Fig. S6.** Leading eigenvalue of the E subnetwork and the E-VIP subnetwork subnetwork as a function of bottom-up input $\alpha$.

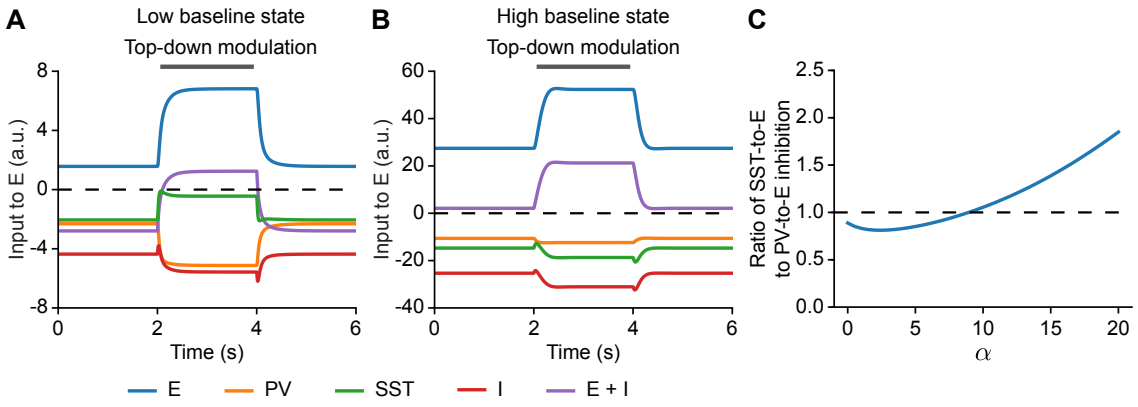


**Fig. S7.** Bottom-up inputs shift the prevalence of inhibition received by the excitatory population (E) from PV to SST. (**A**) Input to the E population at $\alpha$ = 0 corresponding to darkness, i.e., without sensory stimulation. Top-down modulation via VIP is applied during the interval from 2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

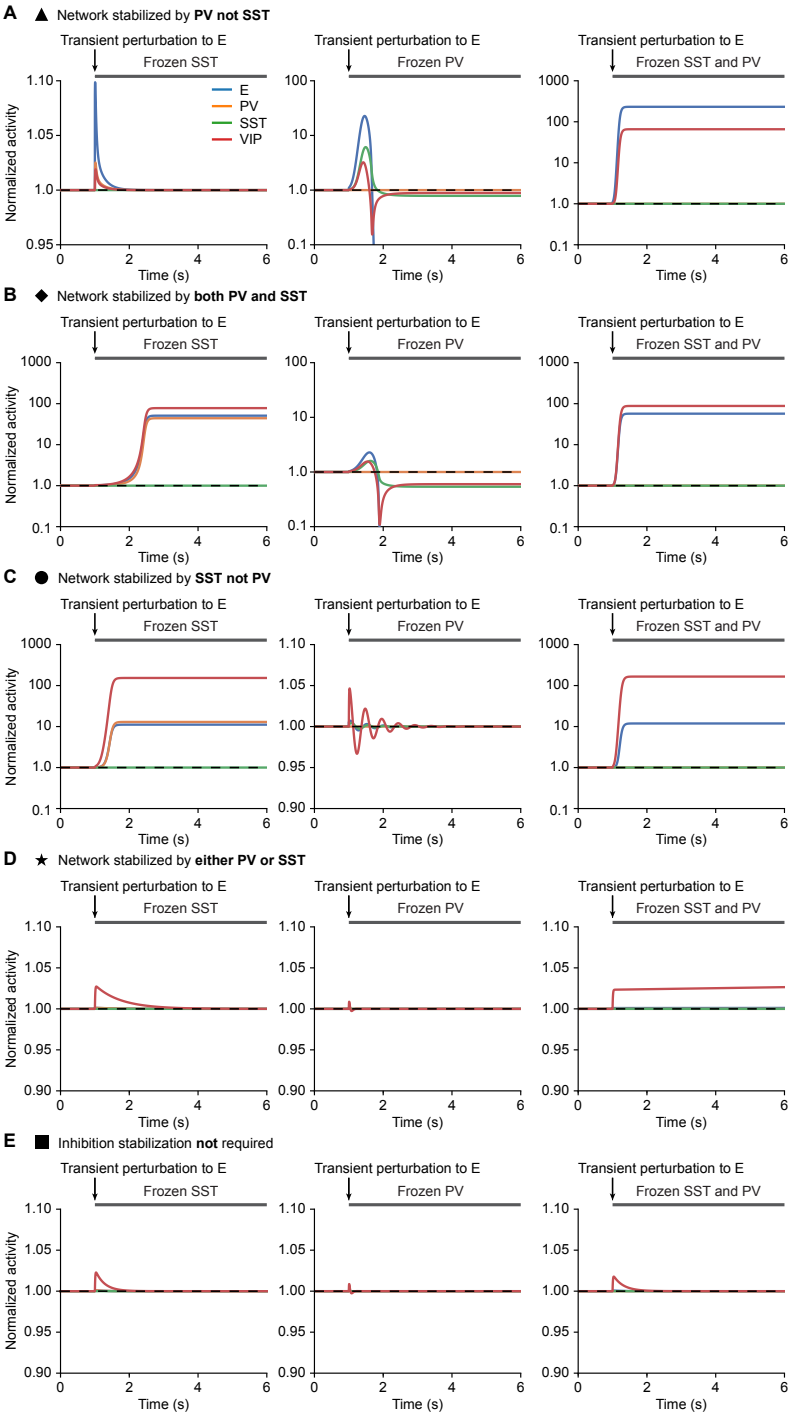

**Fig. S8.** Responses of networks also including E-to-E STD to transient perturbations to the excitatory population while freezing corresponding inhibition at different baseline states, indicated in Fig. 7D. (**A**) Normalized activity when injecting additional excitatory current into E while freezing SST, or PV, or both, respectively, for a bottom-up input of $\alpha$ = 0. A small transient excitatory perturbation to the excitatory population is introduced at the time point marked with arrows. The periods in which inhibition is frozen are marked with the gray bar. The network is inhibition stabilized and stabilized by PV but not SST. Only freezing PV or both PV and SST causes a deviation from the original fixed point. (**B**) Similar to A but for a bottom-up input $\alpha$ = 5. The network is inhibition stabilized and requires SST and PV for stabilization, as a transient perturbation leads to a deviation from the original fixed point in all freezing experiments. (**C**) Similar to A but for a bottom-up input $\alpha$ = 15. The network is inhibition stabilized and stabilized by SST but not PV. Only freezing SST or both PV and SST causes a deviation from the original fixed point. (**D**) Similar to A but for a bottom-up input $\alpha$ = 77. The network is inhibition stabilized and stabilized by either PV or SST. Only freezing both PV and SST causes a deviation from the original fixed point. (**E**) Similar to A but for a bottom-up input $\alpha$ = 90. The network is no longer inhibition stabilized. Despite frozen inhibition, a transient perturbation does not result in a deviation from the original fixed point.

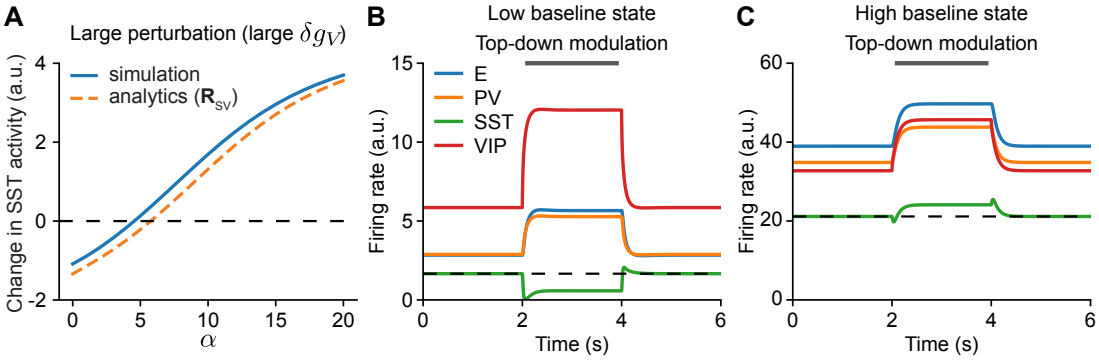


**Fig. S9.** Response reversal in networks also including E-to-SST STF. (**A**) In networks also including E-to-SST STF, analytical prediction of the change in SST population response induced by the perturbation to VIP ($\mathbf{R}_{SV}$, Eq. 10) qualitatively match with numerical simulation for a large perturbation ($\delta g_V$ = 3). (**B**) Network responses, including E-to-SST STF, to top-down modulation without any bottom-up input ($\alpha$ = 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**) Similar to B but at $\alpha$ = 15 corresponding to sensory stimulation. Simulations were performed with top-down modulation $c$ = 3 and $U^{SE}_{\max}$ = 2.

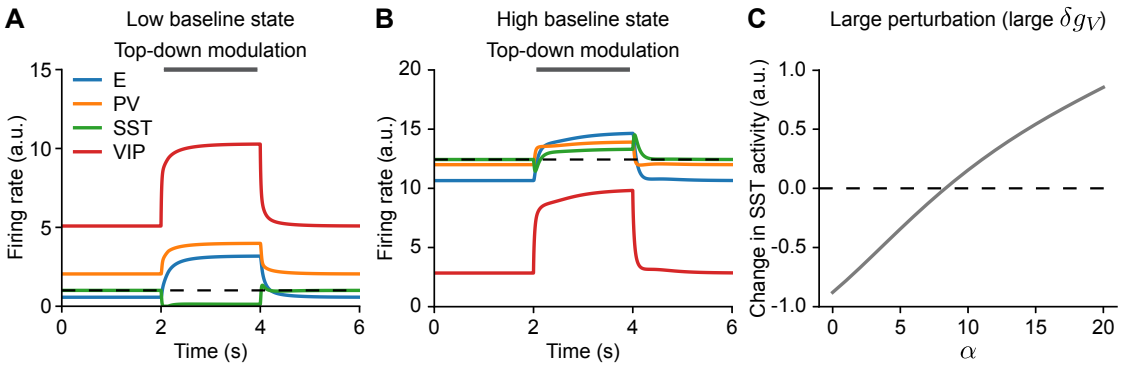


**Fig. S10.** Networks with multiple additional short-term plasticity mechanisms are able to generate response reversal. (**A**) Network responses, including all short-term mechanisms observed in (Campagnola et al., 2022), on existing connections, to top-down modulation without any bottom-up input ($\alpha$ = 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. (**B**) Same as A but at $\alpha$ = 15 corresponding to sensory stimulation. (**C**) Change in SST response induced by top-down modulation to VIP as a function of bottom-up input $\alpha$ in networks with short-term plasticity on all existing connections. Simulations were performed with top-down modulation $\delta g_V$ = 3.

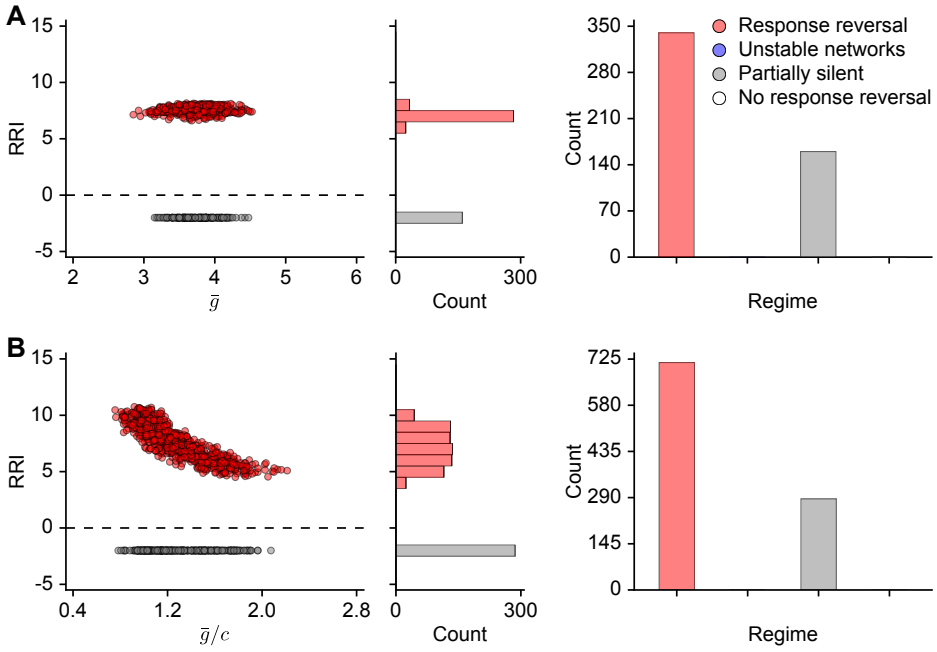


**Fig. S11.** Response reversal is observed for a wide range of background and bottom-up inputs. (**A**) Left: Response Reversal Index (RRI) as a function of the mean background input $\bar{g}$. Background inputs were randomly and independently sampled from the ranges $g_E, g_{PV}, g_{VIP} \in [3, 5]$, and $g_{SST} \in [2, 4]$, each with a step size of 0.1. The mean background input $\bar{g}$ is calculated by averaging the sampled background inputs to different populations. Simulations were performed for $n = 500$ random choices. A positive RRI indicates a response reversal of SST from negative to positive, whereas a zero RRI represents no response reversal. A negative RRI indicates partially silent networks (i.e., at least one population activity is zero) or unstable networks (i.e., network activity explodes). Middle: Histogram of RRI distribution. Right: Counts of different simulation results. (**B**) Similar to A, but for the relationship between RRI and the ratio of the mean background input $\bar{g}$ to the top-down modulatory input $c$. Additionally, $c$ was randomly drawn from 2 to 4 with a step size of 0.1. Simulations were performed for $n = 1000$ random choices. All simulations were performed with top-down modulation $c = 3$.

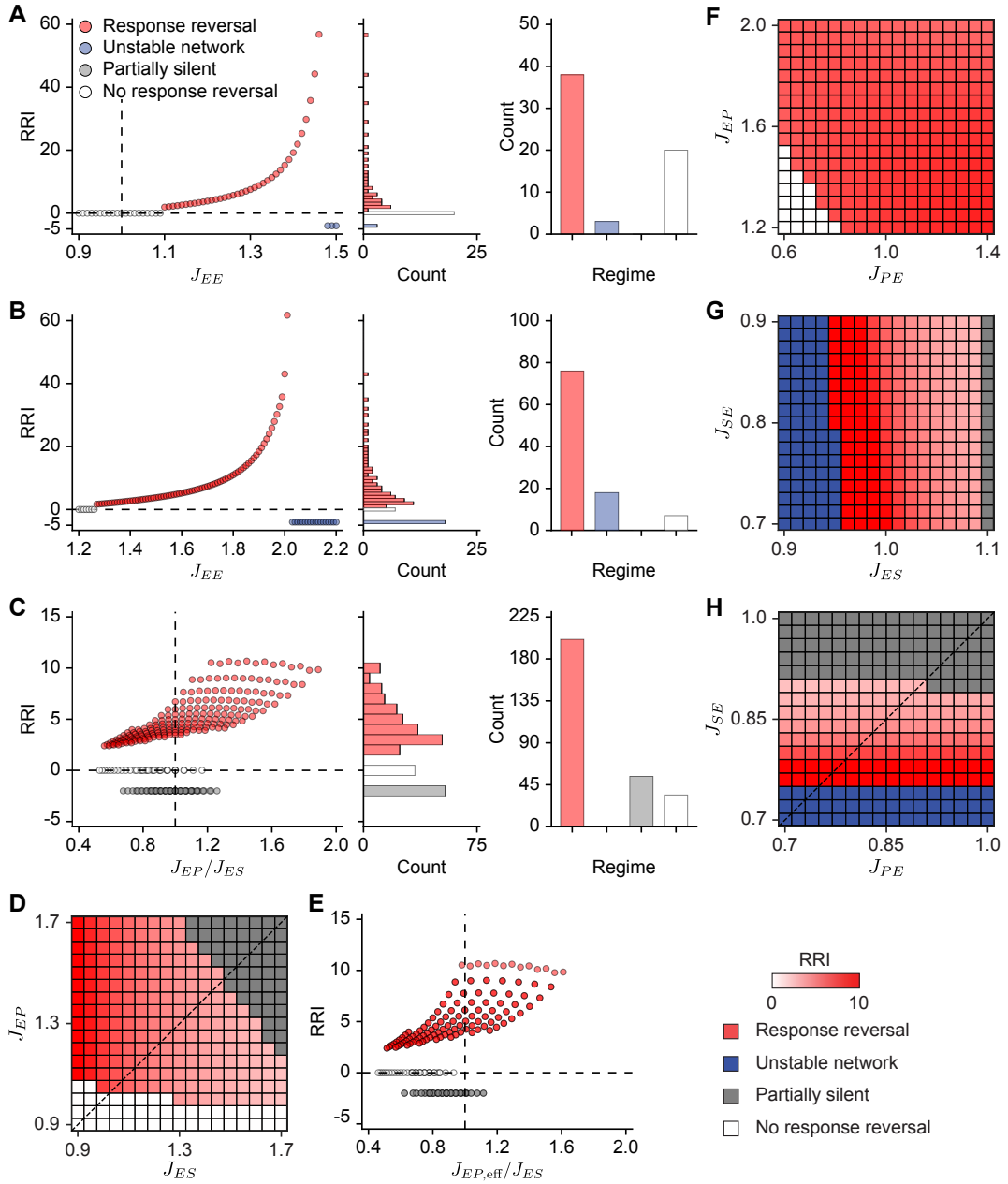


**Fig. S12.** Response reversal is observed for a wide range of network connectivity. (**A**) Left: Response Reversal Index (RRI) as a function of initial weights $J_{EE}$. A positive RRI indicates a response reversal of SST from negative to positive, whereas a zero RRI represents no response reversal. A negative RRI indicates partially silent networks (i.e., at least one population activity is zero) or unstable networks (i.e., network activity explodes). The vertical dashed line indicates $J_{EE}$ = 1. Middle: Histogram of RRI distribution. Right: Counts of different simulation results. (**B**) Similar to A but with higher inhibitory weights (see Tab. S5). (**C**) Similar to A but for the ratio of the $J_{EP}/J_{ES}$ weights. The vertical dashed line indicates a ratio of 1. The initial values of $J_{EP}$ and $J_{ES}$ are taken from 0.9 to 1.7 with a stepsize of 0.05. (**D**) RRI for different initial combinations of $J_{EP}$ and $J_{ES}$. (**E**) Similar to C, but as a function of the effective ratio between $J_{EP}$ and $J_{ES}$ by multiplying the short-term depression variable $x_{EP}$ with $J_{EP}$. Here, $x_{EP}$ is determined by its steady-state value before top-down modulation at the low bottom-up input ($\alpha$ = 0). (**F**) RRI for different initial weights of $J_{PE}$ and $J_{EP}$. (**G**) Similar to F but for different initial weights of $J_{ES}$ and $J_{SE}$. (**H**) Similar to F but for different initial weights of $J_{PE}$ and $J_{SE}$. All simulations were performed with top-down modulation $c$ = 3.

**Table S1:** Parameters for networks with iSTP.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| **Network dynamics and network connectivity** | | | |
| $\tau_E$ | 20 | ms | time constant of E rate dynamics |
| $\tau_P$ | 10 | ms | time constant of PV rate dynamics |
| $\tau_S$ | 10 | ms | time constant of SST rate dynamics |
| $\tau_V$ | 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** | | | |
| $\tau_x$ | 100 | ms | time constant of short-term depression |
| $U_d$ | 1 | a.u. | depression factor |
| $\tau_u$ | 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 |

These values are used elsewhere unless specifically stated otherwise.

**Table S2:** Parameters for networks also including E-to-E STD.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| **Network dynamics and network connectivity** | | | |
| $\tau_E$ | 20 | ms | time constant of E rate dynamics |
| $\tau_P$ | 10 | ms | time constant of PV rate dynamics |
| $\tau_S$ | 10 | ms | time constant of SST rate dynamics |
| $\tau_V$ | 10 | ms | time constant of VIP rate dynamics |
| $J_{EE}$ | 1.8 | a.u. | connection strength from E to E |
| $J_{EP}$ | 2.0 | a.u. | connection strength from PV to E |
| $J_{ES}$ | 1.0 | a.u. | connection strength from SST to E |
| $J_{PE}$ | 1.4 | 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 |
| $J_{SE}$ | 0.9 | 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** | | | |
| $\tau_x$ | 100 | ms | time constant of short-term depression |
| $\tau_x^{EE}$ | 10 | ms | time constant of E-to-E short-term depression |
| $U_d$ | 1 | a.u. | depression factor |
| $U_d^{EE}$ | 0.3 | a.u. | E-to-E depression factor |
| $\tau_u$ | 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** | | | |
| $g_E$ | 7 | a.u. | background input to E |
| $g_P$ | 7 | a.u. | background input to PV |
| $g_S$ | 5 | a.u. | background input to SST |
| $g_V$ | 7 | a.u. | background input to VIP |
| $c$ | 3 | a.u. | top-down input to VIP |

**Table S3:** Parameters for networks also including E-to-SST STF.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| **Network dynamics and network connectivity** | | | |
| $\tau_E$ | 20 | ms | time constant of E rate dynamics |
| $\tau_P$ | 10 | ms | time constant of PV rate dynamics |
| $\tau_S$ | 10 | ms | time constant of SST rate dynamics |
| $\tau_V$ | 10 | ms | time constant of VIP rate dynamics |
| $J_{EE}$ | 1.3 | a.u. | connection strength from E to E |
| $J_{EP}$ | 1.5 | a.u. | connection strength from PV to E |
| $J_{ES}$ | 0.9 | a.u. | connection strength from SST to E |
| $J_{PE}$ | 1.1 | 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 |
| $J_{SE}$ | 0.5 | 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 |
| $J_{VP}$ | 0.3 | a.u. | connection strength from PV to VIP |
| $J_{VS}$ | 0.2 | a.u. | connection strength from SST to VIP |
| **Short-term plasticity** | | | |
| $\tau_x$ | 100 | ms | time constant of short-term depression |
| $U_d$ | 1 | a.u. | depression factor |
| $\tau_u$ | 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 |
| $U_{\max}^{SE}$ | 2 | a.u. | maximum value of the E-to-SST 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 |

**Table S4:** Parameters for networks with short-term plasticity on all existing connections.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| **Network dynamics and network connectivity** | | | |
| $\tau_E$ | 20 | ms | time constant of E rate dynamics |
| $\tau_P$ | 10 | ms | time constant of PV rate dynamics |
| $\tau_S$ | 10 | ms | time constant of SST rate dynamics |
| $\tau_V$ | 10 | ms | time constant of VIP rate dynamics |
| $J_{EE}$ | 1.7 | a.u. | connection strength from E to E |
| $J_{EP}$ | 2.1 | a.u. | connection strength from PV to E |
| $J_{ES}$ | 1.5 | a.u. | connection strength from SST to E |
| $J_{PE}$ | 1.0 | a.u. | connection strength from E to PV |
| $J_{PP}$ | 1.2 | a.u. | connection strength from PV to PV |
| $J_{PS}$ | 1.3 | a.u. | connection strength from SST to PV |
| $J_{SE}$ | 0.7 | a.u. | connection strength from E to SST |
| $J_{SV}$ | 0.4 | a.u. | connection strength from VIP to SST |
| $J_{VE}$ | 0.9 | a.u. | connection strength from E to VIP |
| $J_{VP}$ | 0.5 | a.u. | connection strength from PV to VIP |
| $J_{VS}$ | 0.4 | a.u. | connection strength from SST to VIP |
| **Short-term plasticity** | | | |
| $\tau_x$ | 100 | ms | time constant of short-term depression |
| $\tau_x^{EE}$ | 10 | ms | time constant of E-to-E short-term depression |
| $U_d^{EE}$ | 0.19 | a.u. | E-to-E depression factor |
| $U_d^{EP}$ | 0.49 | a.u. | PV-to-E depression factor |
| $U_d^{ES}$ | 0.12 | a.u. | SST-to-E depression factor |
| $U_d^{PE}$ | 0.04 | a.u. | E-to-PV depression factor |
| $U_d^{PP}$ | 0.5 | a.u. | PV-to-PV depression factor |
| $U_d^{PS}$ | 0.11 | a.u. | SST-to-PV depression factor |
| $U_d^{VP}$ | 0.37 | a.u. | PV-to-VIP depression factor |
| $\tau_u$ | 400 | ms | time constant of short-term facilitation |
| $U_f^{SE}$ | 0.18 | a.u. | E-to-SST facilitation factor |
| $U_f^{VE}$ | 0.03 | a.u. | E-to-VIP facilitation factor |
| $U_f^{VS}$ | 0.28 | a.u. | SST-to-VIP facilitation factor |
| $U_f^{SV}$ | 0.05 | a.u. | VIP-to-SST facilitation factor |
| $U_{\max}$ | 3 | a.u. | maximum value of the facilitation variable |
| $U_{\max}^{SE}$ | 2 | a.u. | maximum value of the E-to-SST facilitation variable |
| **Inputs** | | | |
| $g_E$ | 5 | a.u. | background input to E |
| $g_P$ | 5 | a.u. | background input to PV |
| $g_S$ | 3 | a.u. | background input to SST |
| $g_V$ | 5 | a.u. | background input to VIP |
| $c$ | 3 | a.u. | top-down input to VIP |

**Table S5:** Parameters for sensitivity analysis of network connectivity.

| Symbol | Value | Unit | Description |
|---|---|---|---|
| **Network dynamics and network connectivity** | | | |
| $\tau_E$ | 20 | ms | time constant of E rate dynamics |
| $\tau_P$ | 10 | ms | time constant of PV rate dynamics |
| $\tau_S$ | 10 | ms | time constant of SST rate dynamics |
| $\tau_V$ | 10 | ms | time constant of VIP rate dynamics |
| $J_{EE}$ | [1.2, 2.2] | a.u. | connection strength from E to E |
| $J_{EP}$ | 1.7 | a.u. | connection strength from PV to E |
| $J_{ES}$ | 1.4 | a.u. | connection strength from SST to E |
| $J_{PE}$ | 2.2 | a.u. | connection strength from E to PV |
| $J_{PP}$ | 1.6 | a.u. | connection strength from PV to PV |
| $J_{PS}$ | 1.1 | a.u. | connection strength from SST to PV |
| $J_{SE}$ | 1.0 | a.u. | connection strength from E to SST |
| $J_{SV}$ | 0.6 | a.u. | connection strength from VIP to SST |
| $J_{VE}$ | 1.3 | 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** | | | |
| $\tau_x$ | 100 | ms | time constant of short-term depression |
| $U_d$ | 1 | a.u. | depression factor |
| $\tau_u$ | 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 |