# Neural circuit mechanisms of cognitive flexibility

by

Guangyu Robert Yang

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Neural Science

New York University

January 2018

---

Xiao-Jing Wang

# Dedication

I would like to dedicate this dissertation to my wife, Shuying Luo.

# Acknowledgments

During the winter vacation of my sophomore year at Peking University, I toured around universities in the US, visiting some well-known researchers in computational neuroscience, a subject in which I became interested. When I arrived at New Haven, I went straight to Xiao-Jing Wang's office and knocked on his door. Even though I came totally out of the blue, Xiao-Jing patiently listened to my naïve opinions about neuroscience and theories. He suggested that I should take a dynamical system course and encouraged me to apply to the summer school he is co-organizing in Shanghai. Everything else followed from there. I am forever grateful to my advisor Xiao-Jing for leading me into this field of wonders, encouraging me to be fearless in pursuing the truth, and a constant emphasis on connecting models with experiments.

John Murray was a senior graduate student when I first joined the lab. I collaborated closely with him on my first project in the lab. Years later, after I have been in many collaborations, I realized that what John did was way beyond a second-author-collaborator. He was indeed a wonderful mentor and has always been a great friend.

Francis Song was my most important collaborator in the past five years. His enthusiasm for bridging neuroscience and neural networks coincided with mine. I was extremely fortunate to overlap with him for the most of my time in the Wang Lab.

Members of the Wang Lab have created an incredibly dynamic and stimulating in-

tellectual environment. The many lively discussions I have had in lab meetings, after seminars, and during drinks showed the collective passion to science that sustained me throughout the graduate school and continues to inspire me. In particular, I would like to thank Rishidev Chaudhuri, Jorge Jaramillo, Owen Marschall, and Jian Li. I would also like to thank members of the Theory Suite at Meyer 750, especially Emin Orhan for many thoughtful discussions on neural networks.

I am grateful to the many collaborators I was lucky to have. In particular, I would like to thank Bill Newsome, Pavel Osten, Yongsoo Kim, Luis Carlos Garcia del Molino, Jorge Mejias, Madhura Joglekar, Daeyeol Lee, and Hyojung Seo.

I would also like to thank my committee members Paul Glimcher, John Rinzel, Adam Carter, and Brent Doiron. They have kept me on track and provided thoughtful feedbacks throughout the past years.

I would like to thank my classmates at both Yale and NYU who have formed another pillar of my intellectual and social life.

I would like to give special thanks to Shenglong Wang–whom I have never met in person–for providing timely, professional, much-needed support for the NYU computing cluster.

Finally, I am forever grateful to my parents. Ever since I was a toddler, they have encouraged me to explore and to understand the unknowns. They gently cultured my interests in science, while always respecting and supporting my choices.

# Abstract

Cognitive flexibility allows us to dynamically and adaptively respond to the world depending on the need or context. Along with working memory and inhibitory control, cognitive flexibility is considered one of the core executive functions of the brain. There have been many models for cognitive flexibility, however, they are both biologically simplified and computationally limited. This dissertation presents models can overcome some of these difficulties by including cell-type specificity into neural circuit models of cognitive flexibility and by training recurrent neural networks to flexibly perform challenging cognitive tasks.

The brain consists of many different types of cells, with cell-type specific connectivity between them. We hypothesize that two major types of interneurons play critical roles in one form of cognitive flexibility: pathway-specific gating, which refers to the capability to dynamically route or gate information from different pathways in the brain. We show that biologically-constrained circuit models equipped with several types of interneurons and multi-compartmental pyramidal neurons can support pathway-specific gating.

In combination of experimental data analysis, we further study the functional properties of multi-cell-type circuit models. By analyzing the density of several types of interneurons throughout mouse cortex, we propose a hierarchy of cortical areas based on the ratio of dendrite- and soma-targeting interneurons. Dendrite-targeting interneurons

are relatively enriched in higher-order cortical areas like the prefrontal cortex. We show that such an enrichment can have counter-intuitive functional impact on the circuit that can only be understood with circuit modeling. We develop a theoretical framework explaining these counter-intuitive phenomena. The framework can also be used to explain seemingly contradictory results found in mouse V1.

Most computational models of cognitive functions are computationally limited. However, recurrent neural networks can be trained to perform many complicated tasks and therefore serve as candidate models for cognitive functions. We study a range of recurrent networks, each trained to perform a classical cognitive task. We show that our networks recapitulate important behavioral and neurophysiological features of animals performing similar tasks.

Previous computational models of cognitive flexibility tend to focus on one task at a time. Humans and animals can also flexibly switch between a large number of tasks. We study how neural circuits can support and switch between many different cognitive tasks by training and analyzing a recurrent neural network that performs 20 cognitive tasks. By dissecting the network, we show that the network consists of clusters of neurons, each supporting a cognitive process underlying a group of cognitive tasks. The network can flexibly perform a task by assembling the correct combination of clusters.

Finally, we present preliminary results where we train multi-cell-type recurrent neural networks to perform a cognitive task. We analyze the neural mechanism of the network and show that it can learn to utilize dendrites of pyramidal neurons to perform pathway-specific gating.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Overview

Imagine yourself sitting in a noisy café trying to read. To focus on the book at hand, you need to ignore the surrounding chatter and clattering of cups, with your brain filtering out the irrelevant stimuli coming through your ears and gating in the relevant ones in your visionwords on a page. Now you hear a vague mention of your name. Your brain can disengage you from the visual stimuli and start to pay attention to the auditory stimuli. The external stimuli are the same, yet your brain can flexibly route different streams of information to the central processing system. Or consider yourself visiting the UK from the US. Now when you are crossing the road, instead of looking to your right first, you need to look to your left. You can choose to respond in a completely different way to the same stimulus depending on the context. Or imagine yourself in a heated debate with a colleague, who asked you to "think for one moment from my perspective." You can shift mentally from your own standpoint to, at least an approximation of, someone else's point of view.

These behaviors are all supported by our cognitive flexibility, which involves "the selective use of knowledge to adaptively fit the needs of understanding and decision mak-

ing in a particular situation" (Spiro et al., 1988). Cognitive flexibility, together with inhibitory control and working memory, are considered the three core executive functions of the brain (Diamond, 2013). Cognitive flexibility is also called, or is closely related to, task/set shifting and mental flexibility. It became a formal subject of scientific studies in the 1960s (Scott, 1962), thanks to the cognitive revolution that started in the late 1950s (Miller, 1956). However, flexibility of the mind has been an informal folk concept long before it was a formal subject in psychology, not to mention neuroscience. For example, flexibility was the focus of one subtest in the California Psychological Inventory (Gough, 1956), one of the earliest personality tests.

One of the most important family of tasks used to assess and study cognitive flexibility are the card sorting tasks. In these tasks, the subjects are presented with a set of cards. There are typically a number of colored items on each card, and subjects are required to sort the cards based on one of the dimensions (item type, number, color, etc.). The subjects also need to flexibly switch from one dimension to another based on the experimenters' instructions or feedback. Card sorting tasks were first introduced by the German neurologist Kurt Goldstein and the psychologist Adhémar Gelb around World War I to study the behavior of brain injured patients. They observed that some patients have difficulties sorting cards based on a single dimension: "it is impossible or difficult to guide their attention in a particular direction" (Gelb and Goldstein, 1925; Weigl, 1927). One widely-studied variant of the card sorting tasks is the Wisconsin Card Sorting Task (WCST), which was developed by Esta Berg in 1948 at the University of Wisconsin (Berg, 1948) with the clear goal of studying "flexibility in thinking". In this task, the subjects are presented with a set of stimulus cards, items on which differ in color, shape, and number. The subjects are asked to match another pile of cards to the stimulus cards. The

correct matching is based on a single dimension (color, shape, or number). The subjects are, however, not informed about the rule, and have to figure it out through trial-and-error. Importantly, the matching rule silently changes after a number of trials, and subjects need to switch their strategies accordingly. Human subjects with lesions in dorsolateral frontal cortex have great difficulties performing this task, while patients with lesions in temporal, parietal, or orbitofrontal cortex can perform the tasks as well as control subjects (Milner, 1963). In comparison to working memory, cognitive flexibility matures relatively late. Children gain adult-level performance at WCST around age 10 (Welsh, Pennington and Groisser, 1991).

Another related task that involves cognitive flexibility is the Stroop task (Stroop, 1935). In this task, subjects are presented with words that represent various colors, for example, blue, green, and red. At the same time, these words are written in colored ink. Importantly, the ink-color could be different from the meaning of each word. Subjects are asked to report either the color of each word or the word itself. In this task, the subjects need to flexibility control the information flow from the word-reading and the color-naming pathways. A critical difference between the Stroop task and the card sorting tasks is that in the Stroop task, the word-reading pathway is stronger or prepotent in comparison to the color-naming pathway. Correctly naming colors requires inhibition of the word-reading pathway. Therefore the Stroop task is more commonly used to study inhibitory control. If card-sorting tasks and the Stroop task appear highly related to selective attention, there is another family of tasks used to assess cognitive flexibility that is more closely linked to creativity. A task in the fluency task family requires the subjects to come up with as many answers as possible that fit a certain pattern. For example, in the Thurstone word fluency task (Thurstone, 1938), subjects are asked to write as many words as possi-

ble that begin with a certain letter, "A" for example. In a verbal fluency task, the subjects may be asked to come up with words belonging to the "tools" category or the "animals" category (Baldo et al., 2001). Together, these tasks allow experimenters to gain substantial insights into the effects of brain lesions, psychiatric disorders, and development on cognitive flexibility. These tasks also formed the basis for later tasks that are used to study cognitive flexibility in non-human animals and in computational models.

In comparison to study in humans, the study of cognitive flexibility in non-human animals started relatively recently. Both neural recording and lesioning studies in monkeys confirmed the role of prefrontal and anterior cingulate cortex in cognitive flexibility Johnston et al. (2007); Mansouri, Matsumoto and Tanaka (2006); Isoda and Hikosaka (2007); Rushworth et al. (2003); Mante et al. (2013); Siegel, Buschman and Miller (2015). Neural recordings of monkeys performing a task analog to the WCST showed that prefrontal neural activities are modulated by the current task rule (Mansouri, Matsumoto and Tanaka, 2006). Neural correlates of the task rule in a task switching paradigm have also been reported in the posterior parietal cortex. Stoet and Snyder (2004) trained monkeys to classify colored bars based on either their colors or their orientations. The rule, color or orientation, is provided explicitly at the beginning of each trial. Neural correlates of the current task rule are found in the lateral bank of the intraparietal sulcus (area LIP) in the posterior parietal cortex. Whether or not cognitive flexibility is implemented by modulating representations in sensory areas is less clear. Using a task similar to the one used by Stoet and Snyder (2004), Mirabella et al. (2007) reported that neurons in the visual area V4 are also modulated by the current task rule. In contrast, Sasaki and Uka (2009) trained monkeys to discriminate directions or depths of random dot stereogram stimuli depending on the rule cue. They found little rule selectivity in the visual area MT. In the

past few years, experimenters have started training rodents to perform versions of these tasks (Rodgers and DeWeese, 2014). Rodgers and DeWeese (2014) trained rats to switch between a localization task and a pitch discrimination task. The stimuli contain both spatial and pitch information, and the rats need to focus on one of the two dimensions in each block of trials. Selectivity of the current rule is found in both the prefrontal cortex and auditory cortex. After training mice to focus on either the visual or auditory modality based on a rule cue, Wimmer et al. (2015) found that the prefrontal cortex controls the sensory thalamus to perform the modality selection.

In a study particularly relevant to this thesis, Mante et al. (2013) recorded from the prefrontal cortex of monkeys distinguishing either the direction or the color of a colored random-dot stimulus (Mante et al., 2013). A rule cue is presented at the beginning of each trial informing the monkeys which dimension they should focus on. They reported that the prefrontal cortex can selectively integrate information from different input pathways (motion or color) based on a dynamical system solution.

The dominant conceptual framework for understanding the neural mechanism of cognitive flexibility is the theory of cognitive control summarized by Miller and Cohen (2001). According to this theory, the prefrontal cortex maintains a firing pattern that represents the rule. These prefrontal activities guide the information flow in the brain by providing biasing signals to signal pathways. This mechanism is used in many computational models that involve cognitive flexibility. It can be well demonstrated with the model developed by Cohen, Dunbar and McClelland (1990). Using the then newly developed framework of parallel distributed processing, Cohen, Dunbar and McClelland (1990) trained a simple feedforward network to perform the Stroop task. Their network is able to capture essential features of the Stroop effect, including long reaction times to

perform the color-naming task. In this model, there are two different processing pathways from the sensory inputs to the response, one for word reading and another for color naming. Two input units, representing the word reading and color naming rules respectively, excite corresponding sensory processing units for each pathway. These rule inputs are used to model the effect of top-down control inputs that provide biases for the processing pathways. Later computational models typically follow this tradition and explain flexible switching between different tasks by having selective rule inputs to one set of neurons performing a task (Ardid and Wang, 2013). It has also been hypothesized that the control inputs can influence the processing of sensory inputs not only by increasing or decreasing activation of neurons, but also by dynamically adjusting synaptic weights for different pathways (Olshausen, Anderson and Van Essen, 1993). This kind of dynamic adjustment of synaptic weights can be potentially achieved through a coordinated change of excitatory and inhibitory inputs to neurons, leading to a change in the neuronal gain (Vogels and Abbott, 2009). A notable exception of the guided flow model (Miller and Cohen, 2001) is the computational model presented in Mante et al. (2013). The authors trained a recurrent neural network model to perform the motion/color discrimination task described above. By carefully analyzing the mechanism of the network from a dynamical system perspective, the authors found that the network performs the task not by modulating processing of the sensory inputs. Instead, the network can selectively integrate information from one pathway but not the other depending on the rule cue.

Although much has been done, the neural mechanisms of cognitive flexibility are still far from being understood. The current conceptual understanding stays almost exclusively at the level of simplified neural circuitry. Both experiments and models are typi-

cally concerned about "neurons" in a generic way. There is little differentiation between different types of neurons or different layers within an area. There is a great diversity of cell types in the brain (Cajal, 1911). There are at least 60 identified cell types in mammalian retina (Helmstaedter et al., 2013). In the cortex, where the full diversity of cell types is far from being mapped, more than 50 types of inhibitory neurons have been identified (Markram et al., 2004). Morphologically and physiologically distinct types of cells can have markedly different functional roles even if they reside in the same cortical area (Kepecs and Fishell, 2014). Tools that allow scientists to record and manipulate specific cell types in mammals have only become available in the last 10 years (Sohal et al., 2009; Taniguchi et al., 2011). However, most experiments studying the single-neuron level neural mechanisms of cognitive flexibility have been conducted in monkeys, where cell-type specific tools are not yet available. As a result, past conceptual and computational models for cognitive flexibility (or any other cognitive functions) almost never involve different cell types. In a rare exception, Wang et al. (2004) proposed a computational model of working memory where three major types of interneurons play different roles.

Although there are many cell types in the brain, cortical interneurons, as a first-order approximation, can be classified into three largely non-overlapping groups: cells expressing parvalbumin (PV), somatostatin (SOM or SST), and the serotonin receptor 5HT3a (5HT3aR) (Rudy et al., 2011). SST-positive neurons account for roughly 30% of all interneurons in the cortex. They primarily target dendrites of pyramidal neurons (Markram et al., 2004; Freund and Buzsáki, 1996) and other interneurons. They largely avoid targeting each other through chemical synapses (Jiang et al., 2015), yet nearby ones are frequently connected through gap junctions (Urban-Ciecko and Barth, 2016). Another major type of inhibitory neurons, the PV-positive inhibitory neurons mainly target so-

matic areas of pyramidal neurons (Markram et al., 2004; Freund and Buzsáki, 1996). They also target each other and other types of interneurons (Jiang et al., 2015). In addition to their differences in local connectivities, PV+ and SST+ neurons receive, at least quantitatively, different long-range connections (Wall et al., 2016). These differences strongly suggest distinct functional roles for PV+ and SST+ neurons. The differential roles of PV+ and SST+ neurons in the neural mechanism of cognitive flexibility, if they exist, are almost completely unknown. Here I will attempt to address this issue by hypothesizing that dendrite-targeting inhibitory neurons have a critical role in cognitive flexibility. I will use computational models to demonstrate the plausibility of this hypothesis.

Another issue with previous conceptual and computational models of cognitive flexibility is that they mostly focus on relatively simple tasks. Animal studies of cognitive flexibility frequently involve switching between two sensory modalities or features for a limited set of stimuli. Although these tasks can still be difficult for animals to learn, humans can apparently switch between much more complicated behaviors. Moreover, the tasks used in animal studies can often be framed as studying selective attention, instead of cognitive flexibility. In other words, these tasks tend to focus on a rather specific aspect of cognitive flexibility: the capability to flexibly attend to different sensory modalities or features. Cognitive flexibility can take many forms beyond selective attention. At the same time, it is difficult to build, in a bottom-up fashion, computational models that are capable of complex behaviors. In the bottom-up modeling approach, the overall structure of the model is specified by the modeler. Usually, a small number of free parameters are hand tuned or fitted to explain experimental findings. The major strength of this approach is the interpretability of the model, because there is a better understanding of the functional roles of model components and parameters. However, trying to hand-build

a model for complex behaviors is somewhat like Ptolemy's attempt to explain elliptical orbits with epicycles (circles upon circles). Not surprisingly, the attempts to build successful object recognition systems "by hand" have mostly failed. In addition, traditional bottom-up models frequently have difficulties explaining the variability in neuronal selectivities observed in the brain. In this thesis, I will attempt to address this problem by studying a neural network trained with machine learning techniques in order to perform many cognitive tasks. Such a network provides us with an opportunity to study cognitive flexibility in a broader context.

## 1.1 Chapter Overview

Here I will overview the contents of each chapter. Most chapters are based on published papers so they will be self contained.

While reading a book in a noisy café, how does your brain "gate in" visual information while filtering out auditory stimuli? In **Chapter 2**, we propose a mechanism for such flexible routing of information flow in a complex brain network (pathway-specific gating), tested using a network model of pyramidal neurons and three classes of interneurons with connection probabilities constrained by data. We find that if inputs from different pathways cluster on pyramidal neuron dendrites, a pathway can be gated-on by a disinhibitory circuit motif. The branch-specific disinhibition can be achieved despite dense interneuronal connectivity, even with random connections. Moreover, clustering of input pathways on dendrites can emerge naturally through synaptic plasticity regulated by dendritic inhibition. This gating mechanism in a neural circuit is further demonstrated by performing a context-dependent decision-making task. The model suggests that cog-

nitive flexibility engages top-down signaling of behavioral rule or context which targets specific classes of inhibitory neurons.

In Chapter 2 we propose that dendrite-targeting interneurons are critical for pathway-specific gating. Higher-order cortical areas receive converging inputs from multiple pathways and have a stronger need to perform pathway-specific gating. Therefore our proposal suggests that there should be more dendrite-targeting interneurons in higher-order cortical areas. In **Chapter 3**, we present data analyses that confirm this prediction. We analyze the first quantitative brain-wide map of interneuron cell densities and show that higher-order cortical areas have a higher density of SST neurons and a lower density of PV neurons. We then model the functional implication of this density variation using both a simplified rate-neuron circuit model and a more realistic spiking-neuron circuit model. We found that varying cell densities can have a counterintuitive impact on the neural circuit responses to external inputs. For example, increasing the density of one type of neurons can affect how other types of interneurons interact. These findings emphasize the necessity of circuit modeling in understanding the functions of interneuronal circuit, especially when several types of interneurons are present.

In **Chapter 4**, we further study the counterintuitive nature of interneuronal circuitry and use that to explain puzzling experimental results that seem to contradict the disinhibitory circuit motif. For example, the response of SST cells in mouse V1 to top-down behavioral modulation can change its sign when the visual input changes, a phenomenon that we call response reversal. We developed a theoretical framework to explain these seemingly contradictory effects as emerging phenomena in circuits with two key features: interactions between multiple neural populations and a nonlinear neuronal input-output relationship. Furthermore, we built a cortical circuit model which reproduces the

counterintuitive dynamics observed in mouse V1. Our analytical calculations pinpoint connection properties critical to response reversal, and predict additional novel types of complex dynamics that could be tested in future experiments.

Although results presented in Chapters 3 and 4 may not appear immediately relevant to cognitive flexibility, they provide a more rigorous understanding of the functions and dynamics of soma- and dendrite-targeting interneurons. This understanding will become critical if we are to fully understand interneurons' roles in cognitive flexibility and other cortical functions.

In Chapters 5-7, we will use the approach of trained network analysis to study cognitive flexibility. In Chapter 5 we will prepare the readers with the method to be used and apply the method to classical cognitive tasks and results. The ability to simultaneously record from large numbers of neurons in behaving animals has ushered in a new era for the study of the neural circuit mechanisms underlying cognitive functions. One promising approach to uncovering the dynamical and computational principles governing population responses is to analyze model recurrent neural networks (RNNs) that have been optimized to perform the same tasks as behaving animals. Because the optimization of network parameters specifies the desired output but not the manner in which to achieve this output, "trained" networks serve as a source of mechanistic hypotheses and a testing ground for data analyses that link neural computation to behavior. Complete access to the activity and connectivity of the circuit, and the ability to manipulate them arbitrarily, make trained networks a convenient proxy for biological circuits and a valuable platform for theoretical investigation. However, existing RNNs lack basic biological features such as the distinction between excitatory and inhibitory units (Dale's principle), which are essential if RNNs are to provide insights into the operation of biological circuits. Moreover,

trained networks can achieve the same behavioral performance but differ substantially in their structure and dynamics, highlighting the need for a simple and flexible framework for the exploratory training of RNNs. In **Chapter 5**, we describe a framework for gradient descent-based training of excitatory-inhibitory RNNs that can incorporate a variety of biological knowledge. We validate this framework by applying it to well-known experimental paradigms such as perceptual decision-making, context-dependent integration, multisensory integration, parametric working memory, and motor sequence generation. Our results demonstrate the wide range of neural activity patterns and behavior that can be captured in our trained network models, and suggest a unified setting in which diverse cognitive computations and mechanisms can be studied.

A neural system has the ability to flexibly perform many tasks, the underlying mechanism cannot be elucidated in traditional experimental and modeling studies designed for one task at a time. In **Chapter 6**, we trained a single network model to perform 20 cognitive tasks that may involve working memory, decision-making, categorization and inhibitory control. We found that after training, recurrent units developed into clusters that are functionally specialized to various cognitive processes. Moreover, we introduced a measure to quantify relationships between single-unit neural representations of tasks, and report five distinct types of such relationship that can be tested experimentally. Surprisingly, our network developed compositionality of task representations, a critical feature for cognitive flexibility, whereby one task can be performed by recombining instructions for other tasks. Finally, we demonstrate how the network could learn multiple tasks sequentially. This work provides a computational platform to investigate neural representations of many cognitive tasks.

In **Chapter 7**, we present preliminary results where we combine the approaches from

previous chapters. We train a recurrent neural network equipped with multi-compartmental pyramidal neurons and multiple types of interneurons (PV, SST, and VIP neurons). We trained the network to perform a generalized version of the pathway-specific gating Mante task (Mante et al., 2013). The network has to focus on one of the five pathways provided while ignoring information from others. We investigate the solution our network developed and test whether it learned to utilize dendrite-targeting interneurons as predicted in Chapter 2. These results show how we can combine canonical circuit motives discovered in the cortex with the approach of training neural networks. This combination allows us build circuit models that are both grounded in anatomy and capable of sophisticated cognitive tasks.

Chapter 2 is based on Yang, Murray and Wang (2016). Chapter 3 is based on the data analysis and modeling results in Kim et al. (2017). Chapter 4 is based on del Molino et al. 2017, under review, in which LCG del Molino performed the research, and I participated in the research design and manuscript writing. Chapter 5 is based on Song, Yang and Wang (2016). I participated in the research design. HF Song and I performed the research (HF Song performed the majority of it). Chapter 6 is based on Yang et al. 2017, under review. Chapter 7 is based on the work of JA Li. I participated in the research design. JA Li performed the research as a summer internship student.

# Chapter 2

# Gating through dendritic disinhibition

## 2.1 Introduction

Distinct classes of inhibitory interneurons form cell-type specific connections among themselves and with pyramidal neurons in the cortex (Markram et al., 2004; Jiang et al., 2015). Interneurons expressing parvalbumin (PV) specifically target the perisomatic area of pyramidal neurons. Interneurons expressing somatostatin (SOM) specifically target thin basal and apical tuft dendrites of pyramidal neurons (Urban-Ciecko and Barth, 2016; Chiu et al., 2013). Interneurons expressing vasoactive intestinal peptide (VIP) avoid pyramidal neurons and specifically target SOM neurons (Pfeffer et al., 2013). Long-range connections from cortical (Lee et al., 2013; Zhang et al., 2014) or subcortical (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014) areas can activate VIP neurons, which in turn suppress SOM neurons, and disinhibit pyramidal dendrites. Such dendritic disinhibitory circuit is proposed to gate excitatory inputs targeting pyramidal dendrites (Wang et al., 2004; Kepecs and Fishell, 2014; Sridharan and Knudsen, 2015) (**Fig. 2.1a**).

Insofar as any cortical area receives inputs from tens of other areas and projects to

many other areas, information flow across the complex cortical circuit needs to be flexibly gated (or routed) according to behavioral demands. Broadly speaking, there are three types of gating in terms of specificity. First, all inputs into a cortical area may be uniformly modulated up or down. Recent research in mice has demonstrated that such gating involves the disinhibitory motif mediated by VIP and SOM interneurons (Pinto and Dan, 2015; Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014; Kvitsiani et al., 2013; Pi et al., 2013; Gentet et al., 2012; Lee et al., 2013). These studies generally found that VIP neurons are activated, and SOM neurons are inactivated, in response to changes in the animals' behavioral states, such as when mice receive reinforcement (Pi et al., 2013), or start active whisking (Gentet et al., 2012; Lee et al., 2013) or running (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014). The reported state change-related activity responses can be remarkably homogeneous across the local population of the same class of interneurons (Pinto and Dan, 2015; Kvitsiani et al., 2013).

Second, gating may involve selective information about a particular stimulus attribute or spatial location (for instance, in visual search or selective attention (Zhang et al., 2014)). Whether SOM or VIP neurons are endowed with the required selectivity remains insufficiently known. In sensory cortex, SOM neurons exhibit greater selectivity to stimulus features (such as orientation of a visual stimulus) than PV neurons(Ma et al., 2010). Furthermore, in motor cortex, SOM neurons have been shown to be highly heterogeneous and remarkably selective for forward versus backward movements (Adler and Gan, 2015).

Third, for a given task, neurons in a cortical area may need to "gate in" inputs from one of the afferent pathways, and "gate out" other afferent pathways (Akam and Kullmann, 2010; Vogels and Abbott, 2009), which we call "pathway-specific gating". For instance, imagine yourself sitting in a noisy cafe and trying to focus on your book. Your associa-

tional language areas receive converging inputs from both auditory and visual pathways. Opening the gate for the visual pathway while closing the gate for the auditory pathway allows you to focus on reading (**Fig. 2.1b**). In the classic Stroop task, the subject is shown a colored word, and is asked to either name the color or read the word. One possible solution to this task is for a decision-making area to locally open its gate for the deliberate pathway (color-naming) while closing its gate for the more automatic pathway (word-reading).

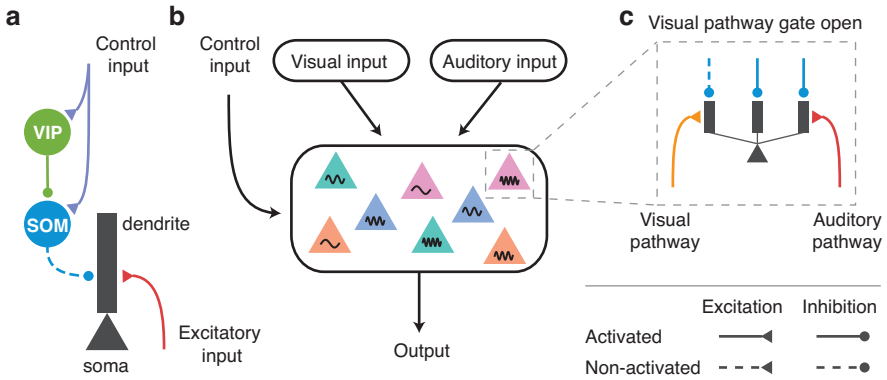


Figure 2.1: Dendritic disinhibitory circuit as a mechanism for pathway-specific gating. (**a**) Subcellular microcircuit motif for gating through dendritic disinhibition. Dendrites of pyramidal neurons are inhibited by SOM interneurons, which are themselves inhibited by VIP interneurons. A control input (representing a context or a task rule) targeting VIP interneurons (and potentially SOM neurons) can thereby disinhibit pyramidal neuron dendrites, opening the gate for excitatory inputs targeting these dendrites. (**b**) Circuit configuration for pathway-specific gating. Pyramidal neurons receive converging inputs from multiple pathways, e.g. visual and auditory. Single neurons in these areas are selective to multiple stimulus features, indicated here by color and frequency. The processing of each pathway is regulated by the control input. (**c**) Inputs from different pathways target distinct subsets of dendrites of these pyramidal neurons. A pathway can be gated-on by specifically disinhibiting the dendrites that it targets, corresponding to an alignment between excitation and disinhibition. Disinhibition is represented by dashed lines.)

Using computational models, we propose that the dendritic disinhibitory circuit can instantiate pathway-specific gating. Each of the many branches of a pyramidal dendrite processes its inputs quasi-independently (Poirazi, Brannon and Mel, 2003) and nonlinearly (Major, Larkum and Schiller, 2013). Feedforward and feedback pathways target different regions (e.g. basal or apical tuft) of dendritic trees of pyramidal neurons (Petreanu

et al., 2009). We hypothesize that excitatory inputs from different pathways can cluster onto parts of dendrites of pyramidal neurons, which we term “branch-specific," even though inputs from a particular pathway may target multiple branches. This hypothesis is supported by mounting evidence for synaptic clustering on dendritic branches (Druckmann et al., 2014; Yang et al., 2014; Kastellakis et al., 2015). A pathway can presumably be “gated-on” by specifically disinhibiting the branches targeted by this pathway (**Fig. 2.1c**), i.e. by a disinhibition pattern aligned with the excitation. This branch-specific disinhibition is motivated by findings showing that synaptic inhibition from SOM neurons can act very locally on dendrites, even controlling individual excitatory synapse by targeting the spine (Chiu et al., 2013) or the pre-synaptic terminal (Urban-Ciecko, Fanselow and Barth, 2015).In this work, we developed a network model with thousands of pyramidal neurons and hundreds of interneurons for each (VIP, SOM, and PV) type, and show that pathway-specific gating can be accomplished by the disinhibitory motif, even though the connectivity from SOM neurons to pyramidal neurons is dense: each SOM neuron on average targets more than 60% of neighboring pyramidal neurons ($< 200$ $\mu$m) (Fino and Yuste, 2011).

We first characterized how branch-specific disinhibition can efficiently gate excitatory inputs onto pyramidal dendrites. We then investigated the plausibility of branch-specific disinhibition in a dendritic disinhibitory circuit model constrained by experimentally measured single-neuron physiology and circuit connectivity. We found that although SOM-pyramidal connectivity is dense at the level of neurons, at the level of dendrites it is sufficiently sparse to support branch-specific disinhibition, and therefore pathway-specific gating, given that SOM neurons can be selectively controlled. We then showed control inputs targeting both VIP and SOM neurons can selectively suppress

SOM neurons as needed. Notably we drew these conclusions under some "worst-case" assumptions to our model such as random interneuronal connectivity. Using a calcium-based synaptic plasticity model, we found that disinhibitory regulation of plasticity can give rise to an appropriate alignment of excitation and disinhibition, which is required for pathway-specific gating in our scheme. Finally, we demonstrated the functionality of this mechanism in a circuit model performing an example context-dependent decision-making task (Mante et al., 2013).

Our results suggest that, in addition to the proposal that SOM neurons act as a "blanket of inhibition" (Karnani, Agetsuma and Yuste, 2014), they can indeed subserve pathway-specific gating. This work argues that top-down behavioral control may involve rule signals targeting specific interneuron types rather than, or in addition to, pyramidal neurons, and that the disinhibitory motif could play a major role in synaptic plasticity.

## 2.2 Results

### 2.2.1 Pathway-specific gating with dendritic disinhibition

To study dendritic disinhibition, we first built a simplified neuron model with a reduced morphology (**Fig. 2.2a**). It comprises one spiking somatic compartment, and multiple dendritic compartments which are electrically coupled to the soma but otherwise independent of each other. The somatic and dendritic compartments have no spatial extent themselves. This choice of morphology is inspired by previous studies showing that different dendritic branches can integrate their local input independently from one another (Poirazi, Brannon and Mel, 2003).

A prominent feature of active processing in thin dendritic branches is their ability to

produce NMDA plateau potentials (Schiller et al., 2000), also called NMDA spikes. The NMDA plateau potential is a regenerative event in which the membrane potential increases nonlinearly and sometimes sharply with the NMDAR input, due to the release of voltage-dependent magnesium block of NMDARs. The reduced neuron model can exhibit NMDA plateau potential in dendrites (**Fig. 2.2b**), in line with simulations of morphologically reconstructed neuron models (**Fig. 2.3**). The mean dendritic voltage in response to a Poisson spike train input is a sigmoidal function of the input rate, due to the NMDA plateau potential (light blue curve in **Fig. 2.2c**).

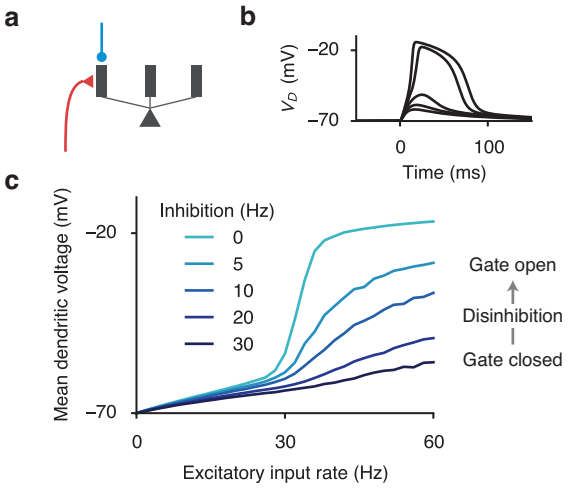


Figure 2.2: Disinhibition of a single dendrite. (**a**) A reduced compartmental neuron with a somatic compartment connected to multiple, otherwise independent, dendritic compartments (only three shown). (**b**) Excitatory inputs can generate a local, regenerative NMDA plateau potential in the dendrite. As number of activated synapses increased, there is a sharp nonlinear increase in the evoked dendritic membrane depolarization ($V_D$). (**c**) Disinhibition of the targeted branch opens the gate for the excitatory input.

The NMDA plateau potential can be prevented by applying a moderate synaptic inhibition, mediated by $GABA_A$ receptors, to the same dendrite (dark blue curve in **Fig. 2.2c**). Inhibition is particularly effective in controlling this dendritic nonlinearity when excitatory inputs are mediated by NMDA receptors with experimentally-observed saturation, in stark contrast to AMPA receptors (**Fig. 2.3**) or NMDA receptors without saturation (**Fig. 2.4**). Inhibitory input also linearizes the relationship between mean dendritic voltage

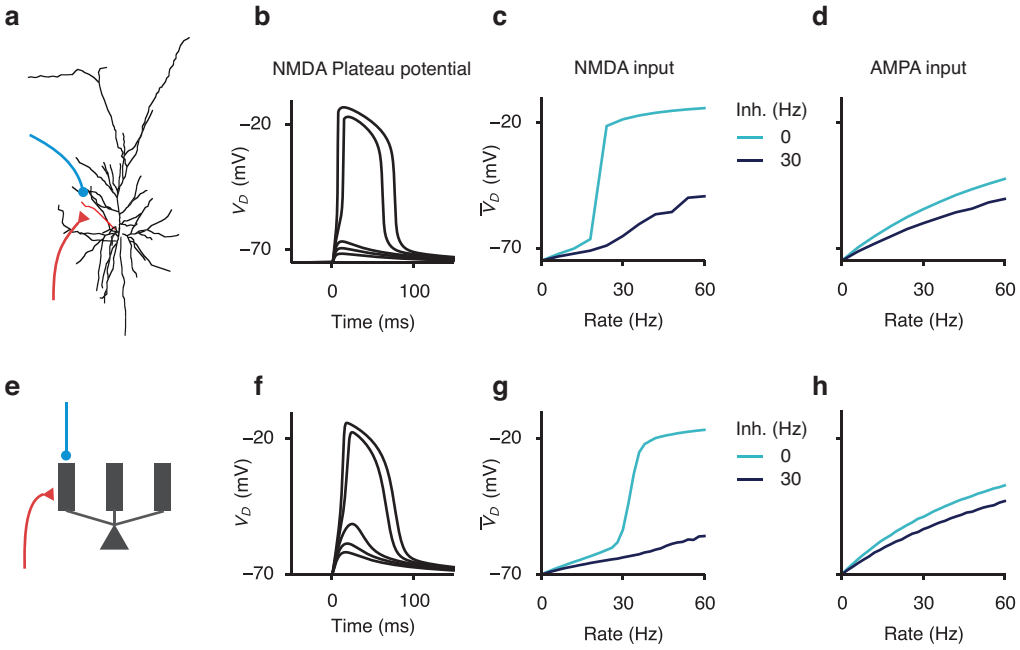


Figure 2.3: Dendritic disinhibition powerfully gates dendritic nonlinearity. (**a-d**) Dendritic disinhibition controls NMDAR-dependent nonlinearity in a reconstructed compartmental neuron model. (**a**) A morphologically reconstructed compartmental model of a layer 2/3 pyramidal neuron (Branco, Clark and Häusser, 2010) receives excitatory and inhibitory inputs uniformly distributed onto one basal dendrite. (**b**) Excitatory inputs can generate a local, regenerative NMDA plateau potential in the dendrite. As the number of activated synapses is increased, there is a sharp nonlinear increase in the evoked dendritic membrane depolarization ($V_D$). (**c-d**) Presynaptic spike times are modeled as Poisson-distributed events. (**c**) In response to synaptic input mediated by NMDAR channels, the mean dendritic voltage across time ($\overline{V}_D$) increases nonlinearly as a function of excitatory rate (light blue). Moderate inhibition largely suppresses NMDA plateau potentials even for high excitatory input rate (dark blue). (**d**) The effect of inhibition is much weaker when excitatory input is mediated by AMPARs. 20 excitatory synapses are used as input in (**c,d**). (**e-h**) A reduced compartmental neuron model captures the nonlinearity of the morphologically reconstructed model. (**e**) A somatic compartment is connected to multiple, otherwise independent, dendritic compartments (only three shown). (**f-g**) Modeling results in the reconstructed neuron model (**b-d**) are reproduced by the simplified model. 15 excitatory synapses are used as input in (**g,h**).

and excitatory input rate (**Fig. 2.2c**), due to stochastic transitions into or out of NMDA plateau potential induced by low-rate inhibition (**Fig. 2.4**). Therefore excitatory inputs to a dendritic branch can be efficiently gated by inhibition (Jadi et al., 2012).

We now consider multiple pathways of inputs targeting distinct sets of dendrites. In the default condition, all dendritic branches receive a high baseline inhibition from dendrite-targeting SOM neurons (Gentet et al., 2012; Lee et al., 2013), closing gates for

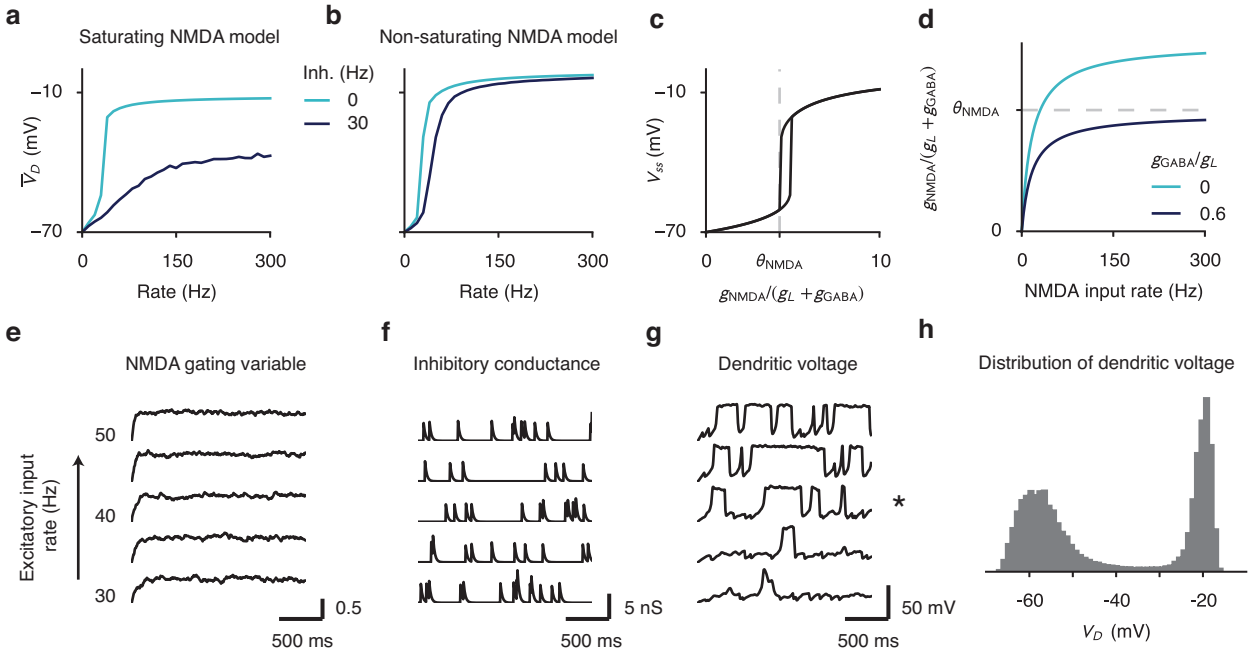


Figure 2.4: Effects of NMDA receptor saturation and low-rate inhibition. (**a-d**) NMDAR saturation allows for much stronger inhibitory control. (**a**) Using an NMDAR model with saturation allows mild dendritic inhibition to powerfully control dendritic voltage. Note that voltage of the inhibited dendrite (dark blue) never reaches the same level as the disinhibited dendrite (light blue). (**b**) The same level of inhibition has a much smaller effect when we used a non-saturating NMDAR model, because inhibition can no longer switches off nonlinear dendritic spikes. (**c**) For constant synaptic conductance, the steady-state voltage of one dendritic branch ($V_{\text{ss}}$) increases sharply with the effective input $g_{\text{NMDA}}/(g_L + g_{\text{GABA}})$, where $g_{\text{NMDA}}$, $g_L$, and $g_{\text{GABA}}$ are the NMDAR, leak, and GABAR conductances, respectively. The dashed line indicates the threshold $\theta_{\text{NMDA}}$ below which $V_{\text{ss}}$ is stably in the low state. (**d**) The NMDR conductance, and therefore $g_{\text{NMDA}}/(g_L+g_{\text{GABA}})$, saturates at high input rates to NMDAR synapses. With moderate inhibition, the saturated value of the effective input can be lower than the threshold $\theta_{\text{NMDA}}$ for an NMDA plateau potential. (**e-h**) Low-rate (temporally sparse) Poisson inhibition generates irregular NMDA plateau potentials and graded encoding of input rate. Inhibition is said to be temporally sparse when the product of the inhibition rate $r_I$ and the time constant $\tau_{\text{GABA}}$ of GABAR is much smaller than 1, i.e. $r_I \cdot \tau_{\text{GABA}} \ll 1$ (**e**) Due to relatively high input rate and long time constant, the NMDAR gating variable averaged across synapses is nearly constant in time. Each trace corresponds to a different excitatory input rate, ranging from 30Hz (bottom) to 50Hz (top); the same applies to (**f,g**). (**f**) Inhibitory conductance is temporally sparse due to a low background inhibition rate of 5 Hz. (**g**) The dendritic voltage switches stochastically in time, into and out of the NMDA plateau potential. (**h**) The dendritic voltage across time exhibits a bimodal distribution, due to stochastic switching. The excitatory rate is set to 40 Hz (asterisk in (**g**)).

all pathways. Disinhibiting the branches targeted by one pathway can selectively open the gate for this pathway while keeping the gates closed for other pathways (**Fig. 2.5a**). When a gate is open, the neuron's output firing rate transmits the stimulus selectivity of the corresponding input pathway most effectively (**Fig. 2.5b**).

When two excitatory pathways are activated simultaneously, we can plot the neuron's response to stimulus variables of both pathways, i.e. the two-dimensional tuning curve (**Fig. 2.5c,d**). In the default condition when all gates are closed, there is little response to either pathway (**Fig. 2.5c**). By specifically disinhibiting the branches targeted by pathway 1, we can open the gate for pathway 1. With gate 1 opened, the neuron is primarily selective to pathway 1 stimuli (**Fig. 2.5d**). The remaining impact of pathway 2 stimuli is due to the fact that the impact of excitatory inputs can never be fully counteracted by dendritic inhibition.

The gating mechanism worsens when a fraction of excitatory input is mediated by AMPARs, but improves when a fraction of inhibitory input is mediated by $GABA_B$ receptors (**Fig. 2.6**). For parsimony, in the following sections, excitatory synaptic inputs are mediated only by NMDARs which are critical to the nonlinear dendritic computations, and inhibitory inputs are mediated only by $GABA_A$Rs.

### 2.2.2 Performance of gating in pyramidal neurons

Which circuit properties determine the effectiveness of pathway-specific gating in our model? A neuron responds to its optimal stimulus from an input pathway with (baseline-corrected) firing rate ($r_{on}$) when the pathway is gated-on, and ($r_{off}$) when the pathway is gated-off, which could be readily measured experimentally. The gating selectivity is then

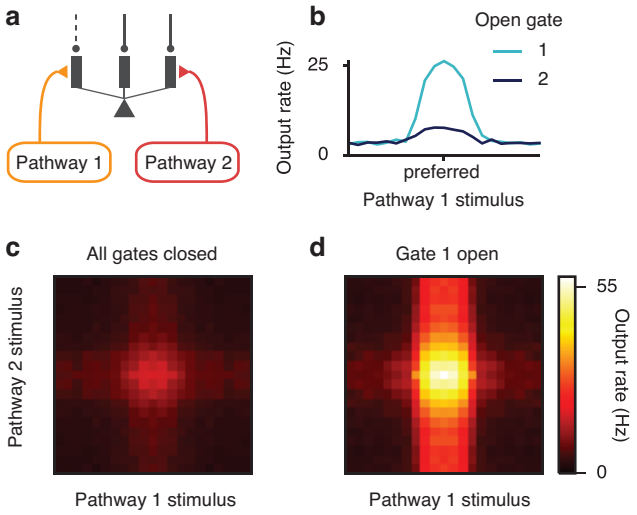


Figure 2.5: Context-dependent gating of specific pathways. (**a**) A pyramidal neuron receives converging inputs from multiple pathways carrying different stimulus features, giving it selectivity to a preferred stimulus for each feature dimension. Each input pathway targets separate dendrites, which are disinhibited correspondingly in each context by top-down control inputs (not modeled here). (**b**) Tuning curve for input pathway 1, when only this pathway is activated. The input pathway encodes a stimulus feature, e.g. motion direction, with a bell-shaped tuning curve for the input. The preferred feature value corresponds to higher input firing rate. When gate 1 is open by disinhibiting the dendrites targeted by input pathway 1, the neuron exhibits strong tuning (light blue). When gate 2 is instead open, the neuron exhibits weak tuning for the feature (dark blue). The amount of inhibition reduced for a disinhibited dendrite, i.e. the disinhibition level, is 30 Hz. (**c,d**) Two dimensional tuning curves when both pathways are activated. (**c**) In the default context, no dendrites are disinhibited and both pathways are gated off. The neuron exhibits weak responses regardless of the stimulus features. (**d**) When gate 1 is open by disinhibiting branches targeted by pathway 1, the response of this neuron is dominated by tuning to the pathway 1 stimulus, although pathway 2 has a residual impact.

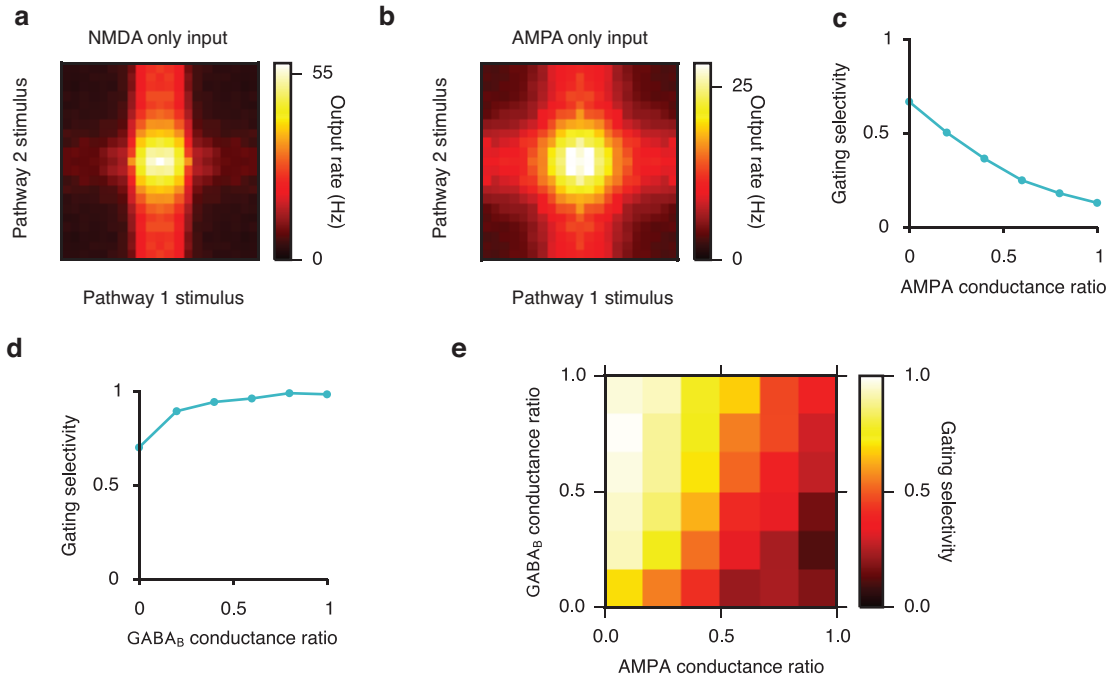


Figure 2.6: Pathway-specific gating with varying levels of AMPAR and GABA$_B$ conductance. In the majority of our work, dendritic excitation is mediated only by NMDARs and dendritic inhibition only by GABA$_B$Rs. Here we show how pathway-specific gating varies with the inclusion of AMPAR and GABA$_B$ inputs. (**a**) Pathway-specific gating when excitatory input is mediated solely by NMDARs, adapted from **Fig. 2g** for comparison. (**b**) When the excitatory input is conducted solely by AMPARs (maximum conductance $\tilde{g}_{\mathrm{AMPA}}$ =2.5 nS for each synapse), the gating performance is strongly degraded. All other conditions are kept the same in (**a**) and (**b**). Disinhibited dendrites receive 30-Hz disinhibition. (**c**) Gating selectivity (which ranges from 0 for no gating to 1 for perfect gating, see Experimental Procedures for the definition) decreases as a function of the AMPA conductance ratio. Here AMPA conductance ratio is defined as $\tilde{g}_{\mathrm{AMPA}}/(\tilde{g}_{\mathrm{AMPA}}+\tilde{g}_{\mathrm{NMDA}})$, which is 0 in the NMDAR-only case and 1 in the AMPAR-only case. $\tilde{g}_{\mathrm{AMPA}}+\tilde{g}_{\mathrm{NMDA}}$ is held constant at 2.5 nS. (**d**) Gating selectivity increases as a function of the GABA$_B$ conductance ratio. This is due to both the slower dynamics of GABA$_B$ receptors and the inward-rectifying potassium (KIR) conductance activated by GABA$_B$ receptors Shoemaker (2011); Sanders et al. (2013). Here excitatory inputs are mediated by NMDARs only. (**e**) Gating selectivity remains high for a wide range of combinations of AMPA and GABA$_B$ conductance ratio.

quantified,

$$\text{Gating Selectivity} = \frac{r_{\mathrm{on}} - r_{\mathrm{off}}}{r_{\mathrm{on}} + r_{\mathrm{off}}}, \tag{2.1}$$

which ranges from 0 (no gating) to 1 (perfect gating). We developed a multi-compartmental rate model (Poirazi, Brannon and Mel, 2003) that greatly improves the efficiency of the circuit model simulation. The rate model is fitted to quantitatively reproduce the activity of the spiking neuron model (**Fig. 2.7**, see Supplemental Information for details).

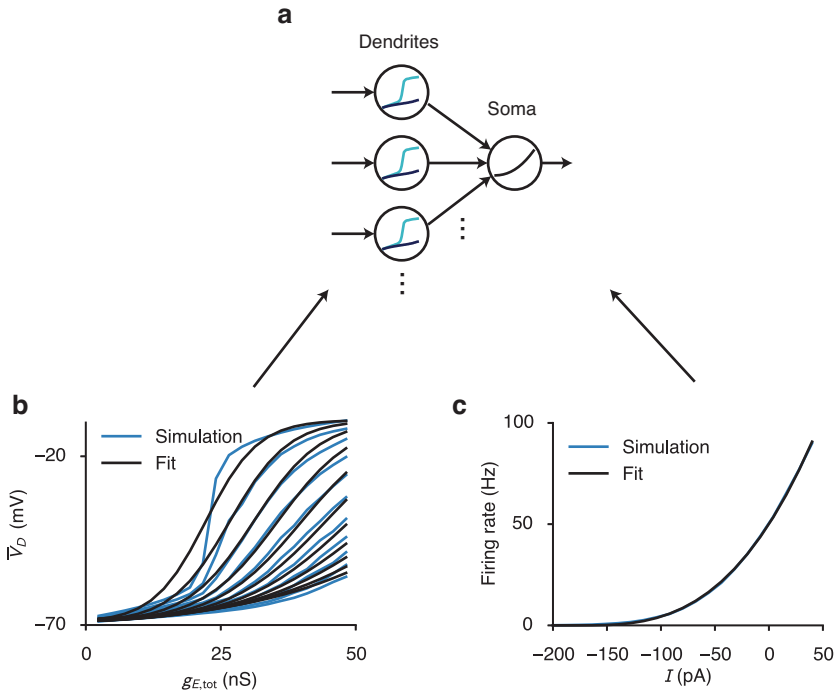


Figure 2.7: Multi-compartment rate model for pyramidal neurons based on the reduced spiking neuron model. (**a**) The neuron model is comprised of multiple dendrite compartments, whose mean voltages are modeled with a family of sigmoidal functions. These dendritic voltages are converted into currents and fed into a somatic compartment, whose firing rate output is modeled with a power-law function. (**b**) The mean dendritic voltage ($\overline{V}_D$) as a function of excitatory and inhibitory inputs. (Blue) Simulation of the reduced-compartmental spiking neuron model. 15 NMDAR inputs fire at a Poisson rate of 30 Hz with conductance ranging from 0.25 to 5.0 nS, resulting in total conductance ($\overline{g}_E$) approximately between 0 and 50 nS. Each curve corresponds to a different inhibitory input rate, ranging uniformly from 0 Hz (top curve) to 100 Hz (bottom curve), in increment of 10 Hz. (Black) Fit of the simulation results. All curves are simultaneously fit with a family of sigmoidal functions, where parameters of the sigmoid, i.e. mid-point and width, are controlled by inhibition. The back-propagating action potential is fixed at a rate of 10 Hz. (**c**) Somatic firing rate as a function of input current from dendrites (and potentially PV neurons). In our model, since at resting state the mean dendritic voltage is lower than the somatic voltage, the input current is negative. The simulation result of the spiking model (Blue) is fit with a power-law function (Black).

We first tested how gating selectivity depends on our assumption of branch-specific disinhibition in a single-neuron setting. Here we assume an alignment of excitation and

disinhibition patterns, which can be achieved through synaptic plasticity as shown later. Each excitatory pathway targets $N_{\text{disinh}}$ randomly chosen dendrites, out of $N_{\text{dend}}$ total dendrites, and this pathway is gated-on by specifically disinhibiting these same $N_{\text{disinh}}$ dendrites (**Fig. 2.8**). Due to the random independent selection of targeted dendrites for each pathway, inputs from two different pathways often overlap.

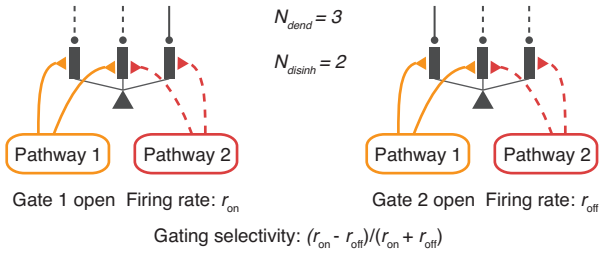


Figure 2.8: Gating selectivity. Schematic of gating, presenting pathway 1 input when gate 1 is opened (left) or gate 2 is opened (right). There are $N_{\text{dend}}$ available dendrites in total. Each input pathway targets $N_{\text{disinh}}$ dendrites. To gate a pathway on, these exact $N_{\text{disinh}}$ dendrites are disinhibited, creating an aligned pattern of disinhibition. Each pathway selects dendrites randomly and independently from other pathways, which can result in overlap of the excitation-disinhibition patterns across pathways. When $N_{\text{disinh}}$ is large, projections from different pathways are more likely to overlap. The neuron's firing rate is $r_{\text{on}}$ and $r_{\text{off}}$ in response to the preferred stimulus of the gated-on (left) and gated-off (right) pathway respectively. The gating selectivity is defined as $(r_{\text{on}} - r_{\text{off}})/(r_{\text{on}} + r_{\text{off}})$, which is 1 for perfect gating and 0 for no gating.

We found that gating selectivity depends critically on the sparseness of the disinhibition (**Fig. 2.9a**), defined as the proportion of targeted/disinhibited dendrites $N_{\text{disinh}}/N_{\text{dend}}$. Gating selectivity improves when disinhibition patterns are sparsened, because the proportion of dendrites that receive overlapping inputs is reduced. We can approximate the limit of $N_{\text{disinh}}/N_{\text{dend}} \to 0$ with non-overlapping disinhibition pattern (diamonds in **Fig. 2.9**). In this case, the gating selectivity is highest but below 1, due to the remaining impact of inputs targeting inhibited dendrites, and is therefore modulated by the level of disinhibition (**Fig. 2.9b**).

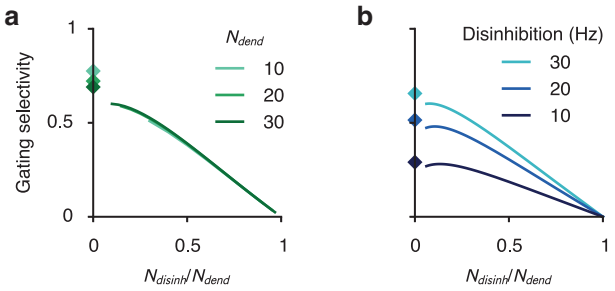


Figure 2.9: Characterization of gating selectivity in pyramidal neurons. (**a**) Gating selectivity increases as excitation/disinhibition patterns become sparser, i.e. with a smaller proportion of targeted and disinhibited dendrites for a pathway ($N_{\text{disinh}}/N_{\text{dend}}$). Diamonds mark the case of non-overlapping excitatory projections, corresponding to the limit of maximal sparseness. (**b**) Gating selectivity is higher with stronger disinhibition, for all sparseness levels.

### 2.2.3 Pathway specific gating with SOM neurons

We have shown that a key determinant of gating performance is the sparseness of innervation patterns onto the dendritic tree. Yet the connectivity from SOM interneurons to pyramidal neurons is dense (Fino and Yuste, 2011). Is it possible for the proposed gating mechanism to function in a cortical microcircuit limited by the dense interneuronal connectivity? To address this issue, we built an interneuronal circuit model, containing hundreds of VIP and SOM interneurons and thousands of pyramidal neurons. We considered "worst-case" conditions in which interneuronal connectivity is completely random (as our gating mechanism can be facilitated by structured connectivity). Surprisingly, we found that relatively high gating performance is achievable under these conditions. We analyzed gating in this circuit in two steps: First, assuming SOM neurons are context-selective, we characterized how the SOM-pyramidal sub-circuit can support high gating selectivity. Second, we characterized how SOM neurons can become context-selective in the VIP-SOM-pyramidal circuit.

First, we built a simplified model of a SOM-pyramidal sub-circuit (**Fig. 2.10a**), which corresponds roughly to a cortical L2/3 column ($400\mu m \times 400\mu m$). The model contains

$N_{\text{pyr}}$ ($\approx 3{,}000$) multi-compartmental pyramidal neurons, each with $N_{\text{dend}}$ ($\approx 30$) dendrites, and $N_{\text{SOM}}$ ($\approx 160$) SOM neurons (see **Table 2.1**). Here we analyze the dependence of gating selectivity on the connectivity from SOM to pyramidal neurons. We consider worst-case conditions in which these connections are random, subject to the SOM-to-pyramidal connection probability of $P_{\text{SOM}\to\text{pyr}}$ ($\approx 0.6$). Assuming that a SOM neuron chooses to target each pyramidal dendrite independently with a SOM-to-dendrite connection probability of $P_{\text{SOM}\to\text{dend}}$, then we have

$$P_{\text{SOM}\to\text{dend}} = 1-(1-P_{\text{SOM}\to\text{pyr}})^{1/N_{\text{dend}}} \tag{2.2}$$

$$\approx P_{\text{SOM}\to\text{pyr}}/N_{\text{dend}} \quad \text{, for small } P_{\text{SOM}\to\text{pyr}} \tag{2.3}$$

Under this assumption, a SOM neuron on average targets $N_{\text{dend}}\cdot P_{\text{SOM}\to\text{dend}}/P_{\text{SOM}\to\text{pyr}} \approx 1.5$ dendrites of a pyramidal neuron given that the two are connected. Each SOM-dendrite connection can correspond to multiple (3-5) clustered synapses (Silberberg and Markram, 2007). So each SOM neuron can make on average 5-8 synapses onto a pyramidal neuron. The connection probability between two neurons is higher at closer proximity (Fino and Yuste, 2011), leading to a even higher number of contacts.

In a default state, SOM neurons fire at a relatively high baseline rate around 10 Hz (Gentet et al., 2012; Lee et al., 2013), closing the gates to all inputs. To open the gate for pathway 1, a randomly chosen subset (50%) of SOM neurons are suppressed, resulting in a pattern of disinhibition across dendrites. Again we assume the excitatory input pattern of pathway 1 is aligned with the corresponding disinhibition pattern. Notably, disinhibition patterns for different pathways generally overlap due to the random selection of SOM neurons and the random connectivity. This overlap can be reduced with either

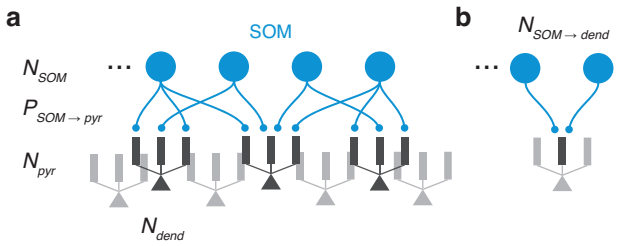


Figure 2.10: Gating selectivity as functions of SOM-pyramidal circuit parameters. (**a**) A simplified model for a cortical column of SOM and pyramidal neurons. We only modeled the SOM-to-pyramidal connections. The model is subject to experimentally-measured constraints of the following parameters: number of SOM neurons ($N_{\text{SOM}}$), connection probability from SOM to pyramidal neurons ($P_{\text{SOM}\to\text{pyr}}$), and the number of dendrites on each pyramidal neuron ($N_{\text{dend}}$). We consider the "worst case" scenario that the SOM-to-dendrite connections are random. Finally we assume for now that control input for each pathway suppresses a random subset of SOM neurons. The different contrasts used are for illustration purpose only. (**b**) A critical parameter for the SOM-to-pyramidal circuit is the number of SOM neurons targeting each dendrite ($N_{\text{SOM}\to\text{dend}}$). This parameter can be calculated using other experimentally-measured parameters under the assumption of random connectivity, $N_{\text{SOM}\to\text{dend}} = N_{\text{SOM}} \cdot [1-(1-P_{\text{SOM}\to\text{pyr}})^{1/N_{\text{dend}}}]$.

structured connections or inhibitory plasticity.

Under the above assumptions, the circuit achieves a mean gating selectivity around 0.5, equivalent to $r_{\text{on}} \approx 3r_{\text{off}}$. We found that the impact of these circuit parameters is determined by one critical parameter: the number of SOM neurons targeting each dendrite $N_{\text{SOM}\to\text{dend}} = N_{\text{SOM}} \cdot P_{\text{SOM}\to\text{dend}} \approx 5$ (**Fig. 2.10b**, see also section **2.5.1**). When we vary parameters while keeping $N_{\text{SOM}\to\text{dend}}$ fixed, the gating selectivity remains largely constant (**Fig. 2.11a-c**). We found that gating selectivity is highest when $N_{\text{SOM}\to\text{dend}}$ is small (**Fig. 2.11d**), and decreases as we increases $N_{\text{SOM}\to\text{dend}}$. Because the overall strength of inhibition has a simple effect on the gating selectivity (**Fig. 2.11b**), we keep it fixed when varying other parameters.

Each dendrite should more appropriately be interpreted as an independent computational unit. When inhibitory connections control individual excitatory connection through pre-synaptic receptors (Urban-Ciecko, Fanselow and Barth, 2015) or by targeting spines (Chiu et al., 2013), the independent unit would be single excitatory synapses.

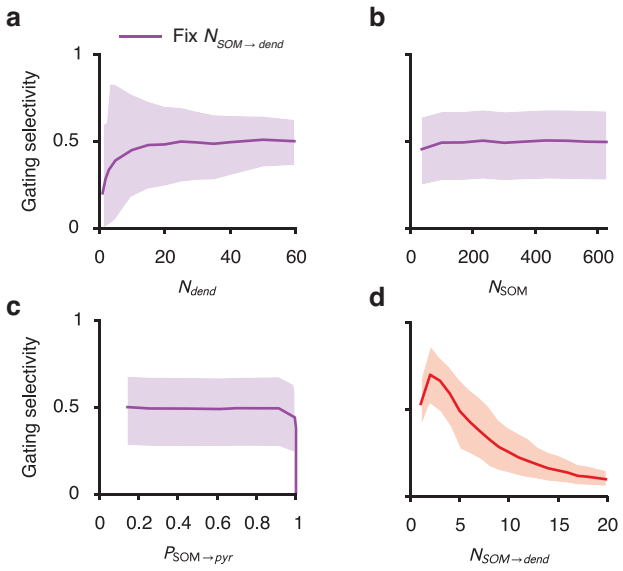


Figure 2.11: Gating selectivity as functions of SOM-pyramidal circuit parameters. (**a-c**) Gating selectivity only weakly depends on $N_{\text{dend}}$ (**a**), $N_{\text{SOM}}$ (**b**), and $P_{\text{SOM}\rightarrow\text{pyr}}$ (**c**) if $N_{\text{SOM}\rightarrow\text{dend}}$ is kept constant by co-varying another parameter. The plotted curve marks the mean and the shaded region marks the bottom 10% to top 10% of the neuronal population. (**d**) Gating selectivity is high when each dendrite is targeted by a few SOM neurons. Given experimental measurements of $P_{\text{SOM}\rightarrow\text{pyr}} \approx 0.6, N_{\text{dend}} \approx 30, N_{\text{SOM}} \approx 160$, we obtained $N_{\text{SOM}\rightarrow\text{dend}} \approx 5$, leading to relatively high gating selectivity $\sim 0.5$. Total strength of inhibition onto each pyramidal dendrite is always kept constant when varying parameters.

This leads to a lower effective value of $N_{\text{SOM}\to\text{dend}}$, then a higher gating selectivity.

### 2.2.4 Pathway specific gating with SOM and VIP neurons

Having analyzed the SOM-pyramidal connectivity, we next examined how SOM neurons can be context-selective, and characterized the gating selectivity in a circuit model containing VIP, SOM, and pyramidal neurons. On top of the previous SOM-pyramidal sub-circuit, We added $N_{\text{VIP}}$ VIP neurons that only target SOM neurons (Pfeffer et al., 2013). Here we assume VIP neurons target all SOM neurons with connection probability $P_{\text{VIP}\to\text{SOM}}$. Broadly speaking, we found two scenarios in which SOM neurons can be suppressed selectively based on the context, depending on the targets of the top-down or locally-generated control inputs (**Fig. 2.12, 2.14**).

In the first scenario, control inputs target VIP neurons solely (**Fig. 2.12a**). In this intuitive scenario, control inputs excite VIP neurons, which in turn inhibit SOM neurons thereby disinhibiting pyramidal dendrites. Gating selectivity is high only if a small proportion of VIP neurons are targeted by control (**Fig. 2.12b**), indicating that VIP neurons must be context-selective, and VIP-SOM connections need to be sparse (**Fig. 2.12c**). VIP-SOM connectivity could possibly be effectively sparse on the scale of a cortical column, since the axonal arbor of VIP neurons are rather spatially restricted (Bayraktar et al., 2000). When varying parameters, we kept fixed the overall baseline inhibition received by each SOM neuron and the overall strength of control inputs.

In the second scenario, excitatory control inputs target both VIP and SOM neurons (**Fig. 2.14a**). If the VIP-SOM connectivity is dense, then VIP neurons activated by control inputs will provide nearly uniform inhibition across all SOM neurons (**Fig. 2.13**). However, SOM neurons can receive selective excitation if the control inputs only directly tar-

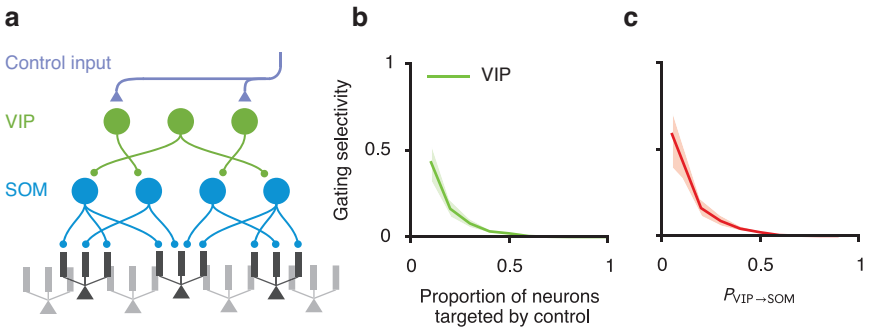


Figure 2.12: Control signals target only VIP neurons. We built a simplified circuit model containing VIP, SOM, and pyramidal neurons. (**a**) In this scheme, for each pathway, control inputs target a random subset of VIP neurons. And the connection probability from VIP to SOM neurons is $P_{\text{VIP}\to\text{SOM}}$. (**b,c**) Good gating selectivity is only achieved when a small subset of VIP neurons is targeted by control inputs (**b**), and when the VIP-SOM connections is sparse (**c**).

get a randomly-chosen subset of SOM neurons. If the inhibition is on average stronger, then the overall effect is a selective suppression of SOM neurons (**Fig. 2.13**). As a result, gating selectivity no longer depends on the proportion of VIP neurons targeted by control inputs, but does depend on the proportion of SOM neurons targeted (**Fig. 2.14b**). Therefore SOM neurons are context-selective, but VIP neurons need not be. Similarly, gating selectivity does not depend on the connection probability from VIP to SOM neurons, $P_{\text{VIP}\to\text{SOM}}$ (**Fig. 2.14c**).

In summary, in order to achieve branch-specific disinhibition, control inputs targeting interneurons have to be selective. Notably, the level of specificity required for the control inputs depends strongly on the neurons they target. When targeting only VIP neurons, the control inputs have to be highly selective (**Fig. 2.12a,b**). However, when control inputs target both VIP and SOM neurons, high gating selectivity can be achieved in a much broader range of parameters, reducing the level of specificity required (**Fig. 2.14a,b**).

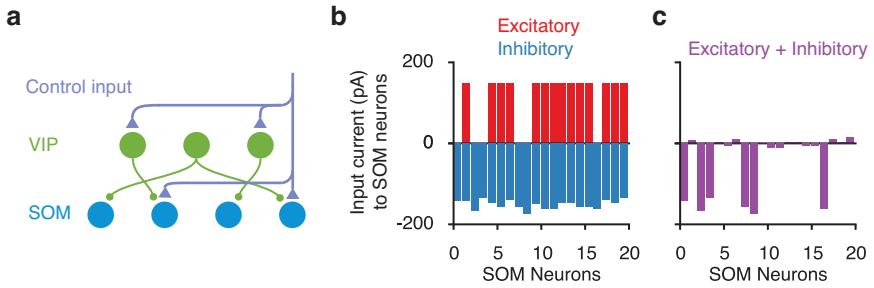


Figure 2.13: Mechanism of control. (**a**) In this scenario, we assume that for each pathway control inputs target a random subset of VIP and SOM neurons. (**b-c**) Input currents onto SOM neurons (only 20 shown). (**b**) 50% of the SOM neurons receive excitatory currents from control (red). 50% of VIP neurons receive excitatory control, but due to the high random connectivity from VIP to SOM neurons, inhibitory currents onto SOM cells are nearly uniform (blue). (**c**) The sum of the excitatory and inhibitory currents onto SOM neurons, i.e. the total currents, are primarily inhibitory and vary strongly across SOM neurons. The overall inhibitory currents are results of overall stronger inhibition. The variability across SOM neurons are mainly inherited from the selective excitatory control input.

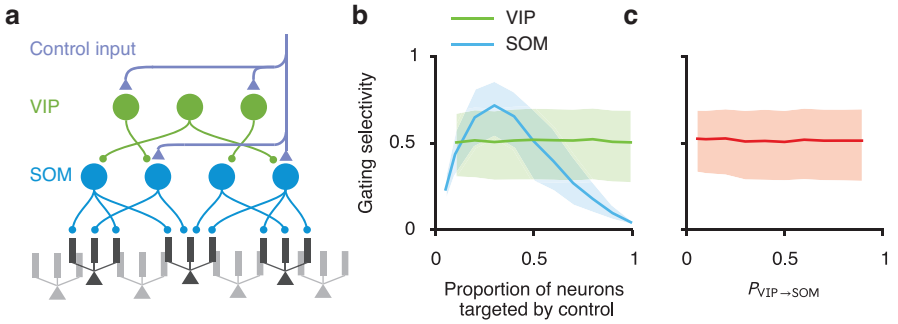


Figure 2.14: Control signals target both VIP and SOM neurons. (**a**) In this scheme, we assume that for each pathway control inputs target a random subset of VIP and SOM neurons. (**b**) Gating selectivity depends on the proportion of SOM (blue) but not VIP (green) neurons targeted by control input. (**c**) Gating selectivity does not depend on $P_{\text{VIP}\to\text{SOM}}$. Curves and shaded regions are as in **Fig. 2.11**.

### 2.2.5 Pathway specific gating with SOM, VIP, and PV neurons

PV neurons receive inhibition from themselves and SOM neurons, and project to perisomatic areas of pyramidal neurons (Markram et al., 2004). Suppression of SOM neurons therefore also leads to disinhibition of PV neurons and an increase of somatic inhibition onto pyramidal neurons. We included PV neurons into our interneuronal circuit model (**Fig. 2.15a**), and found that this inclusion and the consequent increase in somatic inhi-

bition strictly improve gating selectivity in a wide range of parameters (**Fig. 2.15b**). Since the SOM-to-PV and PV-to-pyramidal neuron connections are dense (Karnani, Agetsuma and Yuste, 2014), a selective pattern of SOM suppression will result in an elevated somatic inhibition that is almost uniform across pyramidal neurons (**Fig. 2.16**). Furthermore, we proved that a uniform increase in somatic inhibition will always improve gating selectivity, except when the somatic inhibition is unreasonably strong (see section **2.5.2**).

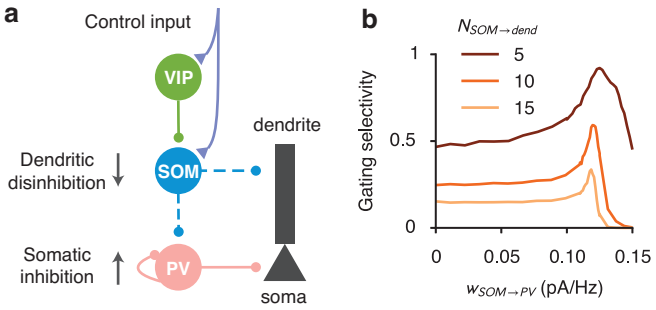


Figure 2.15: Somatic inhibition improves gating selectivity. (**a**) PV neurons project to the somatic areas of pyramidal neurons, and are inhibited by SOM neurons and themselves. Suppression of SOM neurons cause disinhibition of PV neurons, therefore an increase in somatic inhibition onto pyramidal neurons. (**b**) A moderate increase in somatic inhibition always improves gating selectivity. We included PV neurons and their corresponding connections in the model of **Fig. 2.14a**. Gating selectivity increases as a function of the SOM-to-PV connection weights ($w_{\mathrm{SOM}\rightarrow\mathrm{PV}}$) in a wide range (see **Supplementary Note 1** for a proof). However, when gating selectivity is low without PV neurons (light curve), the peak of this increase is lower and the slope is sharper. Gating selectivity starts to decrease when the SOM-to-PV connection, therefore the somatic inhibition, is too strong that the responses of many pyramidal neurons are completely suppressed.

For an intuitive explanation, consider a linear input-output function in the soma. Gating selectivity is based on the relative difference between the pyramidal neuron responses when the gate is open ($r_{\mathrm{on}}$) and when the gate is closed ($r_{\mathrm{off}}$). Providing an equal amount of somatic inhibition in these two conditions is equivalent to subtracting both values by the same constant, which will enhance the relative difference.

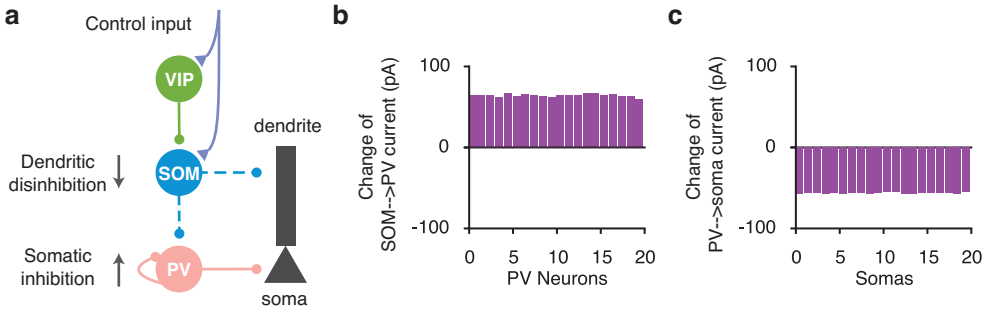


Figure 2.16: Inclusion of PV neurons results in a uniform somatic inhibition across pyramidal neurons. (**a**) SOM suppression lead to PV disinhibition and somatic inhibition. Same as **Fig. 2.15a**. (**b**) The change in the SOM-to-PV input currents after the control input. The change in currents is disinhibitory (net excitatory) with a small standard deviation compared to the mean across PV neurons. Notice that although the control input results in a selective suppression of SOM neurons (**Fig. 2.13c**), the change in the SOM-to-PV currents is almost uniform due to the high SOM-to-PV connection probability. (**c**) The change in the PV-to-soma input currents after the control input is net inhibitory and again uniform across somas. Therefore a selective suppression of SOM neurons results in a non-selective inhibition across somas through PV neurons.

### 2.2.6 Learning pathway-specific gating

A critical feature of our scheme is the alignment between excitation and disinhibition patterns (**Fig. 2.1c**): pyramidal dendrites targeted by an excitatory input pathway are also disinhibited when the gate is open for that pathway. Dendritic disinhibition can regulate synaptic plasticity (Fu, Kaneko, Tang, Alvarez-Buylla and Stryker, 2014; Bar-Ilan, Gidon and Segev, 2012). We hypothesized that such an alignment can naturally arise as a result of the regulated plasticity. To test this hypothesis, we first established a realistic calcium-based plasticity model for dendrites in our reduced spiking neuron model. Pre- and post-synaptic spikes induce calcium transients in dendrites, which determine the synaptic weight changes (Graupner and Brunel, 2012) (**Fig. 2.17a**). We fitted parameters of the model to capture experimental data (Nevian and Sakmann, 2006) (**Fig. 2.18**). Our model also quantitatively predicts findings that were not used in the fitting.

The calcium-based plasticity model allows us to naturally study the effects of dendritic disinhibition on synaptic plasticity and their functional implications. Again we

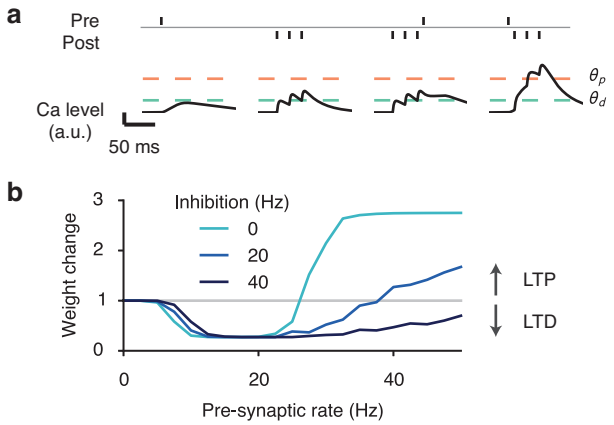


Figure 2.17: Calcium-based plasticity. (**a**) Model schematic. Pre- and post-synaptic spikes both induce calcium influx. The overall synaptic weight change is determined by the amount of time the calcium level spends above thresholds for depression ($\theta_d$) and potentiation ($\theta_p$) (Graupner and Brunel, 2012). The model is fitted to experimental data, and is able to quantitatively predict results not used in the fitting. (**b**) Dendritic inhibition makes potentiation harder to induce. With background-level inhibition (light blue), synaptic weight change shows three regimes as a function of excitatory input rate: no change for low rate, depression for medium rate, and potentiation for high rate. With a medium level of inhibition (dark blue), potentiation requires a higher excitatory input rate. With relatively strong inhibition (black), potentiation becomes impossible within a reasonable range of excitatory input rates. The post-synaptic rate is fixed at 10 Hz.

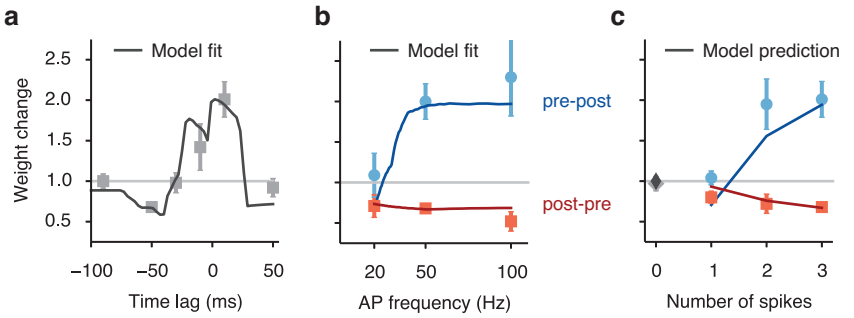


Figure 2.18: Fit and prediction of the plasticity model compared to experimental data. (**a-c**) A pre-synaptic spike is paired with multiple post-synaptic action potentials (AP). Light symbols mark data showing synaptic weight change (weight after learning/weight before learning) when varying the pre-post time lag (**a**), post-synaptic AP frequency (**b**), and number of post-synaptic spikes (**c**). In (**b,c**), the presynaptic spike either precedes (blue) or follows (red) the postsynaptic spikes. Curves in (**a-b**) show the model fit, with the same set of parameters. (**c**) The model generalizes to predict data not used to fit the model. Experimental data are extracted from Nevian and Sakmann (2006).

assume that pre- and post-synaptic firings are Poisson spike trains with specified rates. We found that dendritic inhibition can shift the plasticity from potentiation to depression, even when the pre-synaptic excitatory input rate and the post-synaptic firing rate are both kept constant (**Fig. 2.17b**), consistent with previous modeling findings (Bar-Ilan, Gidon and Segev, 2012). We note that plasticity models based solely on pre- and post-synaptic neuronal firing would not predict the inhibitory modulation of synaptic plasticity.

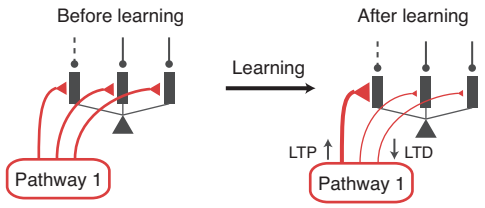


Figure 2.19: Learning to gate specific pathways. (Left) Excitatory synapses from each pathway are initialized uniformly across dendrites. When pathway 1 is activated, specific branches of the neuron are disinhibited (dashed line), i.e. gate 1 is open. During learning, only one pathway is activated at a time. (Right) After learning, activated excitatory synapses onto the disinhibited branches are strengthened, while activated synapses onto inhibited branches are weakened, resulting in an alignment of excitation and disinhibition patterns. Synaptic weights of non-activated synapses remain unchanged (not shown).

We then tested whether disinhibitory regulation of plasticity can support the development of excitation-disinhibition alignment, as needed for pathway-specific gating (**Fig. 2.19**). Importantly, the strength of disinhibition is realistic, similar to those used throughout this paper. Initially, excitatory synapses from each pathway are uniformly distributed across the dendritic branches of single neurons. Different excitatory pathways are then activated one at a time. Whenever a pathway is presented, a particular subset of dendrites are disinhibited, while the rest of the dendrites remain inhibited. Through calcium-based excitatory plasticity, the activated excitatory synapses targeting the disinhibited dendrites become strengthened, whereas those targeting the inhibited dendrites become weakened. Synapses not activated remain the same regardless of the inhibition level (see

**Fig. 2.17b**). After learning, the alignment of excitation and disinhibition patterns support pathway-specific gating (**Fig. 2.20**; compare with **Fig. 2.5b**), with a gating selectivity around 0.7. These findings show that a key aspect of the gating architecture, namely the alignment of excitation and disinhibition patterns, can emerge naturally from the interaction between excitatory synaptic plasticity and context-dependent disinhibition.

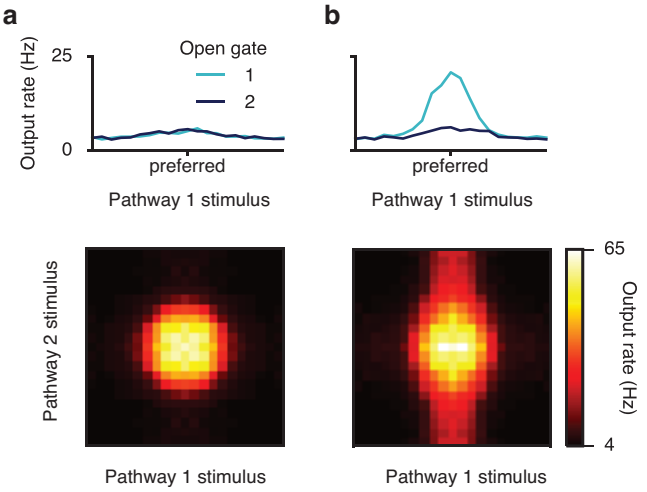


Figure 2.20: Response properties of the neuron before and after learning. (**a**) Before learning. (Top) Tuning curve of the neuron when only pathway 1 is presented. The neuron shows no preference to the gate opened prior to learning. (Bottom) Two-dimensional tuning curve of the neuron when both pathways are simultaneously presented and gate 1 is open. See **Fig. 2** for the definition of the tuning curves. (**b**) After learning. (Top) The neuron shows strong tuning to pathway 1 input when gate 1 is open. (Bottom) When both pathways are presented, the neuron's response is primarily driven by pathway 1 stimulus, although pathway 2 stimulus also affects the neuron's firing.

### 2.2.7 Modeling a flexible behavior with pathway-specific gating

How is gating at the neural level related to gating at the behavioral level? Is moderate gating selectivity (e.g., ~0.5 as above) sufficient to explain performances in flexible cognitive tasks? To address these issues, we applied our model to a context-dependent decision-making task (Mante et al., 2013). In this task, the behavioral response should be based on either the motion direction or the color of a random-dots motion stimulus, depending

on the context cued by a rule signal (**Fig. 2.21a**).

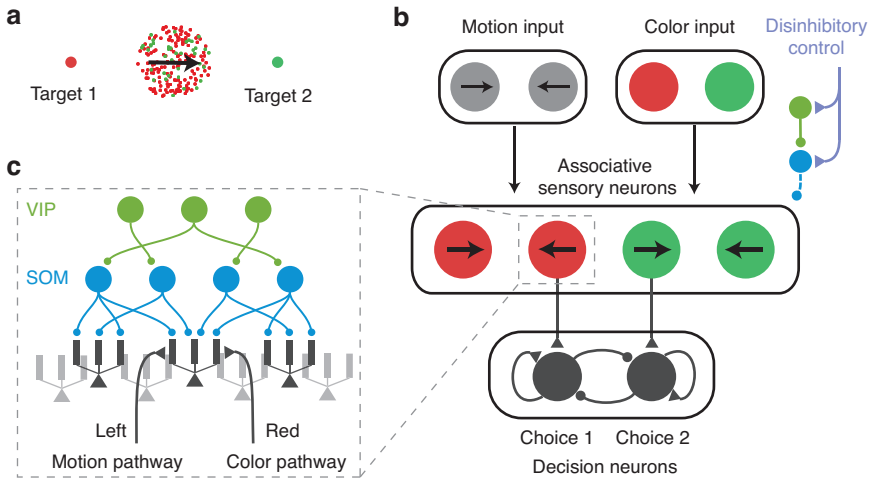


Figure 2.21: Pathway-specific gating in an example context-dependent decision-making task. (**a**) A flexible decision-making task. Depending on the context, subject's behavioral response should be based on either the color or the motion direction of the stimulus. (**b**) The circuit model scheme. Motion and color pathways target associative-sensory neurons, which are subject to context-dependent disinhibitory control. Neurons preferring color and motion evidence for the same target project to the corresponding neural pool in the decision-making circuit. (**c**) Associative-sensory neurons receive converging inputs from both motion and color pathways, and are controlled by the dendrite-targeting interneuronal circuit.

We built a stylized neural circuit model to implement this task using pathway-specific gating through dendritic disinhibition (**Fig. 2.21b**). The local circuit comprises a sensory network and a decision network. The sensory network contains pyramidal neurons that receive convergent sensory inputs from both motion and color pathways, and they group into four pools according to their selectivities to color and motion evidence. The dendrites of pyramidal neurons are controlled by the VIP-SOM interneuronal circuit described above (**Fig. 2.21c, 2.14a**). A subset of pyramidal neurons with high gating selectivity projects to the decision network. Pyramidal neurons representing color and motion evidence for the same target project to the corresponding decision neural pool. The decision network, as modeled previously (Wong and Wang, 2006), is a strongly recurrent network that generates a winner-take-all decision based on its inputs.

We fitted the performance of the model to a monkey's psychometric behavioral data from (Mante et al., 2013), using three free parameters in the model, namely the proportion of sensory neurons that project to the decision network, and the overall connection strengths from the input pathways to the sensory network and from the sensory network to the decision network. By fitting these three parameters, we obtained a quantitative match of the empirical psychometric performance, as a function of relevant (**Fig. 2.22a**) and irrelevant (**Fig. 2.22b**) features. Our model shows that the impact of the irrelevant information should be stronger when the relevant information is more ambiguous (with lower motion coherence, for instance) (**Fig. 2.22c**). Although at its default parameters the interneuronal circuit model can show similar task performance as the empirical data, we found that it can no longer fit the empirical performance if we significantly degrade the neural gating selectivity (**Fig. 2.23**). This simulation therefore serves as a proof-of-principle to demonstrate the potential of dendritic disinhibition as a mechanism for pathway-gating, and as a link to assess the utility of neural gating selectivity in terms of flexible behavioral performance.

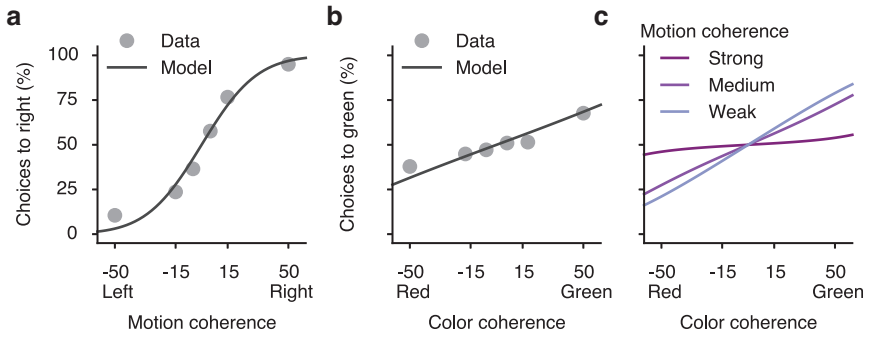


Figure 2.22: Fit and prediction of behavioral performance. Behavioral performance in the motion context as a function of motion coherence (**a**) and color coherence (**b**) for a monkey (dots), and the model's fit (line). Experimental data are extracted from (Mante et al., 2013). The model can capture the behavioral performance of a monkey. (**c**) In the model, impact of the irrelevant pathway (color) is strongest when the relevant pathway signal is weak (with low motion coherence).

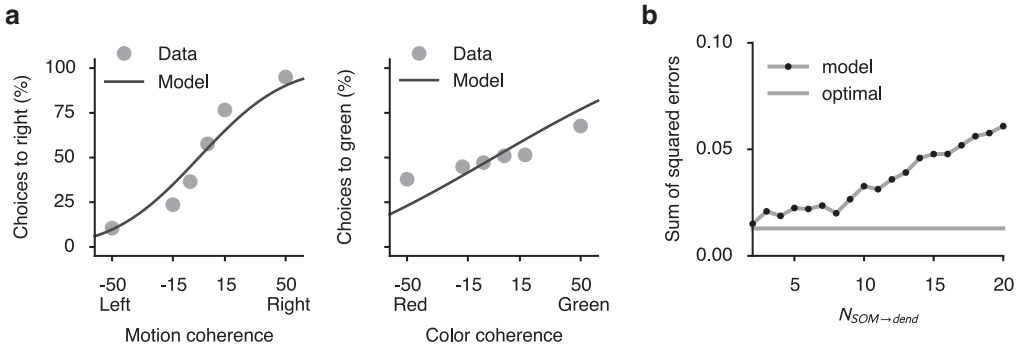


Figure 2.23: Fits of behavioral data as we vary parameters of the interneuronal circuit. (**a**) The model fit to behavioral data in motion context when we set $N_{\mathrm{SOM}\rightarrow\mathrm{dend}} = 20$. The fit is much degraded compared to **Fig. 2.22**. From **Fig. 2.11** we know that $N_{\mathrm{SOM}\rightarrow\mathrm{dend}}$ is the critical parameter for gating selectivity measured on the neural level. (**b**) The sum of squared errors of the model fit as a function of $N_{\mathrm{SOM}\rightarrow\mathrm{dend}}$. For a large range of $N_{\mathrm{SOM}\rightarrow\mathrm{dend}}$, the model can nearly fit the data optimally. The fit starts to degrade when $N_{\mathrm{SOM}\rightarrow\mathrm{dend}} > 10$. Dashed line indicates the error level of the optimal sigmoidal fit, where data are directly fitted to logistic functions. The sum of squared errors shown here is the median error of 50 different model realizations and fits.

## 2.3 Discussion

A canonical cortical microcircuit motif specialized for disinhibition of pyramidal neuron dendrites was proposed theoretically (Wang et al., 2004) and has received strong empirical support from a series of recent experiments (Pi et al., 2013; Zhang et al., 2014; Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014; Lee et al., 2013; Gentet et al., 2012; Pfeffer et al., 2013; Lovett-Barron et al., 2012). Here we explored the functional roles of dendritic disinhibition using computational modeling, at both the single-neuron and circuit levels. In contrast to somatic disinhibition, dendritic disinhibition can gate the inputs to a neuron (Wang et al., 2004; Jadi et al., 2012; Sridharan and Knudsen, 2015). We propose that dendritic disinhibition can be utilized to gate inputs from separate pathways, by specifically disinhibiting dendrites that receive inputs from a target pathway.

We studied the effectiveness of gating in an interneuronal circuit model. Most data used to constrain the model have become available only in recent years thanks to the advance of optogenetics and other experimental tools. Where data are not available, we

considered the "worst-case scenario", namely, connections from VIP to SOM neurons, and from SOM to pyramidal dendrites are completely random, which is most likely not the case (Chen et al., 2015) and any specificity would facilitate our proposed mechanism. Although the SOM-to-pyramidal connections are dense, we found that the connectivity from SOM neurons to pyramidal dendrites are actually sparse enough to support branch-specific disinhibition. We found that the increase of somatic inhibition mediated by the SOM-PV-pyramidal neuron connections further improves gating selectivity. We demonstrated that branch-specific clustering of excitatory pathways can naturally emerge from disinhibitory regulation of synaptic plasticity. As proof of principle, we applied this mechanism to a model for a recent experiment using a context-dependent decision-making task (Mante et al., 2013).

Inhibitory connections in cortex tend to be dense (Fino and Yuste, 2011). This finding has led to the proposal that cortical inhibition functions as a locally non-selective "blanket of inhibition" (Karnani, Agetsuma and Yuste, 2014). Our study offers an alternative perspective, which is compatible with dense interneuronal connectivity, but has different implications for circuit functions. The dense connectivity is measured on a cell-to-cell level. Nonetheless, connections from dendrite-targeting SOM interneurons can be sparse at the level of the dendritic branch, and therefore potentially selective as required for our gating scheme. Our alternative proposal is fundamentally grounded in consideration of dendritic branches as functional units of computation (Poirazi, Brannon and Mel, 2003).

### 2.3.1 Model requirements and assumptions

Our hypothesis has three essential requirements. First, dendritic inhibition must be able to effectively regulate dendritic processing of excitatory inputs. This has been shown in our simulation of a fully reconstructed and a simplified neuron model (**Supplementary Fig. 1**,**Fig. 2**), as well as in previous models (Jadi et al., 2012).

The second requirement states that dendritic disinhibition should be specific at the level of dendritic branches. In a simplified interneuronal circuit model, we showed that the SOM-to-pyramidal neuron circuit can very well support branch-specific disinhibition, mainly because the connectivity from SOM neurons to pyramidal dendrites is effectively sparse. Several circuit components are identified as critical to this requirement and subjects to experimental tests, including the number of SOM neurons targeting each dendrite, and the proportion of SOM neurons targeted by top-down control inputs.

The third and final requirement is an alignment between excitation and disinhibition, which we found can be achieved through synaptic plasticity on excitatory synapses. This feature could also potentially be achieved through inhibitory plasticity (Vogels et al., 2011), by adapting the disinhibition pattern to align with fixed excitatory inputs. These two forms of plasticity are complementary, and both are likely at play. Indeed, a recent study found that during motor learning, spine reorganization on dendrites of pyramidal neurons is accompanied by change in the number of SOM-neuron synapses onto these dendrites (Chen et al., 2015). One appeal of studying excitatory plasticity here is that our calcium-based plasticity model could be quantitatively constrained by data and therefore tested in a biologically plausible regime. At present, much less is known experimentally about the dependence of inhibitory plasticity on pre- and post-synaptic spike timing, dendritic calcium levels, or the class of interneuron (Xue, Atallah and Scanziani, 2014).

Although all necessary, the three requirements only need to be fulfilled to a certain degree. For example, dendritic inhibition needs to effectively regulate, but need not abolish, the effects of excitatory inputs. It would be of interest to investigate which requirements pathway-specific gating is most sensitive to, in future research.

A direct consequence of the branch-specific disinhibition and excitation-disinhibition alignment is branch-specific excitation, namely clustering of excitatory pathways onto pyramidal neuron dendritic branches. The computational benefits of input clustering have been previously proposed (Koch, Poggio and Torre, 1982). There is mounting experimental evidence for input clustering, from anatomical and physiological studies (Druckmann et al., 2014; Yang et al., 2014)(for a review see Kastellakis et al., 2015). Consistent with our model, experimental studies have shown that input clustering can emerge through NMDAR-dependent synaptic plasticity (Kleindienst et al., 2011), and that clustering is functionally related to learning (Fu et al., 2012; Yang et al., 2014). Our model prediction that branch-specific clustering can occur at the level of pathways remains to be directly tested.

We focused on whether the disinhibitory circuit motif can support pathway-specific gating. Our model contains only the minimal elements to answer this question. We did not include all known connections between the four major classes of neurons. In particular, among the missing connections are the pyramidal-to-interneuron and the SOM-to-VIP connections. These connections are unlikely to influence our results, since none affect our essential requirements. It remains to be tested whether the pyramidal-to-pyramidal recurrent excitation will interfere with the plasticity-based induction of excitation-disinhibition alignment. We did not model synaptic dynamics or short-term plasticity, because we focused on the steady-state behavior.

### 2.3.2 Model predictions

Our model makes specific, experimentally testable predictions. One of the most straightforward and testable predictions is that SOM neurons should show context/rule selectivity in some context-dependent or rule-based tasks. Surprisingly, we found that VIP neurons need not be context-selective, as long as SOM neurons are directly receiving context-selective excitatory control inputs (**Fig. 5d-f**). Experimental disruption of these context-selective interneurons should impair the animal's ability to perform context- or rule-dependent choice tasks. The context-selectivity of SOM or VIP neurons is not necessarily presented in every behavioral task. For instance, a recent study, recording in mouse prefrontal cortex during a auditory discrimination task, found highly homogeneous responses within SOM and VIP populations (Pinto and Dan, 2015). We propose that SOM neurons are more likely to exhibit selectivity to context or task in experiments in which the animal performs multiple tasks and branch-specific dendritic spikes also exhibit task selectivity (Cichon and Gan, 2015; Adler and Gan, 2015). A direct test of our model awaits future experiments in a task-switching paradigm in order to examine gating of different pathways into association cortical areas and the selective changes of activity in SOM neural subpopulations. We emphasize that interneuron classes in our model should be more appropriately interpreted according to their projection targets rather than their biochemical markers.

Branch-specific dendritic spikes are already observed experimentally, and SOM neurons are critical for this branch-specificity (Cichon and Gan, 2015). It is however unknown whether SOM-mediated inhibition is also branch-specific. Direct patch-clamping of pyramidal-neuron dendrites *in vivo* (Smith et al., 2013) can isolate inhibitory currents on individual branches, and provide a direct test for our hypothesis, although such an

experiment is technically difficult at present. Our plasticity model predicts that SOM interneurons play a critical role in the learning-related emergence of branch-specific clustering of excitatory synapses on pyramidal neuron dendrites (Yang et al., 2014).

### 2.3.3 Relation to other gating models

Flexible gating, or routing, of information has been a long-standing problem in computational neuroscience (Abbott, 2006), for which a number of models have been proposed. Among proposed ideas are dynamic synaptic weight modulation (Olshausen, Anderson and Van Essen, 1993), gain modulation (Abbott, 2006), synchrony in the input signals (Akam and Kullmann, 2010), perfect balance of excitation and inhibition (Vogels and Abbott, 2009), up/down state-switch in dendrites (Kepecs and Raghavachari, 2007), switching between different neural pools that receive inputs from distinct pathways (Zylberberg et al., 2010), and rule signaling as a selection vector (Mante et al., 2013). Notably, most of these models implement a form of soft gating, which modulates the effective strength of incoming pathways instead of performing a binary on-off switch on them.

These prior models did not exploit the computational power of dendrites (except for (Kepecs and Raghavachari, 2007)) or the roles of specialized classes of interneurons. Harnessing dendrites rather than populations of intermediate neurons saves the number of neurons needed by many-fold. Only in the limit of one dendrite per pyramidal neuron does our mechanism become conceptually similar to gating mechanisms operating on the neuronal level (Zylberberg et al., 2010).

The gating mechanism as studied here is nonlinear but not binary. Typically, the gating selectivity is 0.5 or lower (Figures 4-5) whereas it would be one if gating was perfect. In the biologically plausible regime of inhibitory strength studied here, shunting inhibi-

tion on a dendritic branch still allows synaptic input to appreciably elevate the dendritic voltage and thus impact the soma, which decreases the gating selectivity of the neuron. Gating selectivity is also limited by the number of dendritic branches (or more generally, quasi-independent computational units) on a pyramidal neuron, which is about two dozen. Thus, the proposed mechanism may be especially suitable for “pathway specific gating". Multiple mechanisms may jointly contribute to gating function, and our proposed mechanism is most likely compatible with the aforementioned proposals.

A testable prediction of our model is that a behavioral context or rule guiding routing of information engages top-down signaling which targets specific classes of VIP and SOM inhibitory neurons. This is in contrast to the conventional thinking of executive control as mediated by top-down control signals to pyramidal cells. The present model suggests that a context signal can correspond to a top-down input from a brain area presenting task context onto VIP and SOM cells (leading to effective excitation or inhibition, respectively, of pyramidal neurons).

To conclude, our findings suggest a microcircuit architecture that harnesses dendritic computation and diverse inhibitory neuron types to subserve cognitive flexibility.

## 2.4 Methods

### 2.4.1 Spiking pyramidal neuron models

For the fully reconstructed multi-compartmental pyramidal neuron model (**Fig. 2.3a-d**), we adapted a previously developed model based on a layer 2/3 pyramidal neuron in the rat somatosensory cortex reported by (Branco, Clark and Häusser, 2010). We used the passive membrane parameter set; results are essentially the same with the active mem-

brane parameter set. Simulations were implemented with the NEURON simulator (Hines and Carnevale, 1997).

The reduced multi-compartmental spiking neuron model is comprised of multiple dendritic compartments and one somatic compartment. All dendritic compartments are equivalent, not directly coupled to each other, and coupled to the soma. There are 10 dendritic compartments for all simulations using this model, **Fig. 2.5** and**Fig. 2.20**. The number of dendrites does not change the results as long as we normalize the dendrite-soma coupling strength with respect to the number of dendrites. The soma is modeled as a leaky-integrate-and-fire compartment with dynamics following:

$$C_S \frac{\mathrm{d}V_S}{\mathrm{d}t} = -g_{L,S}(V_S - E_L) - \sum_i g_c(V_S - V_{i,D}) + I_{\mathrm{syn},S} \tag{2.4}$$

where the subscripts $S$ and $D$ correspond to soma and dendrites, respectively. $V_{i,D}$ is the membrane potential of the $i-$th dendrite. $C_S$ is the membrane capacitance, $E_L$ is the resting potential, $g_L$ is the leak conductance, $g_c$ is the coupling between each dendritic compartment and the somatic compartment. We set $C_S = 50.0$ pF and $g_{L,S} = 2.5$ nS, producing a 20-ms membrane time constant for soma. We also set $E_L = -70$ mV and $g_c = 4.0$ nS. The somatic spiking mechanism is integrate-and-fire, with spike threshold $-50$ mV, reset potential $-55$ mV, and refractory period 2 ms. The dynamics of the dendritic membrane potential ($V_D$) follows

$$C_D \frac{\mathrm{d}V_D}{\mathrm{d}t} = -g_{L,D}(V_D - E_L) - g_c(V_D - \hat{V}_S) + I_{\mathrm{syn},D} \tag{2.5}$$

where $\hat{V}_S$ is the somatic shadow potential (Murphy and Miller, 2003), which follows the same equation as $V_S$, except with no spiking and resetting. We set $C_D = 20.0$ pF and

$g_{L,D}$ = 4.0 nS, producing a 5-ms membrane time constant (Nevian et al., 2007). After a somatic spike, the back-propagating action potential is modeled as an 3-ms-delayed voltage increase of 10 mV in all dendrites (Larkum, Senn and Lüscher, 2004).

The main free parameters of the reduced-compartmental model, $g_c$ and $g_{L,D}$, were chosen to match *in vitro* properties reported by (Nevian et al., 2007). Specifically, a single-synapse dendritic EPSP of 1-mV peak is attenuated to about 0.05 mV in the soma, and a dendritic NMDA plateau potential evokes a somatic depolarization with the peak around 10 mV. We also made several efforts to adapt our model to mimic physiological *in vivo* conditions, including excitation-inhibition balanced background inputs and reduced soma-dendrite coupling. We used an *in-vivo* set of parameters whenever appropriate (**Fig. 2.5** and **Fig. 2.20**). The soma-dendrite coupling is reduced five-fold to $g_{c,vivo}$ = 0.8 nS, to achieve the stronger signal attenuation observed in high-conductance state (Rudolph and Destexhe, 2003). In this regime, the soma also receives excitatory and inhibitory background inputs, 500 Hz of 2.5-nS AMPAR input and 150 Hz of 4.0-nS GABAR input, to approximate the excitation-inhibition balanced background input that gives the neuron a baseline Poisson-like firing rate around 3 Hz. Reduced spiking neuron simulations were implemented with the BRIAN neural simulator (Goodman and Brette, 2008).

We used four types of synapses, AMPAR, NMDAR, $GABA_AR$, and $GABA_BR$. Since $GABA_BRs$ are only used briefly in (**Supplementary Fig. 3**), we denote $GABA_A$ simply as GABA. AMPAR and GABAR synapses are modeled as linear:

$$I_{\text{syn}} = -\tilde{g}_{\text{syn}} s_{\text{syn}} (V - E_{\text{syn}}) \tag{2.6}$$

$$\frac{\mathrm{d}s_{\text{syn}}}{\mathrm{d}t} = -\frac{s_{\text{syn}}}{\tau_{\text{syn}}} + \sum_i \delta(t - t_i) \tag{2.7}$$

where $s_{\text{syn}}$ is the gating variable representing the proportion of open channels, $\tilde{g}_{\text{syn}}$ is the maximum synaptic conductance, $E_{\text{syn}}$ is the synaptic reversal potential, $\tau_{\text{syn}}$ is the synaptic time constant, and $t_i$ are pre-synaptic spike times. We set $\tau_{\text{AMPA}} = 2$ ms, $E_{\text{AMPA}} = E_E = 0$ mV, $E_{\text{GABA}} = E_I = -70$ mV, and $\tilde{g}_{\text{GABA}} = 4.0$ nS. For dendrite-targeting inhibitory synapses $\tau_{\text{GABA,dend}} = 20$ ms, whereas $\tau_{\text{GABA,soma}} = 10$ ms for soma-targeting inhibitory synapses. These are based on the observations that dendrite-targeting inhibition tend to be slower (Marlin and Carter, 2014; Ali and Thomson, 2008). In **Fig. 2.3d,h**, $\tilde{g}_{\text{AMPA}} = 2.5$ nS. In **Fig. 2.6**, $\tilde{g}_{\text{AMPA}}$ ranges from 0 to 2.5 nS. Otherwise $\tilde{g}_{\text{AMPA}}$ is set as 0 nS (no AMPAR input).

$\text{GABA}_\text{B}$R synapses are post-synaptic. Each spike at time $t_i$ increases the gating variable $s_{\text{GABA}_\text{B}}(t)$ by $\gamma_{\text{GABA}_\text{B}}[\exp[(t-t_i)/\tau_{\text{GABA}_\text{B},\text{decay}}] - \exp[(t-t_i)/\tau_{\text{GABA}_\text{B},\text{rise}}]]$, where $\gamma_{\text{GABA}_\text{B}}$ is a normalizing factor such that the peak of the above expression is 1. Then the total input current voltage dependent is

$$I_{\text{GABA}_\text{B}} = -\tilde{g}_{\text{GABA}_\text{B}} s_{\text{GABA}_\text{B}} f_{\text{GABA}_\text{B}}(V) \tag{2.8}$$

where $f_{\text{GABA}_\text{B}}(V) = 33.33\text{mV} \cdot (0.5 - 2/(1+\exp((V+98.73)/12.5)))$, as obtained from (Shoemaker, 2011).

NMDAR synapses include a voltage-dependent magnesium block $f_{\text{Mg}}(V)$ and saturating gating variable $s_{\text{NMDA}}$:

$$I_{\text{NMDA}} = -\tilde{g}_{\text{NMDA}} s_{\text{NMDA}} (V - E_E) f_{\text{Mg}}(V) \tag{2.9}$$

$$f_{\text{Mg}}(V) = \left[1 + \exp\left(-\frac{V - V_{\text{half}}}{V_{\text{width}}}\right)\right]^{-1} \tag{2.10}$$

with $V_{\text{half}} = -19.9$ mV and $V_{\text{width}} = 12.48$ mV (Ascher and Nowak, 1988). The NMDA con-

ductance $\tilde{g}_{\mathrm{NMDA}} = 2.5$ nS. The NMDAR gating variable dynamics follow:

$$\frac{\mathrm{d}s_{\mathrm{NMDA}}}{\mathrm{d}t} = -\frac{s_{\mathrm{NMDA}}}{\tau_{\mathrm{NMDA,decay}}} + \alpha_{\mathrm{NMDA}} x_{\mathrm{NMDA}}(t)(1 - s_{\mathrm{NMDA}}) \tag{2.11}$$

$$\frac{\mathrm{d}x_{\mathrm{NMDA}}}{\mathrm{d}t} = -\frac{x_{\mathrm{NMDA}}}{\tau_{\mathrm{NMDA,rise}}} + \sum_i \delta(t - t_i) \tag{2.12}$$

with $\tau_{\mathrm{NMDA,decay}} = 100$ ms, $\tau_{\mathrm{NMDA,rise}} = 2$ ms, and $\alpha_{\mathrm{NMDA}} = 0.3$ ms$^{-1}$. This choice of $\alpha_{\mathrm{NMDA}}$ sets $s_{\mathrm{NMDA}}$ to be roughly 0.4 at its peak after a single spike (Ishikawa, Sahara and Takahashi, 2002; Popescu et al., 2004). With this value of $\alpha_{\mathrm{NMDA}}$, the saturation of NMDA starts to get prominent around firing rate $r = 1/(\alpha_{\mathrm{NMDA}} \tau_{\mathrm{NMDA,rise}} \tau_{\mathrm{NMDA,decay}}) \approx 16$ Hz. By default in simulations with the reduced spiking model, the excitatory inputs are 15 independent NMDAR synapses with the same rate. Fewer number of excitatory synapses can become insufficient to elicit NMDA plateau potential. Since GABAR and AMPAR synapses are linear, their inputs are directly represented by the overall rates.

Each excitatory synapse also has a calcium concentration level with arbitrary unit, which consists of two components, one NMDAR-dependent and one voltage-gated calcium-channel (VGCC) dependent: $[\mathrm{Ca}^{2+}] = [\mathrm{Ca}^{2+}]_{\mathrm{NMDA}} + [\mathrm{Ca}^{2+}]_{\mathrm{VGCC}}$. The NMDAR-dependent component is modeled as leaky integration of the NMDAR current:

$$\tau_{\mathrm{Ca,decay}} \frac{\mathrm{d}[\mathrm{Ca}^{2+}]_{\mathrm{NMDA}}}{\mathrm{d}t} = -[\mathrm{Ca}^{2+}]_{\mathrm{NMDA}} + \kappa_{\mathrm{NMDA}} I_{\mathrm{NMDA}} \tag{2.13}$$

where $\kappa_{\mathrm{NMDA}}$ is a scaling parameter with unit pA$^{-1}$. The VGCC component is evoked by post-synaptic spikes that back-propagate into dendrites. Each spike induces a bi-exponential increase:

$$[\mathrm{Ca}^{2+}]_{\mathrm{VGCC}}(t) = \kappa_{\mathrm{VGCC}} \gamma_{\mathrm{Ca}} \sum_i \left[ \exp\left(-\frac{t - t_i}{\tau_{\mathrm{Ca,decay}}}\right) - \exp\left(-\frac{t - t_i}{\tau_{\mathrm{Ca,rise}}}\right) \right] \tag{2.14}$$

Here $\gamma_{\text{Ca}}$ is a normalizing constant so that the peak response to one spike is $\kappa_{\text{VGCC}}$. And $\kappa_{\text{VGCC}}$ is again a scaling parameter. $\tau_{\text{Ca,decay}} = 30$ ms is estimated from (Nevian and Sakmann, 2006). $\tau_{\text{Ca,rise}} = 2$ ms is used mainly to make $[\text{Ca}^{2+}]$ continuous.

### 2.4.2 NMDA plateau potential

The voltage of a dendrite receiving NMDAR and GABAR inputs follows

$$\begin{aligned} C_D \frac{\mathrm{d}V_D}{\mathrm{d}t} &= -g_{L,D}(V_D - E_L) - g_c(V_D - \hat{V}_S) \\ &\quad - \sum_j \tilde{g}_{\text{NMDA}} s_{\text{NMDA,j}}(t)(V_D - E_E) f_{\text{Mg}}(V_D) - \sum_k \tilde{g}_{\text{GABA}} s_{\text{GABA,k}}(t)(V_D - E_I) \end{aligned}$$

where $j$ and $k$ are indices of NMDAR and GABAR synapses respectively. Denote

$$g_{\text{NMDA}}(t) = \sum_j \tilde{g}_{\text{NMDA}} s_{\text{NMDA,j}}(t) \tag{2.15}$$

as the total NMDA input conductance onto this dendrite. The maximum value of $g_{\text{NMDA}}(t)$ is simply $g_{\text{NMDA,max}} = \sum_j \tilde{g}_{\text{NMDA}} = N_{\text{NMDA}} \tilde{g}_{\text{NMDA}}$, where $N_{\text{NMDA}}$ is the number of NMDAR synapses. Similarly

$$g_{\text{GABA}}(t) = \sum_k \tilde{g}_{\text{GABA}} s_{\text{GABA,k}}(t) \tag{2.16}$$

If we ignore the coupling between this dendrite and its soma for now, and consider constant synaptic conductances $g_{\text{NMDA}} = g_{\text{NMDA}}(t)$, $g_{\text{GABA}} = g_{\text{GABA}}(t)$. Then we have

$$C_D \frac{\mathrm{d}V_D}{\mathrm{d}t} = -g_{L,D}(V_D - E_L) - g_{\text{NMDA}}(V_D - E_E) f_{\text{Mg}}(V_D) - g_{\text{GABA}}(V_D - E_I) \tag{2.17}$$

Since we have $E_I = E_L$, the steady-state dendritic voltage $V_{D,\mathrm{ss}}$ satisfies

$$0 = -(V_{D,\mathrm{ss}} - E_L) - \frac{g_{\mathrm{NMDA}}}{g_{L,D} + g_{\mathrm{GABA}}}(V_{D,\mathrm{ss}} - E_E) f_{\mathrm{Mg}}(V_D) \tag{2.18}$$

This equation can be solved numerically, resulting in the curve in **Fig. 2.4d**.

### 2.4.3 Pathway-specific gating in single pyramidal neuron

Gating is performed by disinhibiting a specific subset of dendrites. Disinhibited dendrites always receive 5 Hz background inhibition. The disinhibition level is defined as the difference between the rates of inhibition received by inhibited and disinhibited dendrites.

In **Fig. 2.5**, each pathway targets two dendrites with 15 NMDAR synapses on each dendrite. The dendrites targeted by each pathway do not overlap. For each pathway, the input rate ($u_E$) follows a bell-shaped tuning to the stimulus value ($z$): $u_E = 40\exp(-z^2)$ Hz, where $z$ ranges between $-2.4$ and 2.4. The disinhibition level is 30 Hz (from 35 Hz to 5 Hz).

Presented alone, the preferred stimulus ($z = 0$) from one pathway increases the output firing rate by $r_{\mathrm{on}}$ ($r_{\mathrm{off}}$) when the pathway is gated on (off). The gating selectivity is defined as

$$\frac{r_{\mathrm{on}} - r_{\mathrm{off}}}{r_{\mathrm{on}} + r_{\mathrm{off}}}, \tag{2.19}$$

For **Fig. 2.9**, excitatory pathways can overlap. In the context with gate 1 open, $N_{\mathrm{disinh}}$ dendrites are disinhibited. Excitatory pathway 1 targets these $N_{\mathrm{disinh}}$ dendrites, each with

strength 25 nS, and similarly for gate 2 and pathway 2. The $N_{\text{disinh}}$ dendrites disinhibited for gate 2 are chosen randomly and independently from the $N_{\text{disinh}}$ dendrites disinhibited for gate 1. For each $N_{\text{disinh}}$ and $N_{\text{dend}}$, $r_{\text{on}}, r_{\text{off}}$ are averaged across all possible projection patterns.

### 2.4.4 Rate pyramidal neuron model

The rate model is fitted with simulation data from the spiking model with *in-vivo* parameters (**Fig. 2.7**). The time-averaged voltage of a dendritic compartment ($\overline{V}_D$) is modeled as a sigmoidal function of total excitatory input conductance ($\overline{g}_E$, see below for definition) following:

$$\overline{V}_D = f_V(\overline{g}_E, \overline{g}_I) = 30 \cdot \left[1 + \tanh\left(\frac{\overline{g}_E - g_{1/2}}{\beta}\right)\right] + V_0 + E_L \tag{2.20}$$

The mid-point $g_{1/2}$ is proportional to the total inhibitory conductance $\overline{g}_I$ plus the leak conductance of the dendrite $g_{L,D}$, as expected from the constant conductance scenario (**Fig. 2.4c**)

$$g_{1/2} = b_g \cdot \left(g_{L,D} + \overline{g}_I\right) \tag{2.21}$$

Based on our observation of the reduced spiking model, we modeled the width $\beta$ as an exponentially increasing function of inhibition:

$$\beta = k \cdot \exp\left(\overline{g}_I / \gamma\right) \tag{2.22}$$

This increase of width $\beta$ as a function of $\overline{g}_I$ captures the linearization effect of sparse

inhibition on the voltage-input function (**Fig. 2.2c**). Fit values of the parameters are $b_g$ = 5.56, $k$ = 9.64 nS, $\gamma$ = 6.54 nS, $V_0$ = 0.78 mV. The model is fitted to a simulated 10-dendrite spiking neuron model. Since we hope to build a rate model for the multi-compartmental neuron and we want dendrites to feed forward into somas in this model, it is necessary to fix the somatic activity at a reasonable level when we calibrate the dendrite model, such that the dendrite activity will no longer depend on somatic activity in our model. So when simulating dendrites of the spiking model, somatic shadow voltage is clamped at -60 mV, and back-propagating action potential is fixed as a Poisson spike train of 10 Hz.This phenomenological model allows us to interpolate the dendritic voltage for a large range of excitatory and inhibitory inputs very rapidly.

The firing rate of the soma is modeled as a power law function of input current $I$:

$$r = f_r(I) = [\max(0, I + 174.86)/45.16]^{2.89} \tag{2.23}$$

Here $I$ is the sum of the input current from dendrites and also the somatic inhibition from PV neurons whenever applicable. The parameters are fitted from simulation of the reduced spiking model. We assume the somatic voltage fluctuates around $E_{\text{reset}}$, and denote the mean dendritic voltage $\langle V_D \rangle$. Then the input current from dendrites is $I_{\text{dend}\to\text{soma}} = G_c \cdot (\langle V_D \rangle - E_{\text{reset}})$, where $G_c$ is the total dendrite-soma coupling of all dendrites. $G_c$ = 8 nS. Since we assume $G_c$ is fixed whenever we vary the number of dendrites, the somatic function does not depend on the number of dendrites and need not be re-parametrized. So $I = I_{\text{dend}\to\text{soma}} + \Delta I_{\text{PV}\to\text{soma}}$, where $\Delta I_{\text{PV}\to\text{soma}}$ is the change in somatic inhibition from PV neurons.

For inputs to the rate model, $\overline{g}_E$ and $\overline{g}_I$ are the time-averaged total conductance of all

excitatory and inhibitory synapses, respectively. For NMDAR-only excitatory input, the approximated time-averaged gating variable $\overline{s}_{\mathrm{NMDA}}$ of a single synapse receiving input rate $r_E$ follows,

$$\overline{s}_{\mathrm{NMDA}} = 1 - (1 + r_E \tau_{\mathrm{NMDA,rise}} \tau_{\mathrm{NMDA,decay}} \alpha_{\mathrm{NMDA}})^{-1} \tag{2.24}$$

For $N_{\mathrm{NMDA}}$ synapses each with maximal conductance $\tilde{g}_{\mathrm{NMDA}}$, the total excitatory conductance is

$$\overline{g}_E = N_{\mathrm{NMDA}} \overline{s}_{\mathrm{NMDA}} \tilde{g}_{\mathrm{NMDA}} \tag{2.25}$$

Therefore, $\overline{g}_E$ saturates as $\overline{s}_{\mathrm{NMDA}}$ does. Because the GABAR conductance is linear in its input rate, the total inhibitory conductance is

$$\overline{g}_I = r_I \tau_{\mathrm{GABA}} \tilde{g}_{\mathrm{GABA}} \tag{2.26}$$

where $r_I$ is the overall inhibitory input rate onto that dendrite.

### 2.4.5 Interneuron Models

SOM neurons are modeled as simple rate neurons with a rectified linear f-I curve. The firing rate of a SOM neuron is

$$r_{\mathrm{SOM}} = \max(0, \beta_{\mathrm{SOM}}(I_{\mathrm{syn,SOM}} - I_{\mathrm{rh,SOM}})) \tag{2.27}$$

where $\max(x,0)$ is a rectified linear function of $x$. $I_{\text{rh,SOM}} = 40$ pA is the rheobase, i.e. the minimum current required to activate the neuron, and $\beta_{\text{SOM}} = 90$ Hz/nA is the f-I curve slope for SOM neurons, which we matched to data from (Lu et al., 2007). SOM neurons typically display adapting responses to constant input, and the synapses of SOM neurons show short-term-plasticity. We ignored these aspects of temporal dynamics because here we are interested in the steady-state response. SOM neurons receive 150 pA input current in the default state, leading to a baseline firing of SOM neurons around 10 Hz as observed experimentally (Gentet et al., 2012; Lee et al., 2013).

For VIP neurons, we assume that the control input targets $N_{\text{control,VIP}} = \text{round}(N_{\text{VIP}} \cdot P_{\text{control,VIP}})$ of them. On average VIP neurons are assumed to fire at $\overline{r}_{\text{VIP}} = 5$ Hz. Therefore the VIP neurons non-activated by the control input fire at 0 Hz, while those targeted by the control input fire at $(5 \cdot N_{\text{VIP}} / N_{\text{control,VIP}})$ Hz.

PV neurons are modeled simply as linear rate neurons with a f-I curve slope of $\beta_{\text{PV}} = 220$ Hz/nA, because their activities never reach zero in our model. Since we are only interested in their change of activities in response to SOM neuron suppression, the spontaneous activity of PV neurons is irrelevant to our model.

### 2.4.6 Interneuronal Network

The full interneuronal network model contains pyramidal, SOM, VIP, and PV neurons. The network model is roughly based on a L2/3 cortical column microcircuit, and contains 3000 pyramidal neurons, 160 SOM neurons, 140 VIP neurons, and 200 PV neurons (Lee et al., 2010). However, the analysis applies more generally. Pyramidal neurons are modeled as multi-compartmental rate neurons as described above. We typically used $N_{\text{dend}} = 30$ dendrites, approximately corresponding to pyramidal neurons in associative

areas (Larkman, 1991). The number of pyramidal neurons does not affect our results.

The SOM-to-dendrite connections are set randomly. Instead of drawing each connection randomly and independently with a fixed probability, we assume that each dendrite is targeted by precisely $N_{\text{SOM}\to\text{dend}}$ SOM neurons, when $N_{\text{SOM}\to\text{dend}}$ is an integer, so that each dendrite receives the same amount of net inhibition in the default state. The identities of SOM neurons targeting each dendrite is randomly chosen. The total inhibitory conductance received by each dendrite is denoted and fixed as $G_{\text{SOM}\to\text{dend}} = 40$ nS, then for each SOM-dendrite connection the conductance is $G_{\text{SOM}\to\text{dend}}/N_{\text{SOM}\to\text{dend}}$. Each SOM-dendrite connection can contain multiple synapses, which we are not explicitly modeling here because GABAergic synapses are linear such that we only need to consider the total conductance. When $N_{\text{SOM}\to\text{dend}}$ is not an integer, we interpolate. Each dendrite is targeted by $\lceil N_{\text{SOM}\to\text{dend}} \rceil$ SOM neurons, where all synapses but one have weights $G_{\text{SOM}\to\text{dend}}/N_{\text{SOM}\to\text{dend}}$, while one has weight $G_{\text{SOM}\to\text{dend}} \cdot (1 - \lfloor N_{\text{SOM}\to\text{dend}} \rfloor / N_{\text{SOM}\to\text{dend}})$. Given the connection probability from SOM to pyramidal neurons $P_{\text{SOM}\to\text{pyr}}$, the number of SOM neurons $N_{\text{SOM}}$, and the number of dendrites on each pyramidal neuron $N_{\text{dend}}$, we set

$$N_{\text{SOM}\to\text{dend}} = N_{\text{SOM}} \cdot [1 - (1 - P_{\text{SOM}\to\text{pyr}})^{N_{\text{dend}}}] \tag{2.28}$$

This is the mean number of SOM neurons targeting each dendrite if the SOM-to-pyramidal connections were completely independent and random.

The VIP-to-SOM connections are set in the same way as the SOM-to-dendrite connections. Since SOM neurons only have one compartment each, we have $N_{\text{VIP}\to\text{SOM}} = N_{\text{VIP}} \cdot P_{\text{VIP}\to\text{SOM}}$. When control inputs target both VIP and SOM neurons, we have $P_{\text{VIP}\to\text{SOM}} =$

0.6. When control inputs only target VIP neurons, we have $P_{\text{VIP}\to\text{SOM}} = 0.1$. Within $100\mu$m the connection probability is measured to be around 0.6 (**Table 2.1**). However, note that the connection probability from VIP to SOM neurons on a column scale is unknown. The spatially-restricted axonal arbors of VIP neurons (Bayraktar et al., 2000) suggest that the connection probability may fall quickly as a function of the VIP-SOM distance. Therefore on the scale of a column, the connection probability could still be as low as 0.1. Total inhibitory connection weight from VIP neurons received by each SOM neuron is 30 pA/Hz, and is distributed onto each synapse the same way SOM-to-dendrite connection weights are set. For $N_{\text{VIP}} = 140$ and $P_{\text{VIP}\to\text{SOM}} = 0.6$, the connection strength of each synapse is about 0.4 pA/Hz. This is close to the unitary VIP-to-SOM IPSQ of 0.7 pA/Hz measured in (Pfeffer et al., 2013). Notice here the connection is current-based because SOM neurons are described by a f-I curve.

The SOM-to-PV, PV-to-PV, and PV-to-pyramidal soma connections are all set similar to the connections above. We set $P_{\text{SOM}\to\text{PV}} = 0.8$, $P_{\text{PV}\to\text{PV}} = 0.9$, $P_{\text{PV}\to\text{soma}} = 0.6$ (Pfeffer et al., 2013). The total inhibitory connection strength from SOM neurons to each PV neuron is varied in **Fig. 2.15b**. The total inhibitory connection from PV neurons to each PV neuron is 30 pA/Hz, and from PV neurons to each pyramidal neuron is 30 pA/Hz. Denote the resulting connection weight matrix $W_{\text{SOM}\to\text{PV}}$, $W_{\text{PV}\to\text{PV}}$, $W_{\text{PV}\to\text{soma}}$, then in steady state the change in somatic inhibition $\Delta\mathbf{I}_{\text{pyr}}$ across pyramidal neurons is

$$\Delta\mathbf{I}_{\text{soma}} = W_{\text{PV}\to\text{soma}}\Delta\mathbf{r}_{\text{PV}} \tag{2.29}$$

where $\Delta\mathbf{r}_{\text{PV}}$ is the change in PV activities. And we have

$$\Delta\mathbf{r}_{\text{PV}} = (\mathbf{1}/\beta_{\text{PV}} - W_{\text{PV}\to\text{PV}})^{-1} W_{\text{SOM}\to\text{PV}}\Delta\mathbf{r}_{\text{SOM}} \tag{2.30}$$

where $\Delta \mathbf{r}_{\text{SOM}}$ is the change in SOM activities before and after control inputs. $\mathbf{1}$ is identity matrix. The precise values of these PV-related parameters do not matter.

Control inputs are excitatory. Here we are agnostic about their origin. They could be locally generated or from long-range projections. Control inputs can target subsets of SOM and VIP neurons. The mean strength of the control inputs across the whole population is always kept as a constant. When control inputs target SOM neurons, $N_{\text{control,SOM}} = \text{round}(N_{\text{SOM}} \cdot P_{\text{control,SOM}})$ of SOM neurons are targeted, with current $75 \cdot N_{\text{SOM}} / N_{\text{control,SOM}}$ pA. Therefore across the whole population the averaged current input is 75 pA. When control inputs target VIP neurons, each of the $N_{\text{control,VIP}}$ targeted VIP neurons fire at $(5 \cdot N_{\text{VIP}} / N_{\text{control,VIP}})$ Hz. For **Fig. 2.12** when control inputs only target VIP neurons, we set $P_{\text{control,SOM}} = 0, P_{\text{control,VIP}} = 0.1$. $P_{\text{control,VIP}}$ has to be low otherwise the gating selectivity would be very close to 0. For **Fig. 2.14**, when control inputs target both SOM and VIP neurons, $P_{\text{control,SOM}} = 0.5, P_{\text{control,VIP}} = 0.5$. Setting $P_{\text{control,SOM}} = 0.5$ does not result in the highest gating selectivity. We did not make particular efforts to fine-tune these parameters.

Finally, excitatory inputs carrying stimulus information for one pathway target the corresponding disinhibited dendrites. When we opened the gate for pathway 1, suppose one dendrite receives averaged inhibitory conductance $\overline{g}_I$. Then the total excitatory conductance $\overline{g}_E$ from pathway 1 onto this dendrite is

$$\overline{g}_E = \begin{cases} (1 - \overline{g}_I / g_{I,\text{th}}) \cdot g_{E,\max} & , \overline{g}_I < g_{I,\text{th}} \\ 0 & , \overline{g}_I \geq g_{I,\text{th}} \end{cases} \tag{2.31}$$

$g_{I,\text{th}}$ is a inhibitory conductance threshold we defined. Therefore when inhibition is weak (disinhibition is strong), excitation is inversely proportional to inhibition level. How-

ever, when disinhibition is weak, there will be no excitatory input at all. Having a cut-off threshold $g_{I,\text{th}}$ prevents excitatory inputs from targeting every dendrite and therefore being overly dense. We set $g_{I,\text{th}}$ = 4.0 nS. Since we have set the sum of conductances of all inhibitory synapses to be $G_{\text{SOM}\to\text{dend}}$ = 40 nS, each SOM neuron fires around 10 Hz prior to disinhibition, and $\tau_{\text{GABA,dend}}$ = 20 ms, the time-averaged conductance received by each dendrite in default is $\overline{g}_I = r_I \tau_{\text{GABA,dend}} G_{\text{SOM}\to\text{dend}}$ = 8.0 nS. Therefore by setting $g_{I,\text{th}}$ = 4.0 nS, excitatory synapses only target dendrites that are at least disinhibited by half. We set the maximum time-averaged excitatory conductance targeting each dendrite to be $g_{E,\text{max}}$ = 25 nS. This value is chosen so that excitation is strong enough to excite a disinhibited dendrite, but not strong enough to excite a strongly inhibited dendrite **Fig. 2.7**.

In **Table 2.1**, we summarized the raw experimental data used to constrain the model.

### 2.4.7 Synaptic plasticity model and learning protocol

The synaptic plasticity model is calcium-based. The calcium dynamics is described above, and the synaptic weight change given these calcium dynamics is modeled with the formalism from (Graupner and Brunel, 2012), restated below for clarity with slightly modified notations.

Over the time of stimulation, the calcium trace spends time $\alpha_p$ above the potentiation threshold $\theta_p$, and time $\alpha_d$ above the depression threshold $\theta_d$. Then the average potentiation is $\Gamma_p = \gamma_p \alpha_p$, and the average depression is $\Gamma_d = \gamma_d \alpha_d$, where $\gamma_p$ and $\gamma_d$ are the potentiation and depression strengthes respectively. In this model, the synapse is assumed to be bistable (DOWN or UP states), so we denote $\rho$ as the probability of the synapse staying in the UP state, which evolves over time in response to the calcium trace

| Parameter | Value | References | Layer | Area | Animal |
|---|---|---|---|---|---|
| Proportion of SOM neurons among inhibitory neurons | 0.208 | Lee et al. (2010) | L2/3 | S1 | mouse |
| Proportion of VIP neurons among inhibitory neurons | 0.2 | Lee et al. (2010) | L2/3 | S1 | mouse |
| Total number of inhibitory neurons in a column | 676 ± 116 | Meyer et al. (2011) | L2/3 | S1 | rat |
| Baseline activity of SOM neurons | 6.2 ± 0.7 Hz | Gentet et al. (2012) | L2/3 | S1 | mouse |
| Unitary IPSQ from SOM to pyramidal neurons | 1.5 ± 0.3 pC | Pfeffer et al. (2013) | L2/3 | V1 | mouse |
| Unitary IPSQ from VIP to SOM neurons | 0.69 ± 0.33 pC | Pfeffer et al. (2013) | L2/3 | V1 | mouse |
| Connection probability from SOM to pyramidal neurons (within 200$\mu$m) | 0.71 ± 0.03 | Fino and Yuste (2011) | L2/3 | frontal cortex | mouse |
| Connection probability from VIP to SOM neurons (within 25 – 100$\mu$m) | 0.625 ± 0.12 | Pfeffer et al. (2013) | L2/3 | V1 | mouse |
| Number of basal dendrites on each pyramidal cell (number of total tips) | 28.8 ± 2.4 | Larkman (1991) | L2/3 | V1 | rat |
| Number of basal dendrites on each pyramidal cell (maximum branches at fixed radius) | 20 ± 2.6 | Elston, Benavides-Piccione and Defelipe (2005) | L3 | V1 | monkey |
| Number of basal dendrites on each pyramidal cell (maximum branches at fixed radius) | 34.2 ± 4.9 | Elston, Benavides-Piccione and Defelipe (2005) | L3 | anterior cingulate cortex | monkey |

Table 2.1: Raw experimental data used to constrain the VIP-SOM-pyramidal disinhibitory circuit. The error estimates are also taken from the references when available. Some of the data are extracted from their figures since the value is not reported in texts. Specifically, the proportion of VIP neurons is inferred from the proportion of 5HT3a neurons among interneurons and proportion of VIP neurons among 5HT3a neurons.

crossing thresholds. Then define $\overline{\rho}$ as the long-term time average of $\rho$, and $\sigma_\rho^2$ as the

standard deviation of $\rho$. Then

$$\overline{\rho} = \frac{\Gamma_p}{\Gamma_p + \Gamma_d} \tag{2.32}$$

$$\sigma_\rho^2 = \frac{\sigma^2(\alpha_p + \alpha_d)}{\Gamma_p + \Gamma_d} \tag{2.33}$$

where $\sigma$ is the amplitude of noise and $\tau$ is the time constant of weight change. In long term, the probability to switch from DOWN to UP state $\mathscr{U}$ and from UP to DOWN states $\mathscr{D}$ are given by

$$\mathscr{U} = \frac{1}{2}\left(1 + \mathrm{erf}\left(-\frac{0.5 - \overline{\rho} + \overline{\rho}e^{-(\Gamma_p+\Gamma_d)/\tau}}{\sqrt{\sigma_\rho^2(1 - e^{-2(\Gamma_p+\Gamma_d)/\tau})}}\right)\right) \tag{2.34}$$

$$\mathscr{D} = \frac{1}{2}\left(1 - \mathrm{erf}\left(-\frac{0.5 - \overline{\rho} + (\overline{\rho} - 1)e^{-(\Gamma_p+\Gamma_d)/\tau}}{\sqrt{\sigma_\rho^2(1 - e^{-2(\Gamma_p+\Gamma_d)/\tau})}}\right)\right) \tag{2.35}$$

erf$(\cdot)$ is the standard error function. For convenience, we set the weight of DOWN state to $w_0 = 0$, and the weight of UP state $w_1 = 3$. Then following stimulation, the weight after learning $w_{\mathrm{post}} = w_{\mathrm{pre}}(1 - \mathscr{D}) + (w_1 - w_{\mathrm{pre}})\mathscr{U}$, given the weight before learning $w_{\mathrm{pre}}$.

We fitted the free parameters of the model with experimental data from (Nevian and Sakmann, 2006). In simulation of the plasticity experiment, we modeled the pre-synaptic extracellular stimulation by 40 NMDAR synapses simultaneously receiving one spike. This stimulation alone causes a 2.8 mV depolarization on the soma, which is within the range of observed values (1–3 mV) for that experiment. It also brings the dendrite close to the NMDA plateau threshold, allowing for strong interaction between pre- and post-synaptic spikes. As in the experiment, all pairings are repeated 60 times. The somatic shadow voltage is clamped at −60 mV.

The model is fitted to data points from **Fig. 2 and 3d** in (Nevian and Sakmann, 2006),

and is used to predict data from **Fig. 3b** therein. Notice that two data points in the test dataset (their **Fig. 3b**) are already included in their **Fig. 2 and 3d**. In all of these cases, there is one pre-synaptic spike, and usually a burst of post-synaptic spikes. The time lag in **Fig. 2.18a** is defined as the timing difference between the first post-synaptic spike in the burst and the pre-synaptic spike. In **Fig. 2.18a,b**, there are 3 post-synaptic spikes. In **Fig. 2.18b,c**, the pre-synaptic spike is either 10 ms earlier than the first post-synaptic spike in the burst, or 10 ms later than the last post-synaptic spike. In **Fig. 2.18a,c**, the post-synaptic burst, when there is one, has frequency of 50 Hz (inter-spike-interval of 20 ms). The fit parameters are the following. The scaling parameters for calcium traces, $\kappa_{\mathrm{NMDA}} = 0.371$ and $\kappa_{\mathrm{VGCC}} = 0.957$. The depression and potentiation rates are $\gamma_d = 39.9$ and $\gamma_p = 177.6$. The potentiation threshold for calcium is $\theta_p = 2.78$. In fitting this particular dataset, we found that there is a certain level of redundancy in parameters; the number of parameters needed to be free is less than the total number of potentially free parameters. We therefore fixed two parameters using values in (Graupner and Brunel, 2012) which were fitted to another dataset: the amplitude of the noise $\sigma = 3.35$ and the time constant of synaptic weight change $\tau = 346.36$ s. The depression threshold is $\theta_d = 1$. Before the plasticity-inducing experiment, we have $w_{\mathrm{pre}} = 1$ which corresponds to $\tilde{g}_{\mathrm{NMDA}} = 2.5$ nS for each NMDAR synapse. So after learning, the actual synaptic conductances are $\tilde{g}_{\mathrm{NMDA}} = w_{\mathrm{post}} \cdot 2.5$ nS.

Just like the spiking pyramidal neuron model, the plasticity model fitted with *in-vitro* data needs to be recalibrated to behave properly under *in-vivo*-like conditions (Higgins, Graupner and Brunel, 2014). We reduced the scaling parameters for calcium traces to $\hat{\kappa}_{\mathrm{NMDA}} = 0.75\kappa_{\mathrm{NMDA}}$, mimicking a reduced extracellular calcium concentration, and to $\hat{\kappa}_{\mathrm{VGCC}} = 0.2\kappa_{\mathrm{VGCC}}$, resembling attenuated effect of back-propagating action potentials in

high-conductance *in-vivo* state. These changes also ensure the weights of non-activated synapses do not change substantially throughout the simulation. In **Fig. 2.17b**, the plasticity inducing protocol is 300-s long. The post-synaptic firing is fixed at 10 Hz.

In **Fig. 2.20**, inputs from both pathways initially target every dendrite with 15 synapses of the same weight $\tilde{g}_{\mathrm{NMDA}} = 2.5$ nS. Each gate is opened by disinhibiting 2 distinct dendritic branches. During learning, all synapses from the gated-on pathway are activated at 50 Hz, whereas the gated-off pathway is not activated. The post-synaptic rate is set at 10 Hz. To measure gating selectivity before learning, 8 of the 15 synapses on each dendrite are activated for both pathways. After learning 5 of 15 synapses were activated, the number is chosen so that before and after learning the total excitatory conductance is the same.

### 2.4.8 Context-dependent decision-making network

We modeled the context-dependent decision-making task from (Mante et al., 2013). In the experimental task, the stimulus is a mixture of random dots that are leftward- or rightward-moving and are red or green. The stimulus can be described by its motion and color coherence. Motion coherence for rightward motion can take 6 values ($-0.5$, $-0.15$, $-0.05$, 0.05, 0.15, 0.5). Color coherence for color red also takes 6 values ($-0.5$, $-0.18$, $-0.06$, 0.06, 0.18, 0.5). On each trial, the color and motion coherence are independently and randomly chosen, resulting in 36 possible stimuli. In **Fig. 2.22a**, the performance with respect to motion coherence is averaged across all color coherences. Similarly for **Fig. 2.22b**, the performance with respect to color coherence is averaged across all motion coherences. In **Fig. 2.22c**, the curve for strong motion coherence is averaged across motion coherence $-0.5$ and 0.5. Similarly for medium and weak coherences.

The context-dependent decision-making circuit model contains two components. The first is a mixed-selective sensory network, which uses the VIP-SOM-pyramidal neuron circuit model described above. The mixed-selective sensory neurons receive motion and color inputs from the sensory stimulus. Here the motion and color inputs do not signal the actual motion and color of the stimulus, but rather the motion and color evidence for a particular target. For convenience, the motion direction corresponding to target 1 is denoted left, and the color corresponding to target 1 is denoted red, and similar for motion right and color green. This treatment follows the analyses and modeling of (Mante et al., 2013). There are four pools of neurons in this network. Each pool prefers a particular combination of motion and color, e.g. left and red. Each neuron pool is modeled exactly as those in **Fig. 2.14a**, where the circuit connectivity is random and control inputs target both VIP and SOM neurons, using the base parameter set described above. The input to each dendrites is 15 NMDAR synapses with rate determined by the coherence (coh) of their preferred motion and color input, as $40 \cdot (1 + \text{coh})$ Hz (Wong and Wang, 2006; Britten et al., 1993). For example, a neuron that prefers left and red inputs would receive $40 \cdot (1 + \text{coh}_{\text{left}})$ Hz input on its dendrites targeted by motion pathway, and $40 \cdot (1 + \text{coh}_{\text{red}})$ Hz on its dendrites targeted by color pathway. Note that $\text{coh}_{\text{left}} = -\text{coh}_{\text{right}}$. The excitatory input for each pathway is set the same way as it is above, however now the maximum conductance of these input synapses $g_{\text{sen}}$ is one free parameter.

The second component of the network is a decision network. This network is a two-pool rate model (Wong and Wang, 2006), using the parameter set therein with no recurrent AMPAR current. The pool representing choice 1 receives input from a subset of the left-red neuron pool in the mixed-sensory network. Sensory neurons are sorted according to their gating selectivity, and only the top $P_{\text{project}}$ fraction of these sensory neurons

project to the decision networks. $P_{\text{project}}$ is also a free parameter. The right-green pool projects to the choice 2 pool. The other two pools do not project to the decision network because only the left-red and the right-green pools have congruent preferences for choice 1 and choice 2, respectively, based on the how color and motion evidence are defined. In order to fit experimental behavioral choice data efficiently, we further approximated the decision network with a decision function. We assumed that the probability of selecting choice 1 ($P_1$) is determined by the difference $\Delta I_{\text{dec}}$ (pA) between the input currents to the two choice pools. We fitted this function by simulating the decision network with mean input current 15.6 pA to both pools, yielding

$$P_1 = \left[1 + \exp\left(\frac{-\Delta I_{\text{dec}}}{\sigma}\right)\right]^{-1} \tag{2.36}$$

with $\sigma = 0.99$ pA. The second free parameter of the model is the projection strength $J_{\text{dec}}$ of the mixed-sensory input, such that $\Delta I_{\text{dec}} = J_{\text{dec}}(r_{\text{left,red}} - r_{\text{right,green}})$. $r_{\text{left,red}}$ is the average firing rate of the left/red-preferring pool.

The three free parameters $g_{\text{sen}}, P_{\text{project}}, J_{\text{dec}}$ are fitted to behavioral data of each monkey in (Mante et al., 2013). The fit parameter values are $g_{\text{sen}} = 1.21$ nS, $P_{\text{project}} = 0.36$ and $J_{\text{dec}} = 15.0$ pA/Hz for monkey F, and are $g_{\text{sen}} = 1.80$ nS, $P_{\text{project}} = 0.083$ and $J_{\text{dec}} = 4.37$ pA/Hz for monkey A. Importantly, the data used to fit the model is far from being sufficient. Also our circuit model is simplistic. Therefore these fitted parameter values do not reflect our estimates of these quantities in the brain. Rather, these fittings demonstrate that the proposed circuit architecture can potentially capture behavioral performance. As shown in **Fig. 2.23**, if neural gating is strongly degraded, then no set of these fit parameters can capture behavioral performance.

### 2.4.9 Model fitting in general

Model parameters are fitted to experimental or simulation data in various contexts. These fitted models include the rate pyramidal neuron, the calcium-based plasticity model, and the context-dependent decision-making network. In all these cases, parameters are chosen to minimize the squared-error between the model and data using sequential least squares programming (SLSQP) method from the SciPy library (scipy.optimize.minimize, with method 'SLSQP').

## 2.5 Appendix

### 2.5.1 Gating selectivity critically depends on $N_{\text{SOM}\rightarrow\text{dend}}$

The gating selectivity is defined as the mean gating selectivity across neurons,

$$\text{Gating selectivity} \quad = \quad \text{E}_{\text{neuron}}\left[\frac{r_{\text{on}} - r_{\text{off}}}{r_{\text{on}} + r_{\text{off}}}\right] \tag{2.37}$$

For each neuron, the neural activity given the gated-on pathway is

$$\tilde{r}_{\text{on}} \quad = \quad f_r(\langle \overline{V}_{D,\text{on}} \rangle) \tag{2.38}$$

, where $\langle \overline{V}_{D,\text{on}} \rangle$ is the mean dendritic voltage across all the dendrites on that neuron for the gated-on pathway. Notice here for simplicity we used an input-output formulation for the somatic compartment that is slightly different from the one used in the main text (the results are the same)

$$r = f_r(\langle \overline{V}_D \rangle) = r_0 + \left(\frac{\langle \overline{V}_D \rangle - E_L}{V_r}\right)^{n_r} \tag{2.39}$$

After correcting for the baseline, we have

$$r_{\text{on}} \quad = \quad f_r(\langle \overline{V}_{D,\text{on}} \rangle) - f_r(E_L) \tag{2.40}$$

$$\quad = \quad \left(\frac{\langle \overline{V}_{D,\text{on}} \rangle - E_L}{V_r}\right)^{n_r} \tag{2.41}$$

Similarly

$$r_{\text{off}} = \left( \frac{\langle \overline{V}_{D,\text{off}} \rangle - E_L}{V_r} \right)^{n_r} \tag{2.42}$$

So

$$r_{\text{off}} / r_{\text{on}} = \left[ (\langle \overline{V}_{D,\text{off}} \rangle - E_L) / (\langle \overline{V}_{D,\text{off}} \rangle - E_L) \right]^{n_r} \tag{2.43}$$

In the limit of large number of dendrites on each pyramidal neuron, we can replace the averaged dendritic voltage with its expectation over dendrites $\mathrm{E_D}[\cdot]$.

$$\langle \overline{V}_{D,\text{on}} \rangle \approx \mathrm{E_D} \left[ \overline{V}_{D,\text{on}} \right] \tag{2.44}$$

Under this approximation, $r_{\text{on}}$ and $r_{\text{off}}$ would be the same for every neuron, therefore we have

$$\text{Gating selectivity} = \frac{r_{\text{on}} - r_{\text{off}}}{r_{\text{on}} + r_{\text{off}}} \tag{2.45}$$

$$= -1 + 2 / \left( 1 + \left[ (\mathrm{E_D}[\overline{V}_{D,\text{off}}] - E_L) / (\mathrm{E_D}[\overline{V}_{D,\text{on}}] - E_L) \right]^{n_r} \right) \tag{2.46}$$

Since dendritic voltage is determined by the total excitatory and inhibitory conduc-

tance received,

$$\begin{aligned}
\overline{V}_D &= f_V(\overline{g}_E, \overline{g}_I) && (2.47)\\
&= 30 \cdot \left[1 + \tanh\left(\frac{\overline{g}_E - g_{1/2}}{\beta}\right)\right] + V_0 + E_L && (2.48)\\
&= 30 \cdot \left[1 + \tanh\left(\frac{\overline{g}_E - b_g \cdot \left(g_{L,D} + \overline{g}_I\right)}{k \cdot \exp\left(\overline{g}_I/\gamma\right)}\right)\right] + V_0 + E_L && (2.49)
\end{aligned}$$

Remember that for each pathway, we assume that the excitatory input conductance is a deterministic function of the inhibitory conductance received when the corresponding gate is open.

$$\overline{g}_E = \begin{cases} (1 - \overline{g}_I/g_{I,\text{th}}) \cdot g_{E,\max} & , \overline{g}_I < g_{I,\text{th}} \\ 0 & , \overline{g}_I \geq g_{I,\text{th}} \end{cases} \quad (2.50)$$

Denote this rectified linear function as $\overline{g}_E = f_E(\overline{g}_I)$. For convenience consider two pathways, the inhibitory conductance for gate 1 is $\overline{g}_{I,1}$ and for gate 2 is $\overline{g}_{I,2}$. And excitatory conductance for pathway 1 and 2 are $\overline{g}_{E,1} = f_E(\overline{g}_{I,1})$ and $\overline{g}_{E,2} = f_E(\overline{g}_{I,2})$ respectively. Then

$$\mathrm{E_D}\left[\overline{V}_{D,\text{on}}\right] = 30 \cdot \left[1 + \mathrm{E_D}\left[\tanh\left(\frac{f_E(\overline{g}_{I,1}) - b_g \cdot \left(g_{L,D} + \overline{g}_{I,1}\right)}{k \cdot \exp\left(\overline{g}_{I,1}/\gamma\right)}\right)\right]\right] + V_0 + E_L \quad (2.51)$$

and

$$\mathrm{E_D}\left[\overline{V}_{D,\text{off}}\right] = 30 \cdot \left[1 + \mathrm{E_D}\left[\tanh\left(\frac{f_E(\overline{g}_{I,2}) - b_g \cdot \left(g_{L,D} + \overline{g}_{I,1}\right)}{k \cdot \exp\left(\overline{g}_{I,1}/\gamma\right)}\right)\right]\right] + V_0 + E_L \quad (2.52)$$

We assumed that each dendrite is targeted strictly by $N_{\text{SOM}\to\text{dend}}$ SOM neurons, and since we are keeping the total amount of inhibition $G_{\text{SOM}\to\text{dend}}$ received by each dendrite fixed, the time-averaged conductance of each connection is $G_{\text{SOM}\to\text{dend}}/N_{\text{SOM}\to\text{dend}}$. We also assumed that each SOM neuron gets suppressed with probability $1-p$. Then the number of non-suppressed SOM neurons targeting each dendrite $n_{\text{SOM}\to\text{dend}}$ follows a binomial distribution

$$n_{\text{SOM}\to\text{dend}} \sim \text{B}(N_{\text{SOM}\to\text{dend}}, p) \tag{2.53}$$

And

$$\overline{g}_{I,1} = G_{\text{SOM}\to\text{dend}}/N_{\text{SOM}\to\text{dend}} \cdot n_{\text{SOM}\to\text{dend}} \tag{2.54}$$

Therefore $N_{\text{SOM}\to\text{dend}}$ determines the distribution for $\overline{g}_{I,1}, \overline{g}_{I,2}, \overline{g}_{E,1}, \overline{g}_{E,2}, \text{E}_\text{D}\left[\overline{V}_{D,\text{on}}\right], \text{E}_\text{D}\left[\overline{V}_{D,\text{off}}\right]$, and finally the gating selectivity. In summary, in the limit of a large number of dendrites, we have shown that gating selectivity only depends on the parameter $N_{\text{SOM}\to\text{dend}}$.

### 2.5.2 Gating selectivity strictly improves with somatic inhibition

Denote the f-I response function of the somatic compartment as $f(\cdot)$, and assume the dendritic input current to the soma is $I_{\text{on}}$ and $I_{\text{off}}$ when the gate is open or closed respectively. Also denote the somatic inhibitory current as $I_{\text{PV}}$. For convenience, assume $I_{\text{PV}} > 0$, so the outputs of the pyramidal neuron are

$$r_{\text{on}} = f(I_{\text{on}} - I_{\text{PV}}) \tag{2.55}$$

$$r_{\text{off}} = f(I_{\text{off}} - I_{\text{PV}}) \tag{2.56}$$

respectively. We consider only the case when $r_{\text{on}}, r_{\text{off}} > 0$, which means input stimuli have a net excitatory effect. Also we have $I_{\text{PV}} < I_{\text{off}}$. Since $r_{\text{on}}, r_{\text{off}}$ are baseline corrected, we should have $f(0) = 0$. Here we derive the necessary and sufficient condition for gating selectivity

$$S = \frac{r_{\text{on}} - r_{\text{off}}}{r_{\text{on}} + r_{\text{off}}} \tag{2.57}$$

to strictly increase with $I_{\text{PV}}$.

We have

$$\begin{aligned}
& \frac{\partial S}{\partial I_{\text{PV}}} && (2.58)\\
= \; & \frac{\partial}{\partial I_{\text{PV}}}\left[\frac{r_{\text{on}} - r_{\text{off}}}{r_{\text{on}} + r_{\text{off}}}\right] && (2.59)\\
= \; & \frac{1}{(r_{\text{on}} + r_{\text{off}})^2} \cdot \left[(r_{\text{on}} + r_{\text{off}})\frac{\partial}{\partial I_{\text{PV}}}(r_{\text{on}} - r_{\text{off}}) - (r_{\text{on}} - r_{\text{off}})\frac{\partial}{\partial I_{\text{PV}}}(r_{\text{on}} + r_{\text{off}})\right] && (2.60)\\
= \; & \frac{2}{(r_{\text{on}} + r_{\text{off}})^2} \cdot \left[r_{\text{off}}\frac{\partial r_{\text{on}}}{\partial I_{\text{PV}}} - r_{\text{on}}\frac{\partial r_{\text{off}}}{\partial I_{\text{PV}}}\right] && (2.61)
\end{aligned}$$

So

$$\frac{\partial S}{\partial I_{\text{PV}}} < 0 \tag{2.62}$$

is equivalent to

$$r_{\text{off}}\frac{\partial r_{\text{on}}}{\partial I_{\text{PV}}} < r_{\text{on}}\frac{\partial r_{\text{off}}}{\partial I_{\text{PV}}} \tag{2.63}$$

In a few more steps, we can easily derive that the necessary and sufficient condition

for gating selectivity to improve with somatic inhibition is that

$$(f'(I))^2 - f(I) \cdot f''(I) > 0 \quad , \forall I > 0 \tag{2.64}$$

where $f'(I) = \frac{\mathrm{d}f(I)}{\mathrm{d}I}$.

We can easily see that for any power law function $f(I) = aI^b$,

$$\begin{aligned} (f'(I))^2 - f(I) \cdot f''(I) &= (abI^{b-1})^2 - aI^b ab(b-1)I^{b-2} && (2.65) \\ &= a^2 bI^{2b-2} && (2.66) \end{aligned}$$

is strictly larger than 0, as long as $b > 0$.

# Chapter 3

# Functional impact of cell density changes

## 3.1 Quantifying and comparing cell type distributions in the mouse brain

Cell types are the fundamental building blocks of the brain. The numbers and ratios of distinct cell types underlie the unique processing capacities of brain circuits, yet it has not been possible to quantitatively compare cell types across animals or genders due to the size and complexity of the mammalian brain. Our collaborators, Yongsoo Kim and Pavel Osten with others, produced the first quantitative whole-brain (qBrain) cell type maps in male and female mice, focusing on three major cell types of mainly inhibitory circuitsthose expressing parvalbumin (PV+), somatostatin (SST+), and vasoactive intestinal peptide (VIP+)and four of their subtypes.

Our collaborators have developed a rapid and robust method for quantitative mapping and statistical comparison of distributions of fluorescently labeled neural cell types across the entire mouse brain, which has not previously been possible. To generate a

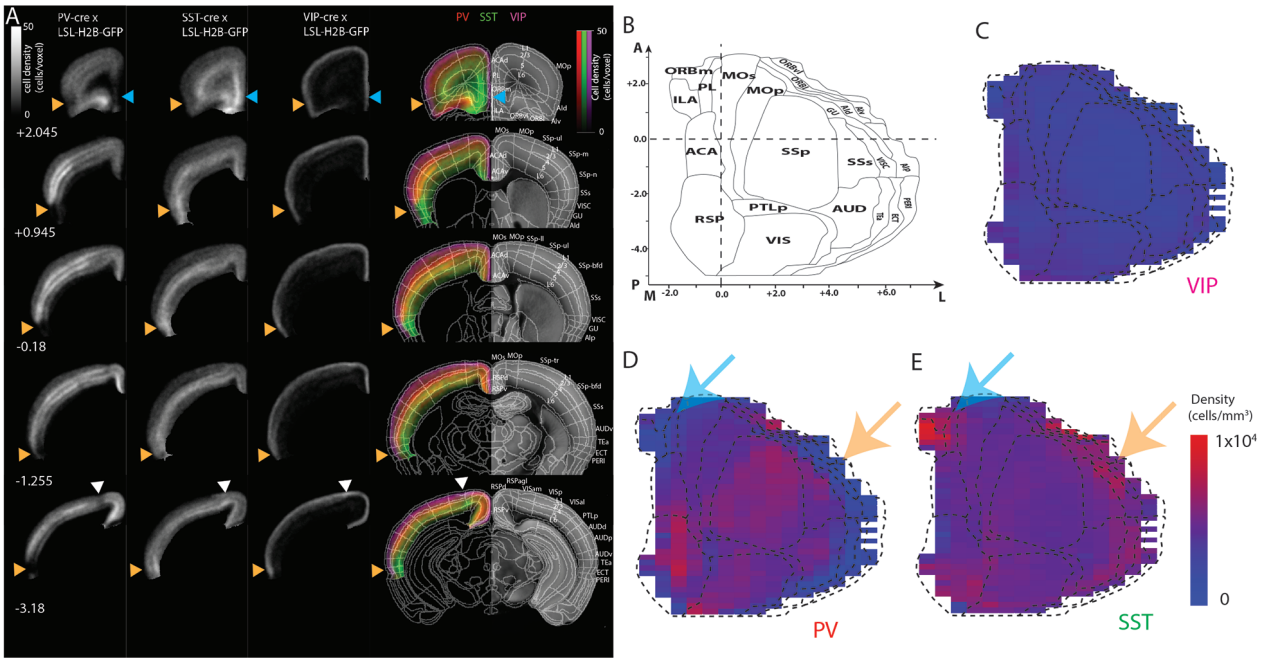


Figure 3.1: Uneven distribution of the three major interneurons in the isocortex. (**a**) Unbiased voxel-based quantitative mapping. The first three columns show the PV+, SST+, and VIP+, and the last column the combined signals overlaid on the RSTP brain with anatomical segmentation information. The heatmap represents number of cells per voxel (a sphere of 100m diameter). The A/P bregma position is shown in the first column. (**b-e**) Cell density mapping using a cortical flatmap (**b**). The heatmap displays of cell density per mm3 for the VIP+ (**c**), PV+ (**d**), and SST+ (**e**) cell populations. Note the low density of PV+ and high density of SST+ in the medial frontal (blue arrow) and lateral association cortices (orange arrow).

quantitative whole-brain (qBrain) cell type resource for the mouse and provide a proof of principle for its broad utility, they have mapped and quantified the distribution of seven cell type populations of predominantly inhibitory neurons. GABAergic inhibitory neurons comprise a great diversity of cell types that play critical roles in a broad range of brain functions and are strongly implicated in human neurologic disease. The three major cell types they mapped include neurons expressing SST, PV and VIP, which in the cortex comprise the majority of inhibitory interneurons and are thought to control the inputs, outputs and long range modulation of cortical excitatory circuits respectively (Kepecs and Fishell, 2014; Wang et al., 2004).

To map the cell densities across the whole mouse brain, our collaborators developed a cell counting and distribution mapping platform, based on automated imaging by serial two-photon tomography (STPT) and data analysis by machine learning algorithms (Kim et al., 2015; Ragan et al., 2012). They applied this platform to generate the first quantitative whole-brain maps for three major cell typesthe PV+, SST+, and VIP+ neurons, and two subtypes of the SST+ and VIP+ populations each (**Fig. 3.1**).

## 3.2 Areal hierarchy in the cortical PV+ to SST+ cell density ratios

To understand the cortical interneuron differences quantitatively, we plotted each layer 2/3 area within a two-dimensional PV+ to SST+ cell density plane, with the cell-type specific density normalized by its average across areas, and we color-coded the cortical areas following the five subnetworks described by Zingg et al. (2014): the motor-somatosensory, audio-visual, medial associational, medial prefrontal, and lateral subnetworks. This representation of the PV+/SST+ data revealed that the local PV+/SST+ cell densities are most similar within the connectivity-based cortical networks (**Fig. 3.2**).

Similar analysis of the VIP+/PV+ and VIP+/SST+ interneuron distribution revealed lesser separation (**Fig. 3.3**), suggesting that the link between local circuits and cortical networks is more specific to the balance between the output-controlling PV+ interneurons and input-controlling SST+ interneurons (note that a similar clustering is also found for PV+ and SST+ cell densities in the cortical layer 5; **Fig. 3.4**).

Finally, we propose a cortical area hierarchy Felleman and Van Essen (1991) related to the PV+/SST+ cell densities, as seen when the cortical areas are sorted according to

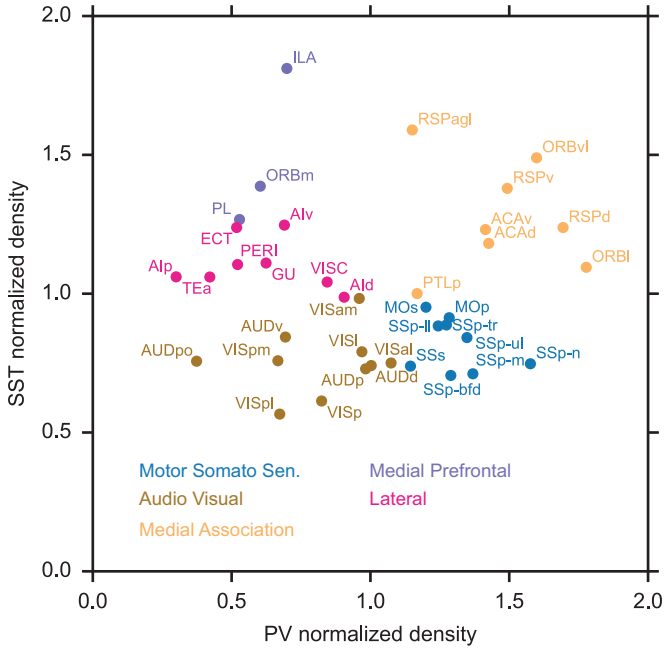


Figure 3.2: Cortical areas in the PV+/SST+ density space. Cortical areas are segregated in space of L2/3 PV+ and SST+ density according to their cortical subnetworks (color coded).

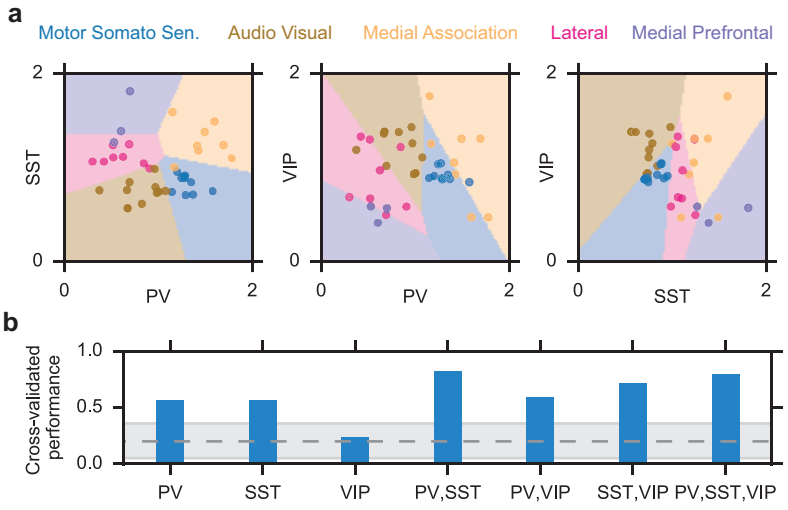


Figure 3.3: Classifying regions. (**a**) Decision boundaries of linear discriminant analysis classifiers using, from left to right, PV+/SST+, PV+/VIP+, or SST+/VIP+ cell densities. (**b**) Cross-validated classifier performances on left-one-out data, when different combinations of density information are used. Gray area indicates the 95% confidence interval of classifier performances on shuffled data.

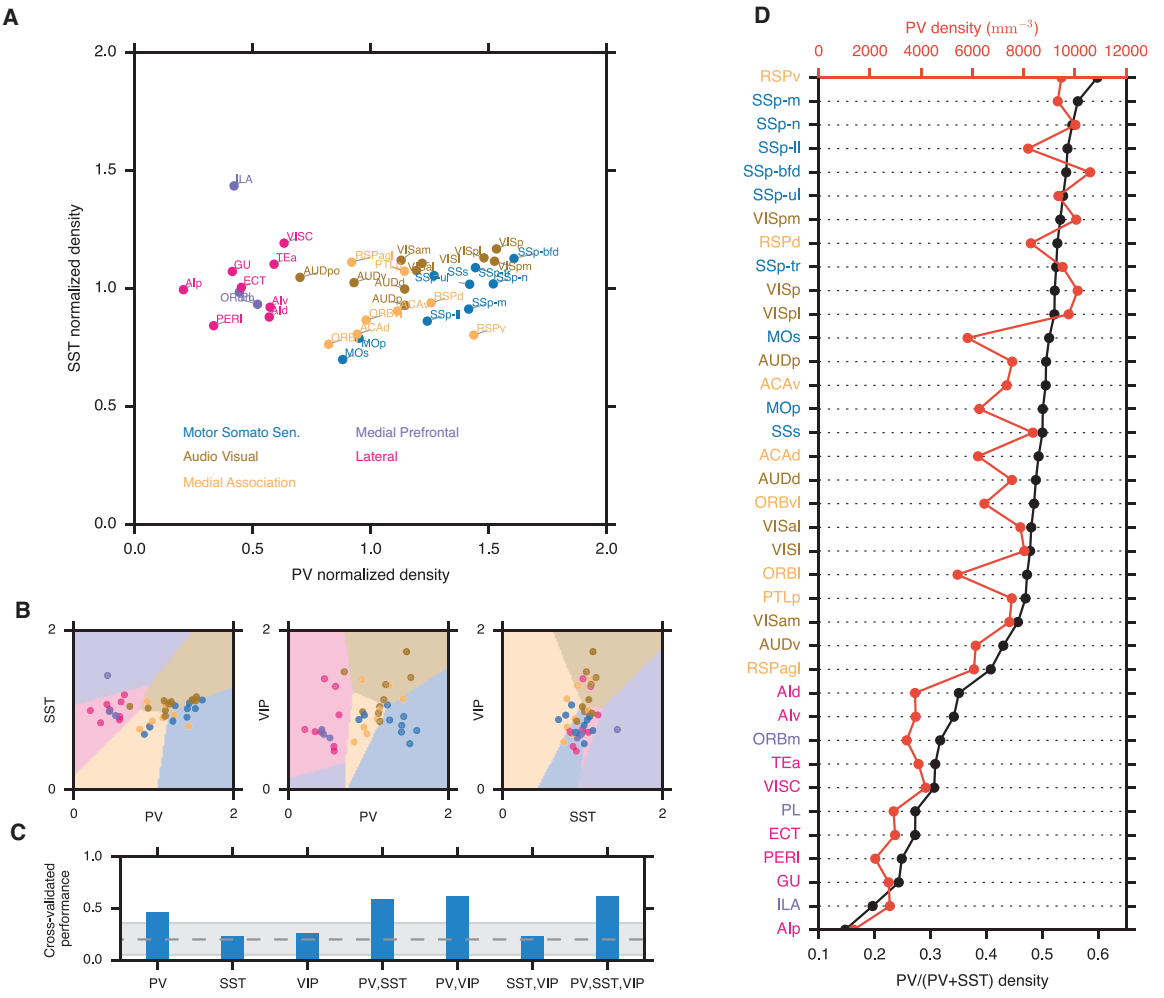


Figure 3.4: Cortical areas in L5 cell density spaces. (**a-d**) Same analyses as in **Fig. 3.2**, **Fig. 3.3**, **Fig. 3.5**, but with L5 density data. Segregation of cortical subnetworks is still present but is weaker in L5 data in comparison to L2/3 data.

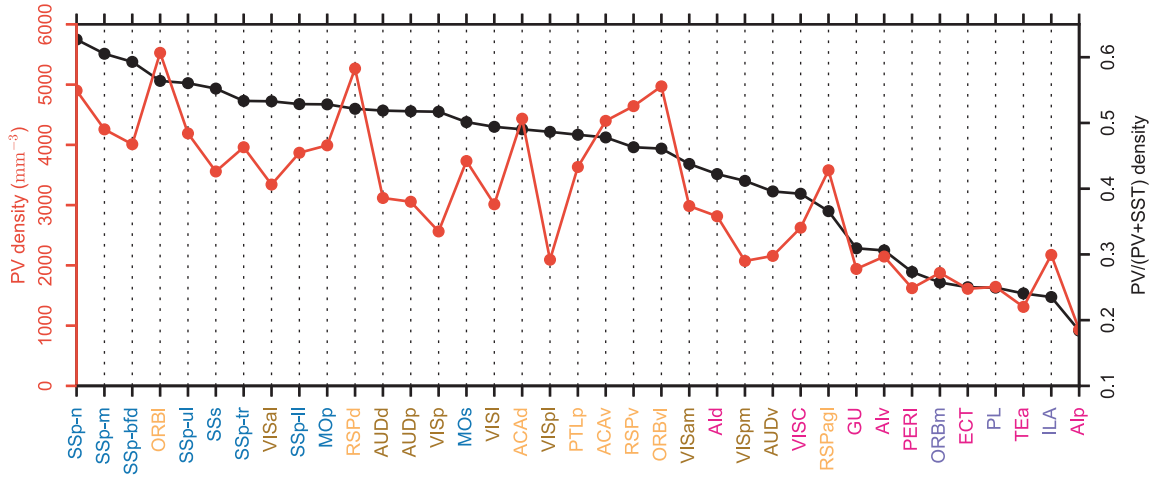


Figure 3.5: Cortical areas ranked by their PV+/SST+ cell density ratios.

their PV+/SST+ ratio: areas belonging to the motor-somatosensory subnetwork tend to have the highest PV+/SST+ density ratios, corresponding to lowest levels in the hierarchy, while the areas from the lateral and the medial frontal subnetworks occupy the highest levels in the hierarchy, with the lowest PV+/SST+ density ratios (**Fig. 3.5**).

## 3.3 Modeling the consequences of the distinct PV+ and SST+ cell densities on local cortical circuits

What may be the functional consequences of the measured cortical interneuron densities? A straightforward prediction is that an increase in the density will correspondingly strengthen the cell types role in controlling cortical excitatory neurons.

To test this prediction, we studied a simple linear rate-based circuit model consisting of excitatory (E), PV+, SST+, and VIP+ cell populations connected according to a circuit diagram measured in cortical layers 2/3 and layer 5 (**Fig. 3.6**) Pfeffer et al. (2013), with the cell density modeled as a scaling factor to all output projection weights of that population. PV+, SST+, and VIP+ neurons all receive long-range inputs from other cortical and

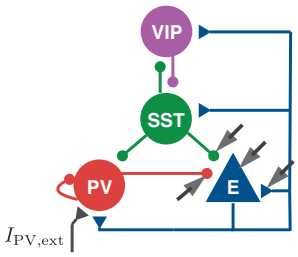


Figure 3.6: Schematic of the cortical circuit model. Cortical circuit model with an excitatory (E) and three inhibitory populations (PV, SST, and VIP) Pfeffer et al. (2013). Output weights of an inhibitory population are proportional to the density of that population.

subcortical areas Wall et al. (2016) that serve a range of important functions, including feed-forward inhibition and feedback disinhibition Kepecs and Fishell (2014). Therefore, we focused on circuit responses, i.e. changes of neural activities and synaptic currents, to external inputs driving the three interneuron cell types, starting with inputs targeting the PV+ populations.

We first compared responses in models of the somatosensory barrel field (SSp-bfd) (low SST to PV L2/3 ratio) and the infralimbic (ILA) cortex (high SST to PV L2/3 ratio) (**Fig. 3.7a**). As expected, in response to the inputs targeting the PV+ population, the steady-state response of the E population is more suppressed in the SSp-bfd, which has higher PV+ cell density than the ILA area (**Fig. 3.7b**, top trace). The circuit model also suggests that the larger suppression of E activity in the SSp-bfd is not due to stronger responses of inhibitory currents (**Fig. 3.7b**, middle traces) but, instead, due to reduced steady-state excitatory recurrent currents reflecting the reduced steady-state E activity (**Fig. 3.7b**, bottom trace).

Next, we compared responses in the posterior auditory (AUDpo) and the ventral retrosplenial (RSPv) areas, which have approximately even SST+ to PV+ ratios, but with densities of both cell types higher in the RSPv (**Fig. 3.7c**). Here, we find that the steady-state response of E population activity is similar in the two areas (**Fig. 3.7d**, top trace),

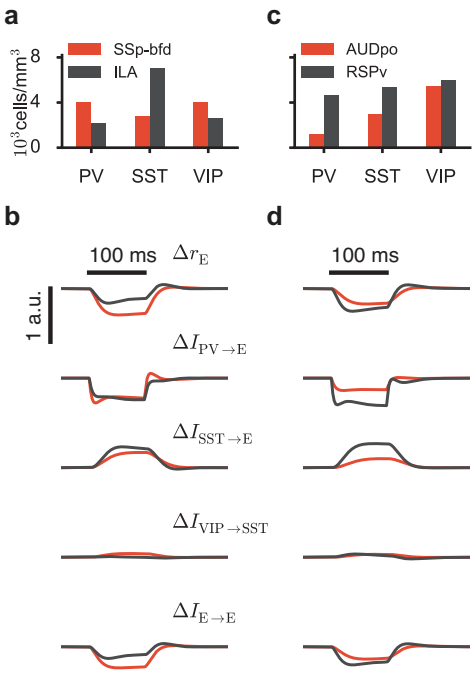


Figure 3.7: Comparison between pairs of areas. Comparing cell densities (**a**) and local circuit responses (**b**) in areas SSp-bfd and ILA. (**b**) From top to bottom: E population activity, PV-to-E, SST-to-E, VIP-to-SST, and E-to-E currents in response to external inputs driving the PV population. Spontaneous activities are kept the same across areas. (**c,d**) Same comparisons as in (**a,b**), for areas AUDpo and RSPv.

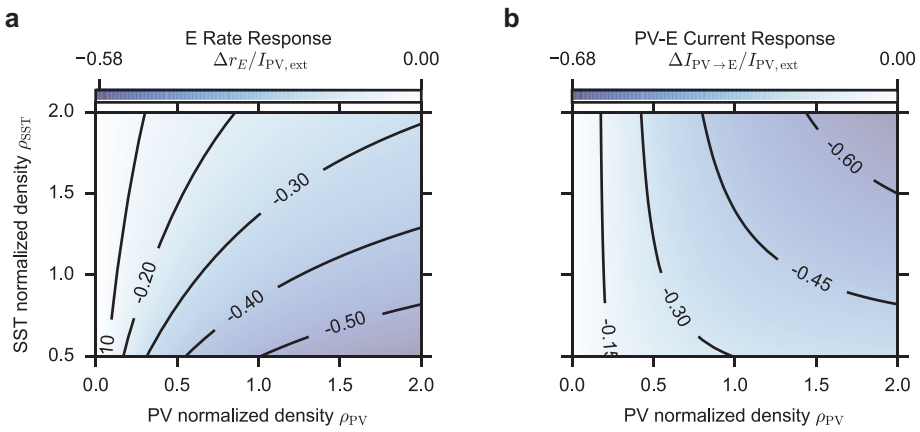


Figure 3.8: Maps of circuit responses. Responses of E population activity (**a**) and PV-to-E current (**b**) depend on both PV+ and SST+ cell densities.

though the density difference is reflected in stronger responses of the PV-to-E and SST-to-E currents in the RSPv (**Fig. 3.7d**).

Moving beyond the selected example areas, we examined the responses to external inputs targeting the PV+ population while systematically varying the PV+ and SST+ cell densities. The steady-state E activity indeed becomes more suppressed with higher PV+ cell density, but it also becomes more disinhibited with higher SST+ cell density (**Fig. 3.8a**). This suggests that both PV+ and SST+ densities are effective at altering the circuit responses when external input targets the PV+ cell population. Under the same conditions, the PV-to-E current response strengthens (becomes more negative) with both higher PV+ and higher SST+ cell density (**Fig. 3.8b**), and the SST-to-E current response also strengthens (becomes more positive/disinhibited) with higher PV+ and SST+ density (**Fig. 3.9**).

These findings are not trivial outcomes of the inhibitory connectivity or the circuit diagram (**Fig. 3.10**), as they require sufficiently strong recurrent excitatory connections (see Methods). Furthermore, these findings remain valid after addition of other putative connections in the cortical circuit, including PV-to-SST, VIP-to-PV, and VIP-to-E connec-

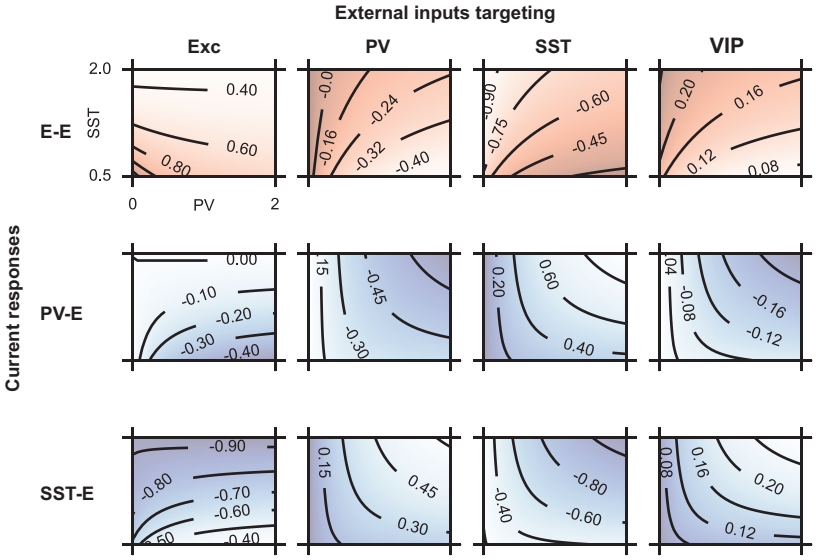


Figure 3.9: All response maps. E-E, PV-E, and SST-E current response maps (top to bottom) when external inputs target the E, PV, SST, or VIP population (left to right). The excitatory population rate response is proportional to the E-E current response.

tions Pi et al. (2013). Finally, these results were also reproduced in a neural circuit model of 5,000 realistic spiking neurons and data-constrained synaptic weights (**Fig. 3.11, 3.12**) Pfeffer et al. (2013); Wang et al. (2004); Litwin-Kumar, Rosenbaum and Doiron (2016).

Based on these results we can interpret the cell-density plots in **Fig. 3.2** in the following way. The largest areal difference defined by the PV+/SST+ ratio is an anti-correlated change in PV+ and SST+ densities, which is equivalent to traveling along the anti-diagonal direction in the PV+/SST+ density plane. Such a change strongly affects the excitatory population response, as the E activity is more suppressed in areas with higher PV+/SST+ ratio (**Fig. 3.13**). Regarding the underlying changes in the local circuit currents, the reduced E activity is reflected in reduced recurrent excitatory currents, but without changes in steady state responses of inhibitory currents. On the other hand, audio-visual and medial association subnetworks are separated along the diagonal direction in the PV+/SST+ density plane, which reflects comparable PV+/SST+ cell ratios and such a change does not affect the E steady-state activity, though it does lead to differences in steady-state

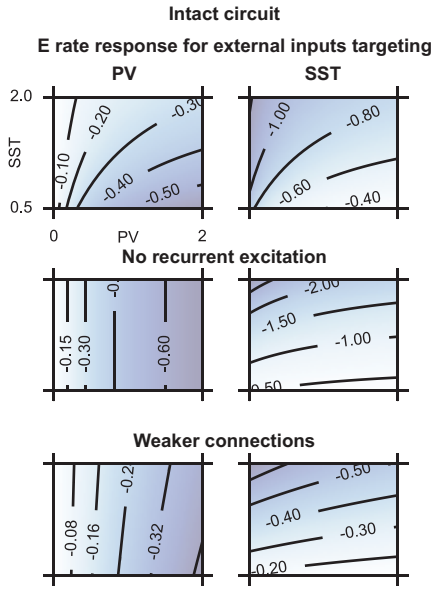


Figure 3.10: Rate response maps are non-trivial. Excitatory population rate responses to external inputs targeting PV (left) or SST (right) population in the intact local circuit (top row), after removing all recurrent excitatory connections (middle row), or after weakening all connections by 70% (bottom row).

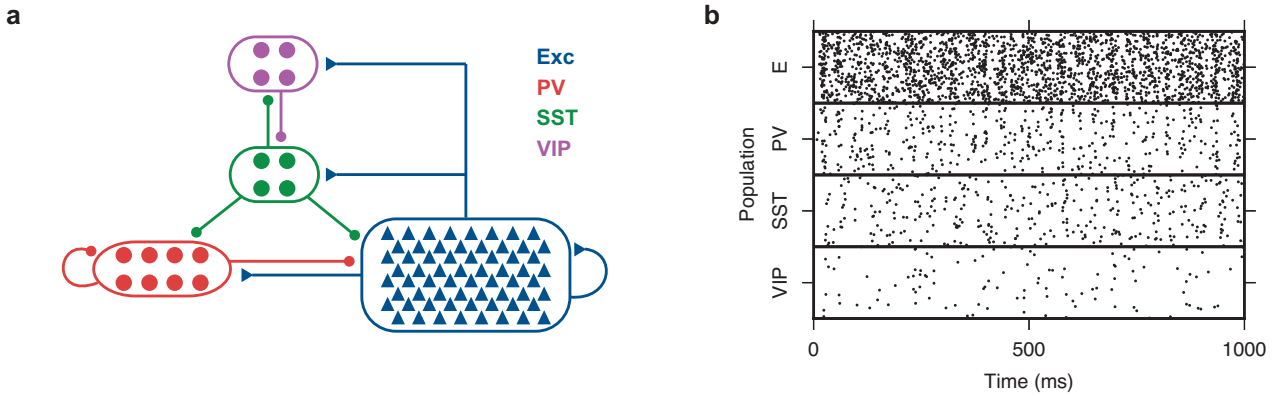


Figure 3.11: Spiking neuronal circuit model. (**a**) Spiking neural network model consisting of 5,000 neurons. (**b**) Raster plot of all neurons at the spontaneous state.

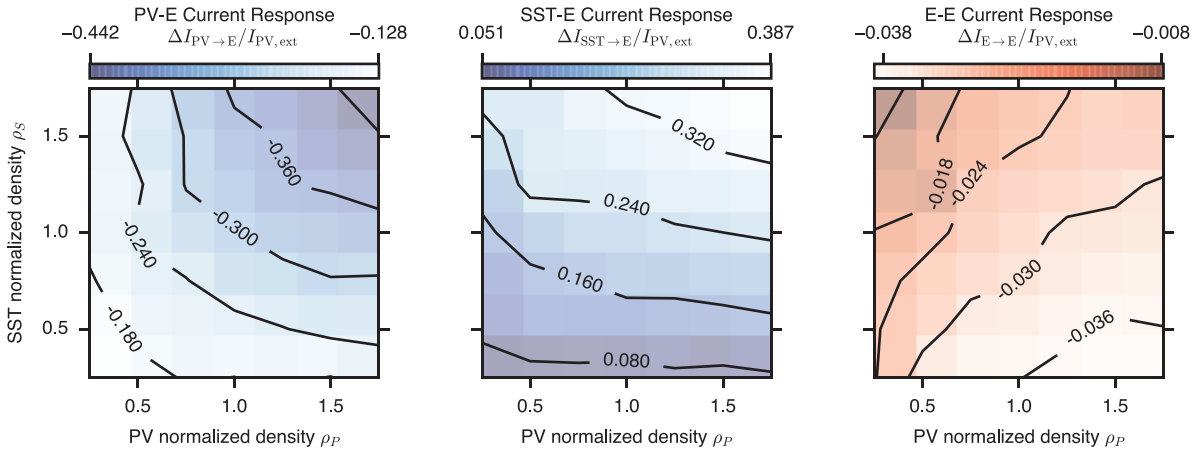


Figure 3.12: Response maps for the spiking circuit model. From left to right: the PV-to-E, SST-to-E, and E-to-E current responses to external input currents driving PV neurons.

responses of inhibitory currents (**Fig. 3.13**).

How can we explain the surprising outcome for the regions lying along the diagonal and anti-diagonal directions of the PV+/SST+ plot? Based on the circuit model (**Fig. 3.6**), inputs onto the PV+ population will suppress the E activity, which, in turn, leads to a suppression of recurrent excitation on the SST+ cells. The loss of SST+ activity then disinhibits the PV+ population and this disinhibitory loop is stronger in areas with higher SST+ density, resulting in a stronger PV-to-E current response (**Fig. 3.14**; see the Method section for more rigorous derivations).

We have also carried out similar analysis of the L2/3 circuit properties in response to external inputs targeting the other cell populations of the local excitatory-inhibitory circuit (Figure S8A). In response to inputs targeting the SST+ population, the steady-state E activity becomes more suppressed with higher SST+ cell density, but more disinhibited with higher PV+ cell density. In response to external inputs targeting the VIP+ population, the steady-state E activity becomes more disinhibited with higher SST+ cell density, and less disinhibited with higher PV+ cell density. Finally, in response to inputs onto the

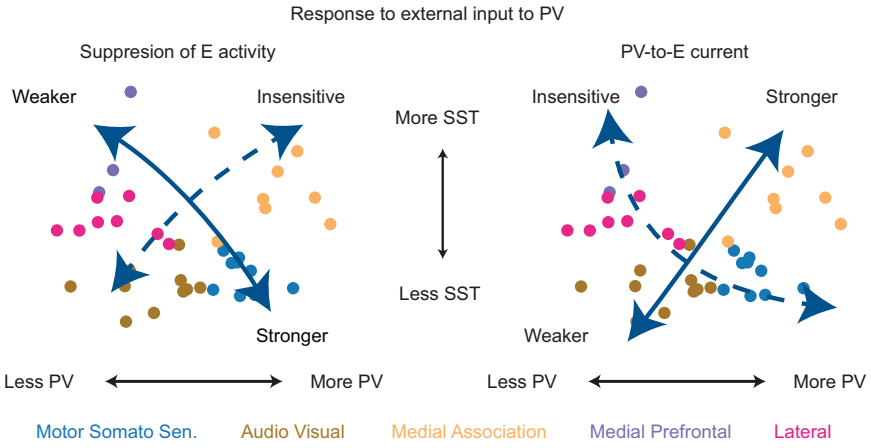


Figure 3.13: Maps of circuit responses overlaid with the distribution of cortical areas in the PV+/SST+ density plane. The diagonal direction refers to the direction between the upper right and the bottom left. The anti-diagonal direction refers to the direction between the upper left and the bottom right corner.

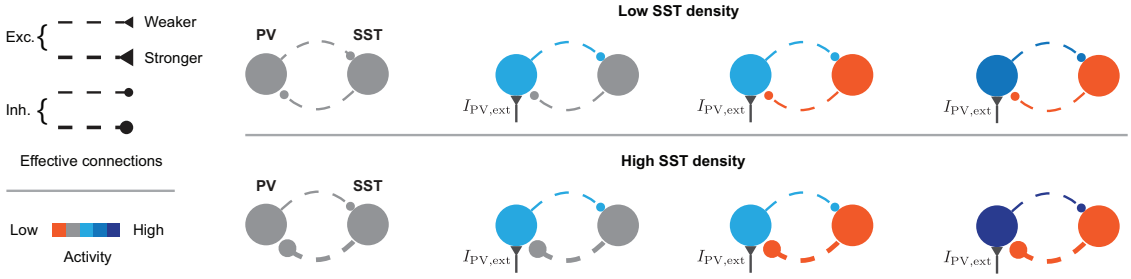


Figure 3.14: Mechanism of counterintuitive findings. Increasing SST+ density strengthens the PV-SST-PV effective disinhibitory loop, leading to a stronger PV-to-E current response.

excitatory population, the steady-state E activity as well as the E-to-E current is reduced with both higher PV+ and SST+ cell density.

In summary, the functional impact of the empirical cell densities needs to be interpreted beyond the immediate effect of each cell type, as for example the SST+ density can strongly influence how PV+ interneurons inhibit excitatory neurons. These results provide a rationale for further studies of other circuit parameters (e.g. connection probability and synaptic weights) in the lateral associational and frontal cortical areas in comparison to the sensorimotor areas (see Discussion).

## 3.4 Methods

### 3.4.1 Statistical analysis

Linear Discriminant Analysis (**Fig. 3.3**) is implemented with the Python package scikit-learn using the default parameters. Other linear classification methods, including linear support vector machine (SVM), yield similar results. The classifier performance is computed using leave-one-out cross-validation. The classifier performance is also tested on 400 instances of shuffled subnetwork labels to obtain the 95% confidence interval for classifier performance on random labeling.

### 3.4.2 Interneuronal circuit model

We studied linear dynamics of a four population rate model.

$$\tau^X \frac{\mathrm{d}r^X}{\mathrm{d}t} = -r^X + \sum_Y W^{XY} r^Y + I^X_{\mathrm{ext}}, \text{ for } X \in E, P, S, V. \tag{3.1}$$

The weight matrix **W** has the form

$$\mathbf{W} = \begin{bmatrix} W^{EE} & W^{EP} & W^{ES} & 0 \\ W^{PE} & W^{PP} & W^{PS} & 0 \\ W^{SE} & 0 & 0 & W^{SV} \\ W^{VE} & 0 & W^{VS} & 0 \end{bmatrix} \tag{3.2}$$

The connection from $Y$ to $X$, $W^{XY}$, is scaled by the normalized density of the projecting population, $W^{XY} = \rho^Y W_0^{XY}$. The connectivity value when $\rho^E = \rho^P = \rho^S = \rho^V = 1$ follows Litwin-Kumar, Rosenbaum and Doiron (2016), therefore

$$\mathbf{W} = \begin{bmatrix} 0.8 & -\rho^P & -\rho^S & 0 \\ 1 & -\rho^P & -0.5\rho^S & \\ 1 & 0 & 0 & -0.25\rho^V \\ 1 & 0 & -0.6\rho^S & 0 \end{bmatrix} \tag{3.3}$$

$\rho^E$ is omitted since they are set to equal to 1. The normalized density $\rho^X$ for each area is the absolute density $\hat{\rho^X}$ normalized by the absolute density averaged across all areas. Typically, we used cell densities from L2/3. We set $\tau^E = 20$ms, $\tau^P = 10$ms, $\tau^S = 20$ms, $\tau^V = 20$ms. These time constant parameters have no effect on the steady-state responses, however, they could change the stability of the fixed point. In this model, the steady state of the system, given it is stable, is

$$\mathbf{r_{ss}} = (\mathbf{1} - \mathbf{W})^{-1}\mathbf{I}_{\text{ext}}. \tag{3.4}$$

**1** is the identity matrix. The linear response of the steady-state activity to external inputs is

$$\frac{\partial \mathbf{r_{ss}}}{\partial \mathbf{I}_{\text{ext}}} = (\mathbf{1} - \mathbf{W})^{-1}. \tag{3.5}$$

The current from population $X$ to $Y$ is $I^{YX} = W^{YX} r^X$. The current response to external inputs targeting population $Z$, $I^Z_{\text{ext}}$, is defined as $\frac{\partial I^{YX}}{\partial I^Z_{\text{ext}}}$.

### 3.4.3 Spiking neural circuit model

Each population $X$ ($= E, P, S, V$) is modeled with $N^X$ adaptive exponential integrate-and-fire neurons, where the $i$-th neuron is described by:

$$\frac{\mathrm{d}V_i}{\mathrm{d}t} = -\frac{(V_i - E_L^X)}{\tau^X} + \frac{\Delta_T^X \exp\left[(V_i - V_T^X)/\Delta_T^X\right]}{\tau^X} - \frac{w_i^X(t)}{C_m^X} - \frac{1}{C_m^X}\sum_Y g_{\mathrm{i,syn}}^{XY}(t)(V_i - E_{\mathrm{syn}}^Y) + \frac{\mu_{\mathrm{ext}}^X}{\tau^X} + \frac{\sigma_{\mathrm{ext}}^X}{\sqrt{\tau^X}}\eta_i(t).$$

$\mu_{\mathrm{ext}}^X + \frac{\sigma_{\mathrm{ext}}^X}{\sqrt{\tau^X}}\eta_i(t)$ is the external input, decomposed into a mean and a noise term. $\eta_i(t)$ is a white noise Gaussian process. The adaptation current follows

$$\tau_w^X \frac{\mathrm{d}w_i}{\mathrm{d}t} = a(V_i - E_L^X) - w_i.$$

The synaptic conductance from population Y to $i$-th neuron in population X follows

$$\frac{\mathrm{d}g_{\mathrm{i,syn}}^{XY}}{\mathrm{d}t} = -\frac{g_{\mathrm{i,syn}}^{XY}}{\tau_{\mathrm{syn}}^Y} + \sum_{j,k} g_{ij}^{XY}\delta(t - t_{j,k}).$$

$t_{j,k}$ is the time of the $k$-th spike from neuron $j$ in population Y. Most parameters can be found in Table 3.1 and are mainly taken from Litwin-Kumar, Rosenbaum and Doiron (2016).

$g_{ij}^{XY}$ is the connection weight from $j$-th neuron of population Y to $i$-th neuron of population X. Ideally, we would like to set $g_{ij}^{XY} = G^{XY}$ with probability $P_{ij}^{XY}$, otherwise $g_{ij}^{XY} = 0$. However, this will introduce additional variability when we compare simulations with different parameters (cell density). In order to reduce this variability, the connections from population Y to X—if exist—are all-to-all. The connection weight is then the original connection weight $G^{XY}$ multiplied by the would-be probability $P_{ij}^{XY}$, $g_{ij}^{XY} = G^{XY}P_{ij}^{XY}$.

For consistency with Litwin-Kumar, Rosenbaum and Doiron (2016), we assume that neurons are tuned to orientation. This assumption has little effect on our results. Neurons' preferred orientations span $(-\pi/2, \pi/2)$ uniformly. And the connection probability is given by

$$P_{ij}^{XY} = P_0^{XY}(1 + P_2^{XY}\cos(\theta_i - \theta_j)).$$

$\theta_i$ is the preferred orientation of the $i$-th neuron from population X. The connection weight matrix, and the would-be connection probability matrices are given below.

| | E | PV | SST | VIP | Unit | Description |
|---|---|---|---|---|---|---|
| $N$ | 4000 | 500 | 250 | 250 | | Number of neurons |
| $C_m$ | 180 | 80 | 80 | 80 | pF | Membrane capacitance |
| $g_L$ | 6.25 | 10 | 5 | 5 | nS | Leak conductance |
| $\tau$ | 28.8 | 8 | 16 | 16 | ms | Membrane time constant |
| $E_L$ | -60 | -60 | -60 | -60 | mV | Resting potential |
| $V_T$ | -40 | -40 | -45 | -45 | mV | Threshold voltage |
| $\Delta_T$ | 1 | 0.25 | 1 | 1 | mV | EIF slope parameter |
| $V_{\text{re}}$ | -60 | -60 | -60 | -60 | mV | Reset potential |
| $\tau_{\text{ref}}$ | 2 | 2 | 2 | 2 | ms | Refractory period |
| $a$ | 4 | 0 | 4 | 4 | nS | Subthreshold adaptation |
| $C_m$ | 8 | 0 | 8 | 8 | pA | Spike-triggered adaptation |
| $\tau_w$ | 150 | 150 | 150 | 150 | ms | Adaptation time constant |
| $\sigma_{\text{ext}}$ | 3.5 | 3.5 | 3.5 | 3.5 | mV | Standard deviation of external input |
| $E_{\text{syn}}$ | 0 | -67 | -67 | -67 | mV | Reversal potential |
| $\tau_{\text{syn}}$ | 2 | 3 | 4 | 4 | ms | Synaptic time constant |

Table 3.1: Spiking network parameters.

$$G = \begin{bmatrix} 0.15 & 1.0 & 1.0 & 0.0 \\ 0.7 & 1.0 & 1.0 & 0.0 \\ 0.35 & 0.0 & 0.0 & 0.25 \\ 0.35 & 0.5 & 0.25 & 0.0 \end{bmatrix} \text{nS}$$

$$P_0 = \begin{bmatrix} 0.05 & 0.3 & 0.3 & 0.0 \\ 0.2 & 0.4 & 0.4 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.4 \\ 0.2 & 0.1 & 0.2 & 0.0 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

In **Fig. 3.12**, we varied the density of PV and SST neurons. If we keep the external background inputs to the system the same while varying cell density, then the spontaneous activities of the network would be very different, which in turn affects how neurons respond to inputs. So a critical step is to ensure the spontaneous activities of the network is more or less the same when varying cell density. In order to achieve this uniform spontaneous activity, we used an optimization algorithm to find the proper mean background inputs $\mu^X_{\text{ext}}, X = E, P, S, V$. The network shows a mild level of synchronization at the spontaneous state.

We compared the spontaneous activity condition with the PV-activated condition, in which the mean level of external input to the PV population $\mu^P_{\text{ext}}$ was increased by 0.5mV. This is a relatively modest increase aimed at probing the linear response of the system. We computed the change in the average currents from PV, SST, and E neurons to E neu-

rons.

## 3.5 Appendix

### 3.5.1 Gradients of circuit responses with respect to cell densities

We studied a simplified linear circuit model

$$\tau \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = -\mathbf{r}(t) + \mathbf{W}\mathbf{r}(t) + \mathbf{u}(t).$$

In component form,

$$\tau \frac{\mathrm{d}r_X}{\mathrm{d}t} = -r_X(t) + \sum_Y W_{XY} r_Y(t) + u_X(t),$$

where $X, Y \in \{E, P, S, V\}$. The output connection weights of a population are scaled by the cell density of that population, $W_{XY} = \rho_Y W_{XY,0}$.

**Rate responses** In this model, the steady state of the system, given it is stable, is $\mathbf{r}_{\mathrm{ss}} = (\mathbf{1} - \mathbf{W})^{-1}\mathbf{u}$, where $\mathbf{1}$ is the identity matrix. We denote

$$\mathbf{M} = \frac{\partial \mathbf{r}_{ss}}{\partial \mathbf{u}},$$

the linear response matrix. An entry of this matrix,

$$M_{XY} = \frac{\partial r_{X,ss}}{\partial u_Y}$$

, represents the change in X population activity when unit-level external inputs target the Y population. $M_{XY}$ takes into account not only the direct connection $W_{XY}$ (if exists), but also all the indirect connections through other types of neurons. For the linear system,

we have

$$\mathbf{M} = \frac{\partial \mathbf{r}_{ss}}{\partial \mathbf{u}} = (\mathbf{1} - \mathbf{W})^{-1}.$$

We are interested in how the response depends on cell densities. Therefore we now calculate the derivative:

$$\frac{\partial M_{XY}}{\partial \rho_Z}.$$

First, we calculate this derivative

$$\begin{aligned}
\frac{\partial M_{XY}}{\partial W_{AB}} &= \left[\frac{\partial M}{\partial W_{AB}}\right]_{XY} \\
&= -\left[M\frac{\partial M^{-1}}{\partial W_{AB}}M\right]_{XY} \\
&= -\left[M\frac{\partial(\mathbf{1}-\mathbf{W})}{\partial W_{AB}}M\right]_{XY} \\
&= \sum_{P,Q} M_{XP}\frac{\partial W_{PQ}}{\partial W_{AB}}M_{QY} \\
&= M_{XA}M_{BY}.
\end{aligned}$$

Notice that

$$\begin{aligned}
\frac{\partial W_{XY}}{\partial \rho_Z} &= \frac{\partial \rho_Y W_{XY,0}}{\partial \rho_Z} \\
&= \delta_{YZ} W_{XY,0} \\
&= \delta_{YZ}\frac{W_{XY}}{\rho_Y} \\
&= \delta_{YZ}\frac{W_{XY}}{\rho_Z}
\end{aligned}$$

Then we have

$$\begin{aligned}
\frac{\partial M_{XY}}{\partial \rho_Z} &= \sum_V \frac{\partial M_{XY}}{\partial W_{VZ}} \frac{\partial W_{VZ}}{\partial \rho_Z} \\
&= \sum_V M_{XV} M_{ZY} \cdot \frac{W_{VZ}}{\rho_Z} \\
&= \frac{M_{ZY}}{\rho_Z} [MW]_{XZ} \\
&= \frac{M_{ZY}}{\rho_Z} \left[ M(\mathbf{1} - M^{-1}) \right]_{XZ} \\
&= \frac{M_{ZY}}{\rho_Z} [M - \mathbf{1}]_{XZ} \\
&= \frac{(M_{XZ} - \delta_{XZ}) M_{ZY}}{\rho_Z}.
\end{aligned}$$

Therefore when $X \neq Z$, we have simply

$$\frac{\partial M_{XY}}{\partial \rho_Z} = \frac{M_{XZ} M_{ZY}}{\rho_Z}.$$

**Current responses** We next study how the current responses depend on cell densities. An unit-level external input targeting a Y population $u_Y$ will change the X population steady-state activity $r_{X,ss}$ by $M_{XY}$, which then changes the steady-state X-to-V current $I_{VX,ss}$ by $W_{VX} M_{XY}$. Mathematically speaking, the current response

$$\frac{\partial I_{VX,ss}}{\partial u_Y} = \frac{\partial (W_{VX} r_{X,ss})}{\partial u_Y} = W_{VX} M_{XY}.$$

We can calculate the derivative of this current response with respect to the density of population Z, $\rho_Z$:

$$\begin{aligned}
\frac{\partial W_{VX} M_{XY}}{\partial \rho_Z} &= W_{VX} \frac{\partial M_{XY}}{\partial \rho_Z} + M_{XY} \frac{\partial W_{VX}}{\partial \rho_Z} \\
&= W_{VX} \frac{M_{ZY}(M_{XZ} - \delta_{XZ})}{\rho_Z} + M_{XY} \frac{\partial W_{VX}}{\partial \rho_Z} \\
&= W_{VX} \frac{M_{ZY}(M_{XZ} - \delta_{XZ})}{\rho_Z} + M_{XY} \frac{W_{VZ}}{\rho_Z} \delta_{XZ} \\
&= \frac{W_{VX} M_{XZ} M_{ZY}}{\rho_Z}
\end{aligned}$$

We can understand the signs of all rate and current responses presented in the main text with these expressions. For example, increasing SST+ density strengthens the PV-to-E current response when external inputs target the PV population. Mathematically, this means

$$\frac{\partial W_{EP} M_{PP}}{\partial \rho_S} < 0.$$

The above equation holds true because

$$\frac{\partial W_{EP} M_{PP}}{\partial \rho_S} = \frac{W_{EP} M_{PS} M_{SP}}{\rho_S},$$

and $M_{PS} < 0, M_{SP} < 0$. $M_{PS} M_{SP}$ describes an effective disinhibitory loop which is strengthened with higher SST+ density.

### 3.5.2 Conditions for findings

We observe several interesting and non-trivial phenomena in the rate and current response maps **Fig. 3.8, 3.9**. The E rate response maps have similar contours when external

inputs target the PV or the SST population (**Fig. 3.9**, top row). When the external input targets the PV population, the PV-to-E current response map has similar contours as the SST-to-E current response map (**Fig. 3.9**, second column). Similar finding is seen when the external input targets the SST population (**Fig. 3.9**, third column). More specifically, the PV-to-E current response and the SST-to-E current response usually have opposite signs but similar magnitudes, regardless of the cell densities.

Below we show that all these phenomena are closely related, and they all hold true when the recurrent E-to-E excitation roughly balances the own leakage of the E population.

**Rate responses** The E rate response is described by $M_{EP}$ when external inputs target the PV population, and $M_{ES}$ when external inputs target the SST population. For these responses to have similar contours in the PV+/SST+ density plane, their gradients with respect to the cell densities

$$(\frac{\partial M_{EP}}{\partial \rho_P}, \frac{\partial M_{EP}}{\partial \rho_S}),$$

and

$$(\frac{\partial M_{ES}}{\partial \rho_P}, \frac{\partial M_{ES}}{\partial \rho_S})$$

should point to the exact opposite directions. Therefore we need

$$\frac{\partial M_{EP}}{\partial \rho_P}\frac{\partial M_{ES}}{\partial \rho_S} = \frac{\partial M_{EP}}{\partial \rho_S}\frac{\partial M_{ES}}{\partial \rho_P}, \tag{3.6}$$

as well as

$$\frac{\partial M_{EP}}{\partial \rho_P} < 0,\ \frac{\partial M_{ES}}{\partial \rho_S} < 0,\ \frac{\partial M_{EP}}{\partial \rho_S} > 0,\ \frac{\partial M_{ES}}{\partial \rho_P} > 0. \tag{3.7}$$

Because

$$\begin{aligned}
(\frac{\partial M_{EP}}{\partial \rho_P}, \frac{\partial M_{EP}}{\partial \rho_S}) &= (\frac{M_{EP}M_{PP}}{\rho_P}, \frac{M_{ES}M_{SP}}{\rho_S}) \\
(\frac{\partial M_{ES}}{\partial \rho_P}, \frac{\partial M_{ES}}{\partial \rho_S}) &= (\frac{M_{EP}M_{PS}}{\rho_P}, \frac{M_{ES}M_{SS}}{\rho_S}).
\end{aligned}$$

So equation 3.6 becomes:

$$0 = M_{PP}M_{SS} - M_{PS}M_{SP}.$$

Below we will compute $M_{PP}M_{SS} - M_{PS}M_{SP}$.

The inverse of a matrix $\mathbf{A}$ can be written as its adjugate adj($\mathbf{A}$) divided by its determinant det($\mathbf{A}$),

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$$

The VIP population plays minimal role when external inputs do not directly target them. So for simplicity, here we ignore the VIP population, and consider only excitatory,

PV, and SST populations. Then the connectivity becomes

$$\mathbf{W} = \begin{bmatrix} W_{EE} & W_{EP} & W_{ES} \\ W_{PE} & W_{PP} & W_{PS} \\ W_{SE} & 0 & 0 \end{bmatrix}.$$

The response matrix also becomes 3-by-3,

$$\mathbf{M} = \begin{bmatrix} M_{EE} & M_{EP} & M_{ES} \\ M_{PE} & M_{PP} & M_{PS} \\ M_{SE} & M_{SP} & M_{SS} \end{bmatrix}.$$

So we have

$$\begin{aligned} M_{PP}M_{SS} - M_{PS}M_{SP} &= \det(\begin{bmatrix} M_{PP} & M_{PS} \\ M_{SP} & M_{SS} \end{bmatrix}) \\ &= \text{adj}(\mathbf{M})_{EE} \\ &= \det(\mathbf{M})(M^{-1})_{EE} \\ &= \det(\mathbf{M})(\mathbf{1} - \mathbf{W})_{EE} \\ &= \det(\mathbf{M})(1 - W_{EE}) \end{aligned}$$

Since $\mathbf{M} = (\mathbf{1} - \mathbf{W})^{-1}$ is invertible, we have $\det(\mathbf{M}) \neq 0$. So for equation 3.6 to hold we need

$$W_{EE} = 1 \tag{3.8}$$

Next we derive conditions for the signs to be correct as in equation 3.7. For the re-

duced 3-dimensional system, we can explicitly calculate the determinant and the adjugate of the matrix $\mathbf{1}-\mathbf{W}$. The determinant is

$$\det(\mathbf{1}-\mathbf{W}) = \quad (W_{PP}-1)\cdot\left[(W_{EE}-1)+\frac{W_{EP}W_{PS}W_{SE}}{(-W_{PP}+1)}+\frac{W_{EP}W_{PE}}{(-W_{PP}+1)}+W_{ES}W_{SE}\right].$$

And the adjugate matrix is

$$\mathrm{adj}(\mathbf{1}-\mathbf{W}) = \begin{bmatrix} -W_{PP}+1 & W_{EP} & W_{EP}W_{PS}+W_{ES}(-W_{PP}+1) \\ W_{PE}+W_{PS}W_{SE} & -(W_{EE}-1)-W_{ES}W_{SE} & -(W_{EE}-1)W_{PS}+W_{ES}W_{PE} \\ W_{SE}(-W_{PP}+1) & W_{EP}W_{SE} & -(W_{EE}-1)(-W_{PP}+1)-W_{EP}W_{PE} \end{bmatrix}.$$

When $W_{EE}=1$ we have

$$\mathrm{adj}(\mathbf{1}-\mathbf{W}) = \begin{bmatrix} -W_{PP}+1 & W_{EP} & W_{EP}W_{PS}+W_{ES}(-W_{PP}+1) \\ W_{PE}+W_{PS}W_{SE} & -W_{ES}W_{SE} & W_{ES}W_{PE} \\ W_{SE}(-W_{PP}+1) & W_{EP}W_{SE} & -W_{EP}W_{PE} \end{bmatrix}.$$

Because

$$\mathbf{M} = (\mathbf{1}-\mathbf{W})^{-1} = \frac{1}{\det(\mathbf{1}-\mathbf{W})}\mathrm{adj}(\mathbf{1}-\mathbf{W}),$$

we have

$$\frac{\partial M_{EP}}{\partial \rho_P} = \frac{M_{EP}M_{PP}}{\rho_P} = \frac{\left[\mathrm{adj}(\mathbf{1}-\mathbf{W})\right]_{EP}\left[\mathrm{adj}(\mathbf{1}-\mathbf{W})\right]_{PP}}{\rho_P\left[\det(\mathbf{1}-\mathbf{W})\right]^2} = \frac{-W_{EP}W_{ES}W_{SE}}{\rho_P\left[\det(\mathbf{1}-\mathbf{W})\right]^2} < 0.$$

Also

$$\frac{\partial M_{ES}}{\partial \rho_P} = \frac{W_{EP} W_{ES} W_{PE}}{\rho_P \left[\det(\mathbf{1} - \mathbf{W})\right]^2} > 0.$$

Therefore the signs of $\frac{\partial M_{EP}}{\partial \rho_P}$ and $\frac{\partial M_{ES}}{\partial \rho_P}$ are fixed. On the other hand,

$$\frac{\partial M_{EP}}{\partial \rho_S} = \frac{M_{ES} M_{SP}}{\rho_S} = \frac{\left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{ES} \left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{SP}}{\rho_S \left[\det(\mathbf{1} - \mathbf{W})\right]^2},$$

$$\frac{\partial M_{ES}}{\partial \rho_S} = \frac{M_{ES} M_{SS}}{\rho_S} = \frac{\left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{ES} \left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{SS}}{\rho_S \left[\det(\mathbf{1} - \mathbf{W})\right]^2}.$$

Because

$$\left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{SP} < 0,$$

and

$$\left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{SS} > 0,$$

in order to satisfy equation 3.7, we need $\left[\text{adj}(\mathbf{1} - \mathbf{W})\right]_{ES} < 0$ or equivalently

$$-W_{ES} W_{SE} > \frac{W_{EP} W_{PS} W_{SE}}{-W_{PP} + 1}. \tag{3.9}$$

Equation 3.9 means that the SST population needs to be overall net inhibitory to the E population. In equation 3.9, the left-hand-side represents the strength of the recurrent E-SST-E inhibitory loop, while the right-hand-side represents the strength of the E-SST-

PV-E recurrent disinhibition loop. Here we see a direct tradeoff between SST's roles in inhibition and disinhibition (through PV). In our model, the parameter choice made SST overall inhibitory, satisfying equation 3.7.

**Current responses** Now we show that when $W_{EE} = 1$ the PV-to-E current response and the SST-to-E current response will always cancel out each other perfectly, leading to a net-zero change of inhibitory currents in response to external inputs onto inhibitory neurons.

Because

$$\sum_Y W_{XY} M_{YZ} = (\mathbf{WM})_{XZ} = (\mathbf{M} - \mathbf{1})_{XZ} = M_{XZ} - \delta_{XZ},$$

so

$$M_{EP} = W_{EE} M_{EP} + W_{EP} M_{PP} + W_{ES} M_{SP}$$

$$M_{ES} = W_{EE} M_{ES} + W_{EP} M_{PS} + W_{ES} M_{SS}.$$

This is simply stating that the rate response is the result of currents from PV, SST, and E populations.

When $W_{EE} = 1$ we have

$$(W_{EE} - 1) M_{EP} = 0,$$

and

$$M_{EP} = W_{EE} M_{EP}.$$

Therefore

$$W_{EP}M_{PP} + W_{ES}M_{SP} = 0,$$

and

$$\frac{\partial(I_{EP,ss} + I_{ES,ss})}{\partial u_P} = 0. \tag{3.10}$$

Similarly,

$$W_{EP}M_{PS} + W_{ES}M_{SS} = 0,$$

and

$$\frac{\partial(I_{EP,ss} + I_{ES,ss})}{\partial u_S} = 0. \tag{3.11}$$

Equations 3.10 and 3.11 mean that the total inhibitory current onto the excitatory population does not change in response to external inputs onto the inhibitory populations.

# Chapter 4

# Paradoxical response reversal in interneuronal circuits

## 4.1 Introduction

Three major non-overlapping classes of interneurons expressing parvalbumin, somatostatin and vasoactive intestinal peptide (henceforth denoted PV, SST and VIP respectively) make up more than 80% of GABAergic cells of mouse cortex (Rudy et al., 2011). These neurons show cell type specific connectivity among themselves and with excitatory (E) neurons (Pfeffer et al., 2013; Jiang et al., 2015) forming a canonical microcircuit in the cortex. This microcircuit motif, initially proposed theoretically (Wang et al., 2004), has been the subject of numerous recent experimental studies using optogenetics tools applied to behaving mice (Lee et al., 2012; Saleem et al., 2013; Kepecs and Fishell, 2014; Lee, Koch and Mihalas, 2017; Hawrylycz et al., 2016) as well as computational studies (Lee, Koch and Mihalas, 2017; Lee and Mihalas, 2016; Yang, Murray and Wang, 2016). However, we still do not fully understand the mechanisms that underlie the behavior of

this microcircuit which are often complex and counterintuitive.

A notable observation was that pyramidal neurons and VIP interneurons concomitantly increase their activities in the primary visual cortex V1 during locomotion in comparison with immobility (Niell and Stryker, 2010), even in the complete absence of visual input (Keller, Bonhoeffer and Hübener, 2012). Moreover, optogenetically activating (respectively inactivating) VIP interneurons mimics (respectively eliminates) the effect of running (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014). Since VIP cells primarily target SST cells, a natural explanation for this phenomenon is disinhibition (Wang et al., 2004; Lee et al., 2013): activation of VIP cells suppresses SST cells, therefore neurons targeted by the SST population are disinhibited, enhancing the overall activity of excitatory neurons. However, recent experiments show that the network behavior might be more complex. Namely, in darkness the activation of VIP cells results in an average decrease of SST population activity (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014; Dipoppa et al., 2016), whereas in the presence of visual stimulation the response of SST cells is reversed and its firing rate increases during locomotion compared to immobility (Pakan et al., 2016; Dipoppa et al., 2016). This finding appears to challenge the disinhibition hypothesis, suggesting that the nature of the interaction between VIP and SST could be stimulus dependent.

These experimental results raise two questions: First, the external activation of a population that directly inhibits a second population can trigger a positive response of the latter. What is the mechanism behind this apparently paradoxical behavior? Second, the same top-down modulation can trigger both a positive and a negative response of certain populations of the circuit depending on the sensory input. Under which conditions can we expect one response or the other?

In this study we model cortical activity and provide comprehensive answers to these two questions. We show that these counterintuitive phenomena rely on two basic features of cortical networks: (i) the presence of multiple populations of interneurons and (ii) nonlinear responses to input. Finally, we use our model to predict complex behaviors that have not yet been experimentally tested. Beyond the mechanistic explanation for the observed behavior in mouse V1, our work provides a very general and powerful framework to explain the dynamics of neural networks with multiple interneuron types, their context-dependent interactions, and the emergence of counterintuitive effects that may occur across different cortical structures and animals.

## 4.2 Results

We simulate microcircuit activity using a four population firing rate model (**Fig. 4.1a**). The average rate of each population is given by a nonlinear function of its input that we refer to as the f-I curve (Abbott and Chance, 2005). The f-I curve is such that when the input is low (below threshold) cells are little responsive to changes in external input. Instead for high input (above threshold) small changes in the input can drive substantial changes in the response (**Fig. 4.1b**). This nonlinearity has been analyzed experimentally and theoretically (Murphy and Miller, 2003; Phillips and Hasenstaub, 2016) and as we will show later, it is a key feature of the model.

Populations are connected according to the microcircuit scheme in (**Fig. 4.1a**) which contains the connections reported in both (Jiang et al., 2015) and (Pfeffer et al., 2013). We also consider three sources of input: (i) top-down modulation that targets VIP cells (ii) local recurrent input and (iii) constant background input set so that the populations

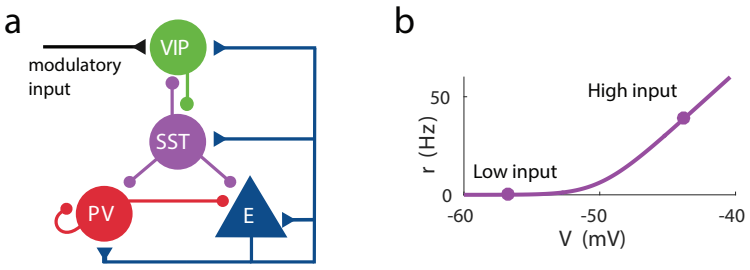


Figure 4.1: Circuit model with multiple types of interneurons. (**a**) Microcircuit connectivity and top-down modulatory input. (**b**) f-I curve. When input is low changes in input have almost no effect on the output rate, instead, when input is high changes in input have a big effect on output rate.

have some fixed baseline activity (see methods for details).

### 4.2.1 Response to top-down modulation depends on baseline activity

To illustrate possible complex behaviors displayed by the network, we first focused on the circuit responses to top-down modulation. The simulation results from our model allow us to identify two qualitatively different scenarios depending on the baseline activity of the network (the baseline activity is the activity before the onset of top-down modulation and we control it by changing the constant background input, see methods for details). On the one hand, when the baseline activity is low, top-down modulation will result in a decrease of the rate of the SST population and an increase of the rates of the other populations (E, PV and VIP) (**Fig. 4.2a**). On the other hand, when baseline activity is high, the rate of all populations increases with top-down modulation (**Fig. 4.2b**). These simulations reveal that population responses to top-down modulation depend in a complex way on the initial state of the network.

The striking behavior exhibited by the SST population can be explained heuristically by analyzing the response of the different populations to external excitatory input targeting VIP cells. When the top-down modulation starts, the rate of the VIP population

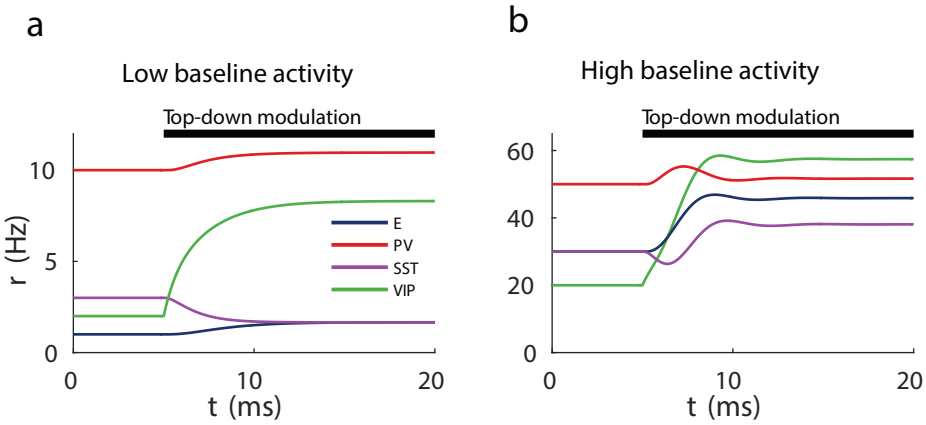


Figure 4.2: Response to top-down modulation depends on baseline activity. Transient dynamics upon the onset of the top-down modulatory current for low baseline activity (i.e. when the rates are low before top-down modulation) (**a**) and high baseline activity (i.e. when the rates are high before top-down modulation) (**b**). Under a low baseline activity condition SST is inhibited and E and PV are slightly disinhibited. The high baseline activity condition shows an example of response reversal in SST activity: it initially goes below the baseline rate but due to significant change in E activity and to the recurrent excitation it eventually reverses to a rate higher than baseline. The baseline activity can be modulated by the visual environment, for example, a bright environment would result in a higher baseline activity compared to a dark environment.

increases. This effect initially results in a reduction of SST activity and therefore a reduction of inhibition to VIP, PV and E cells. When baseline activity is low the E population is below threshold and this change in net input has a small effect on the output. In that situation all populations quickly reach a stationary state. However, when the baseline activity is high the E population is above threshold and a small change in input from SST cells has a big effect on the rate of the E population. If the recurrent excitation in the microcircuit is strong enough it can reverse the initial response of the SST population making it increase its activity to a higher rate than the baseline.

### 4.2.2 Circuit behavior explained by response matrix

In order to formally characterize the steady state response of a population to external input we introduce the response matrix $M$. The intuition behind the response matrix is

that if we change the input to population $j$ (where $j = E, P, S, V$ for excitatory, PV, SST and VIP populations respectively) by a small amount $\delta I_j$, then the change in rate of the population $i$ will be $\delta r_i = \delta I_j M_{ij}$. If $M_{ij}$ is positive (negative), an increase of the external excitation to $j$ will result in an increase (decrease) of the rate of population $i$ (see methods and **Table 4.3** for details). In contrast to the connectivity matrix, which takes into account only the direct path from population $j$ to $i$, the response matrix contains information about all the possible ways in which population $j$ can affect population $i$, namely through indirect connections $j$-$h$-$i$. Due to the complexity of these indirect pathways, for different values of the connectivity matrix (but preserving the excitatory/inhibitory structure) $M_{ij}$ can be positive or negative irrespective of whether the connection from $j$ to $i$ is inhibitory or excitatory. Furthermore due to the nonlinearities in the f-I curve, the response depends on the baseline rate of each of the populations and, as shown before, it can reverse its sign.

As an example we analyze in detail the response of the SST population to external input to VIP cells. As we show in the methods section, this term of the response matrix is given by:

$$M_{SV} = C w_{SV} ((w_{EE} - d_E)(w_{PP} + d_P) - w_{EP} w_{PE}),$$

where $w_{ij}$ are the absolute values of the connection weights and therefore are positive by definition and for the system to be stable $C$ has to be positive (see methods for details). The terms $d_i$ are proportional to the inverse of the first derivative of the f-I curves and are always positive. In particular $d_E$ becomes arbitrarily large when the input is very low and tends monotonically to a positive constant $d_E^\infty$ for high input. Therefore, if $w_{EE} \leq d_E^\infty$ then $M_{SV}$ will always be negative. However, for $w_{EE} > d_E^\infty$ the behavior is much richer: if input is high then $d_E$ will be close to its minimum $d_E^\infty$ and $w_{EE} > d_E$ allowing for $M_{SV}$ to

be positive (provided that the product $w_{EP}w_{PE}$ is small enough). Instead if the input is low, $d_E$ will become very large and $M_{SV}$ will be negative.

It is remarkable that this change in the interaction between VIP and SST populations depends on the activation level of E: modifying the state of one population has a impact in the interactions between other populations. The heuristic explanation is that if the recurrent excitation is strong enough and the E population is already strongly excited (above threshold), a small decrease in the inhibition from SST to the E population can boost its activity and therefore strongly drive the whole microcircuit. If instead, the E population is in a low activation state the change in inhibition will have a weak effect that will not be able to reverse the response of SST.

This observation provides an explanation to the reversal of the response of SST to VIP activation when the baseline activity is changed: as we show in **Fig. 4.3a,c** for low baseline activity, $M_{SV}$ is negative and the presence of an external excitatory current targeting VIP cells will result in a negative response of SST cells and positive response of E, PV and VIP cells, conforming to the disinhibitory hypothesis. On the other hand, for high baseline activity (**Fig. 4.3b,d**), the response of the SST population to input to VIP cells becomes positive leading to the response reversal regime.

A similar analysis can be conducted for all terms in $M$. For example, another case of response reversal in this circuit is that of the excitatory population. **Fig. 4.3** shows that $M_{EE}$ has different signs for different baseline activity levels, meaning that in the low baseline activity scenario input to the $E$ population will result in an increase of its rate whereas in the high baseline activity scenario, increasing the external excitation to the $E$ population will result in a neat decrease of its rate. The explicit expression of $M_{EE}$ (see **Table 4.3**) reveals that if the SST-VIP-SST loop is not strong enough $M_{EE}$ will always be

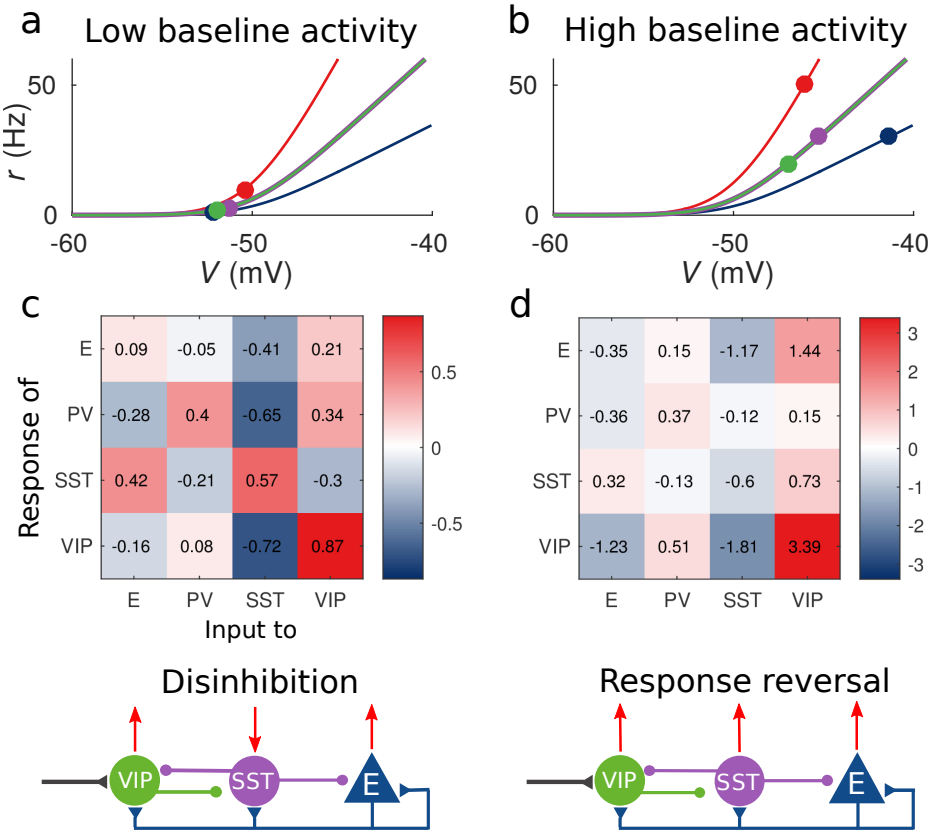


Figure 4.3: Response matrix and disinhibition vs. response reversal regime. (**a,b**) Tuning curves for the different populations and baseline activity in both scenarios (low and high). In the low baseline activity scenario (**a**) all populations are below threshold (flat part of the fI curve), instead in the high baseline activity scenario (**b**) all populations are above threshold, where small changes in input result in large changes in rate. (**c,d**) Response matrices for the two scenarios. In (**c**) the response of SST to external excitation of VIP is negative, while the responses of E and PV are positive. This corresponds to the disinhibition regime. In (**d**) the responses of all populations to external excitation of VIP are positive, in particular, the response of SST is reversed with respect to (**c**) corresponding to the response reversal regime.

positive.

### 4.2.3 Random network model

Experimental recordings showed a great diversity across neural responses even when recording from the same class of cells (Pyramidal, SST, PV or VIP) (Dipoppa et al., 2016; Pakan et al., 2016). Although this diversity can have many origins, such as intrinsic heterogeneity in the cells within the same class, we proposed that random connectivity alone is sufficient to explain it. To do so we develop an extension of our model (**Fig. 4.4**), where each population is composed of multiple identical randomly connected rate units and where the probability that one connection exists from one unit to another depends on the populations of the presynaptic and postsynaptic units according to data extracted from (Jiang et al., 2015; Pfeffer et al., 2013) (see methods for details).

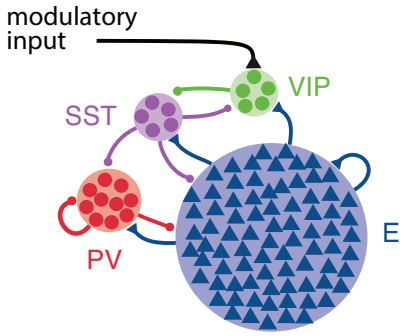


Figure 4.4: Random network model. Schematic of the model. Each population is composed of several rate units and the connectivity between units is random with probabilities extracted from experimental data in the literature.

For each unit we measure the rate modulation (rate during top-down modulation minus baseline activity) for the different baselines. If the rate modulation is positive it means that the neuron is more active in the presence of the modulatory current and vice versa. In **Fig. 4.5** we show scatter plots of the rate modulation under the low baseline condition versus the rate modulation under the high baseline condition for each unit.

These simulations reveal that the behavior of individual neurons can be quite variable while the population average still corresponds to the behavior of the population based model. Since all units of each population are identical, variability in the response has to be due to the heterogeneity in the connectivity. This variability can result in cells within the same population having responses with opposite sign, as has been observed to be the case in mouse V1 (Reimer et al., 2014; Pakan et al., 2016; Dipoppa et al., 2016) and A1 (Kuchibhotla et al., 2016). In addition variability might also have further implications for gating of signals, since variability in inhibitory cells has been proposed to modulate the response gain of neural circuits (Mejias and Longtin, 2014).

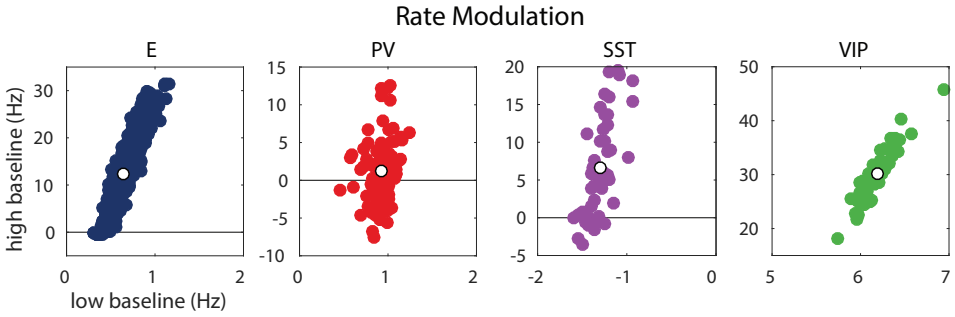


Figure 4.5: Rate modulation in the random network model. Rate modulation (rate after the onset of the modulatory current minus baseline rate) for low and high baseline activities. Each colored point corresponds to one unit. Unit responses are very variable and, in particular within the same population different units might have responses with different sign. White points correspond to the population average. Despite the variability of individual responses the population average corresponds to the population responses in the single unit model in **Fig. 4.2**.

### 4.2.4 Simulation of V1 accounts for experimental measurements

Our framework allows us to easily understand the counterintuitive behavior of V1 during locomotion. In the experiments mice with their head fixed face a screen where different visual stimuli are presented and can run freely on a treadmill (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014; Pakan et al., 2016; Dipoppa et al., 2016).

Different visual stimuli result in different baseline activities in V1 and top-down modulation is triggered when the mice start running.

To model visual input we use external currents. In the case of size-varying gratings this input has two sources: thalamic input that targets excitatory cells and cortical input that targets SST cells. In order to reproduce the surround suppression effect (Ozeki et al., 2009; Adesnik et al., 2012) excitatory cells have a small receptive field and therefore receive center input and SST cells have a large receptive field and receive surround input (see methods for details).

**Fig. 4.6** shows the response reversal phenomenon when a weak visual stimulus is presented. Before the visual stimulation the SST has higher activity for immobility than for locomotion, by contrast, when the visual stimulus is presented, the activity of the SST population is higher for locomotion. In **Fig. 4.7a** we show the experimental data from (Pakan et al., 2016) for three different experimental conditions (darkness, gray screen and grating) and in **Fig. 4.7b** our simulations of V1 under the same conditions. Similarly **Fig. 4.8a** shows the experimental data from (Dipoppa et al., 2016) for gratings of different sizes and **Fig. 4.8b** shows the behavior of our model. Comparisons in **Fig. 4.7** and **Fig. 4.8** show that our simulations reproduce qualitatively the activity of neural populations in mice V1. Namely the activity of all populations is higher during locomotion than during immobility whenever there is visual stimulation and for E, PV and VIP also in the absence of visual stimulation. Our model shows a decrease in activity of SST during locomotion as reported in the experiments (the change in activity of the SST population in darkness in **Fig. 4.7c** is not statistically significant). Our model also exhibits surround suppression for all populations. The quantitative differences might be related to the fact that changes in calcium fluorescence are not proportional to changes in rate.

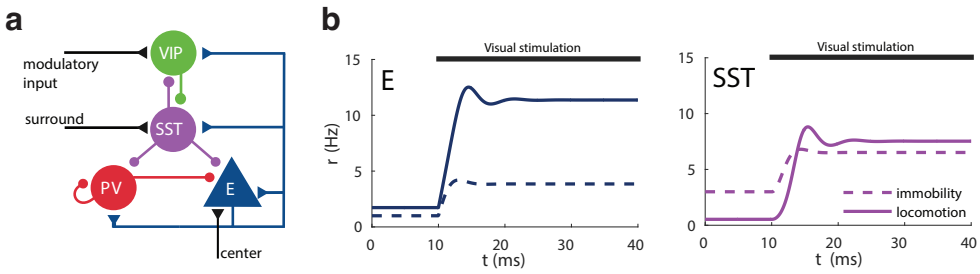


Figure 4.6: Model of mouse V1 behavior. (**a**) Schematic of the microcircuit. Visual input targets E and SST cells. Behavior related top-down modulation targets VIP cells. (**b**) Response of E and SST populations when a weak visual stimulus (6 deg) is presented for locomotion and immobility. The E population always shows a higher response with locomotion. On the other hand, before the visual stimulation the SST population has higher activity for immobility than for locomotion and when the visual stimulus is presented, the activity of the SST population is higher for locomotion.

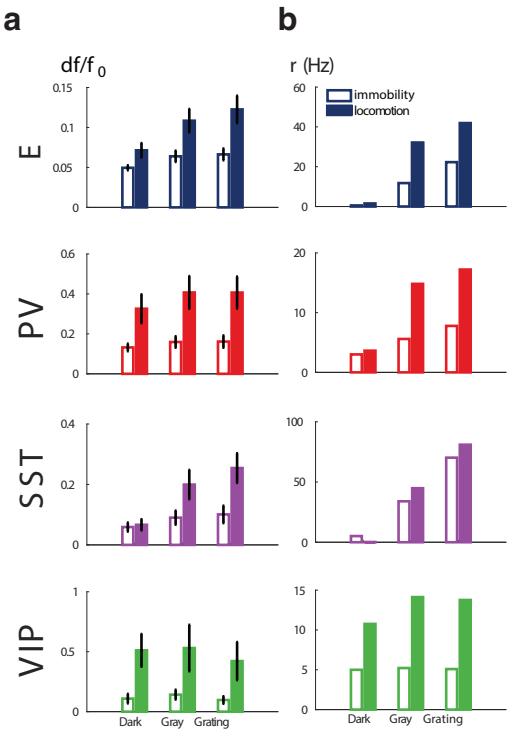


Figure 4.7: Model simulation of Pakan et al. 2016. (**a**) Relative change in calcium fluorescence for three levels of visual stimulation (darkness, gray screen and grating) and two behavioral states: immobility (empty bars) and locomotion (filled bars) extracted from (Pakan et al., 2016). (**b**) Rates (in Hz) of the populations in the V1 simulation for the same conditions as in (**a**).

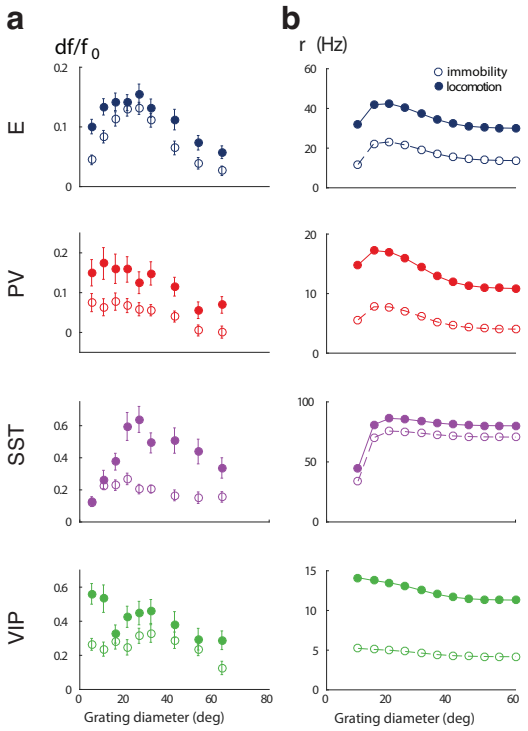


Figure 4.8: Model simulation of Dipoppa et al. 2016. (**a**) Relative change in calcium fluorescence for gratings of diameters ranging from 10 deg to 60 deg for the two behavioral states: immobility (empty dots) and locomotion (filled dots) extracted from (Dipoppa et al., 2016). (**b**) Rates (in Hz) of the populations in the V1 simulation for the same conditions as in (**a**).

Our simulations of this V1 circuit model reproduce the phenomena described in the literature: in darkness, the activities of excitatory, PV and VIP populations increase during locomotion whereas the activity of the SST population decreases with respect to the activity during immobility (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014; Dipoppa et al., 2016). In the presence of visual stimulation the activities of all populations, including SST, increase during locomotion (Pakan et al., 2016; Dipoppa et al., 2016).

To show that our results do not rely on a fine tuning of the connectivity parameters or even on certain details of the microcircuit structure we have run the model with several connectivity matrices and perturbations of them (**Fig. 4.9**) and we find that different connectivity parameters can reproduce the same circuit behavior as has been shown before in other systems (Marder, Goeritz and Otopalik, 2015). We have also considered other microcircuit structures to account for the differences between studies ((Pfeffer et al., 2013) reports projections from PV to VIP and (Jiang et al., 2015) from PV to SST) and we also consider thalamic input to PV (**Fig. 4.10**). In all these cases, the results were consistent with our original findings showing that the phenomenon and the analysis are robust and not a peculiarity of one specific circuit.

## 4.3 Discussion

We have developed a theoretical model of the cortical circuit with multiple interneuron types that accounts for newly identified complex interactions between cell types. The model has been used to reproduce and explain two counterintuitive phenomena observed in mouse cortex. First, in certain cases the activation of VIP cells results in an

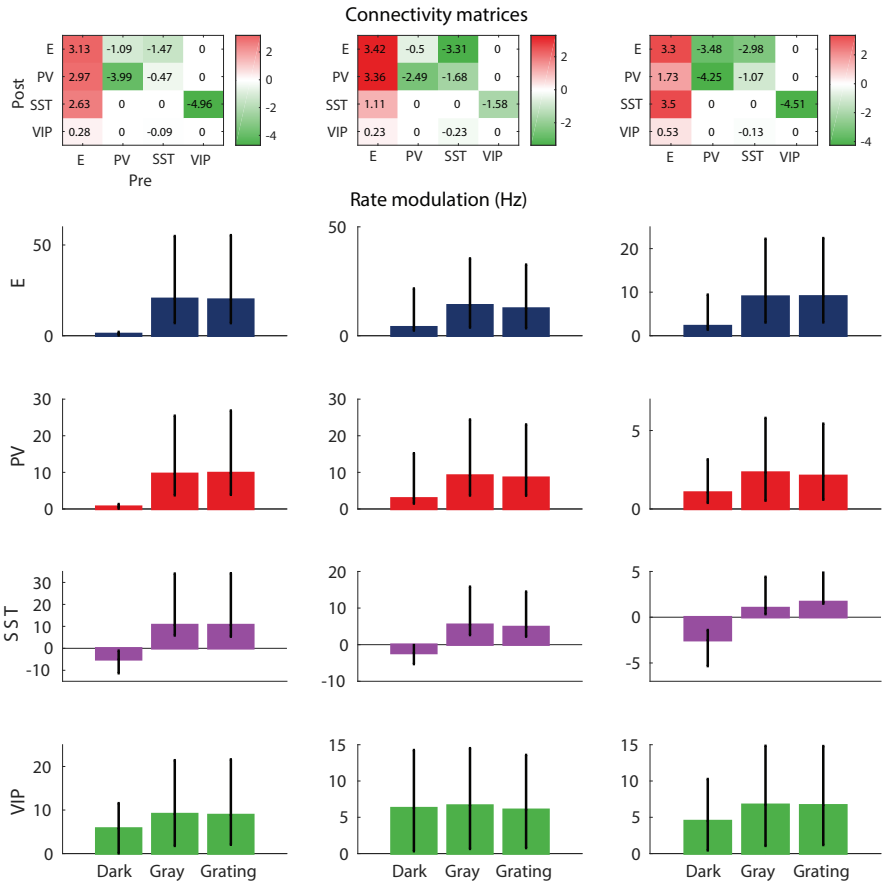


Figure 4.9: Robustness of the behavior. Top: Example of three connectivity matrices that have the same qualitative behavior. Bottom: rate modulation (rate during locomotion minus rate for immobility). Each bar corresponds to the average rate modulation of 20 random perturbations of the matrices on the top where each entry has been multiplied by a random variable uniformly distributed in $[0.9, 1.1]$, which corresponds to random changes of up to $\pm 10\%$. Error bars correspond to the minimum and maximum rate modulations of the 20 realizations. Despite quantitative variations, the qualitative behavior is always the same: rate modulation of SST population in darkness is always negative; rate modulation for all other cases is always positive.

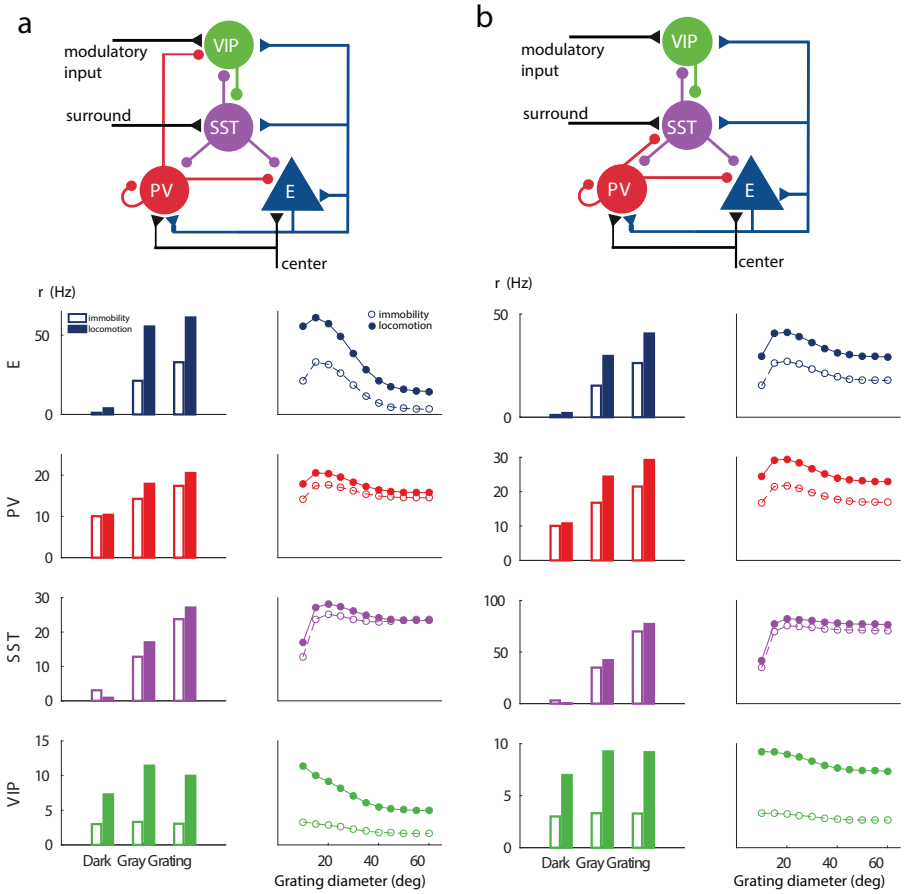


Figure 4.10: Alternative architectures. Two alternative microcircuits with visual input targeting E, SST and PV populations and PV to VIP (a) and PV to SST (b) connections.

overall positive response of the SST population (Dipoppa et al., 2016; Pakan et al., 2016). Second, the sign of the SST population response to excitation of VIP cells depends on the baseline activity of the circuit (Fu, Tucciarone, Espinosa, Sheng, Darcy, Nicoll, Huang and Stryker, 2014; Dipoppa et al., 2016). Two features of the system lead to this behavior: the presence of multiple interneuronal populations and the nonlinearity of f-I curves.

We explained heuristically the response reversal by closely looking at transient dynamics of the circuit. One experimentally-testable prediction of our analysis is that in the response reversal regime, the overall SST population response to top-down modulation should initially decrease and later increase until reaching a higher rate than the baseline.

Based on our model we introduced the response matrix $M$, which is a comprehensive framework to understand counterintuitive steady state responses. It provides explicit information about the contribution of each individual connection. For example by looking at the elements in $M_{SV}$ (see **Table 4.3**), one can readily see that if the recurrent excitation between pyramidal cells is not large enough, $M_{SV}$ can only be negative and therefore response reversal of SST would not happen. As we discussed before, another example is that if both SST and VIP populations have high baseline activities and if the SST-VIP-SST loop is strong enough, $M_{EE}$ can be negative, i.e. the excitatory population can have a negative response to excitatory input (see **Table 4.3** for the explicit expression of $M_{EE}$). If the connections between the SST and the VIP populations are removed (or weakened) or if their baseline activities are sufficiently lowered $M_{EE}$ will always be positive. This constitutes another interesting prediction that can be experimentally tested.

Our calculations also revealed sign correlations between entries of $M$, for example $M_{SV}$ and $M_{SS}$ have opposite signs for any connectivity matrix (given the microcircuit)

and for any baseline activity. This predicts that in the regime where SST activity has a positive response to excitatory input targeting VIP, SST has to have a negative response to external input targeting SST. This prediction means that increased grating size, which provides extra excitation to the SST population (Adesnik et al., 2012), should actually decrease the SST activity, as observed in both biological data (Dipoppa et al., 2016) and our model but not in previous experiments (Adesnik et al., 2012). In addition our results are in line with experimental studies that show that VIP interneurons play an important role in cortical activity modulation (Mesik et al., 2015; Ibrahim et al., 2016; Jackson et al., 2016).

The analysis of the response matrix shows that for the given microcircuit structure we employed, all terms of the matrix can be positive or negative. This is not the case for a network with one excitatory (E) population and only one inhibitory (I) population (Tsodyks et al., 1997; Ozeki et al., 2009). In that case $M_{EE}$ and $M_{IE}$ are always positive, $M_{EI}$ is always negative and only $M_{II}$ can have both signs. In this sense, having more than one inhibitory population results in a much more versatile network.

Our approach constitutes a general conceptual framework in which previous work regarding complex cortical interactions can be better understood (Tsodyks et al., 1997; Ozeki et al., 2009; Litwin-Kumar, Rosenbaum and Doiron, 2016). It provides a parsimonious yet powerful explanation for striking observations of interneuronal circuits in V1 (Pakan et al., 2016; Dipoppa et al., 2016; Lee and Mihalas, 2016) without requiring the assumption of top-down excitatory inputs explicitly targeting SST or PV neurons. Interestingly, both our computational neural network model and the approach presented here (the response matrix analysis) go beyond circuit dynamics in mice V1 and can be easily applied to other cortical areas, for example, it could be extended to explain similar phenomena observed in primary auditory cortex (Seybold et al., 2015; Kuchibhotla et al.,

2016).

We have shown that similarly to the now well-known paradoxical effect that the presence of a single inhibitory neuron type can cause (Tsodyks et al., 1997; Ozeki et al., 2009), the presence of multiple types of interneurons has an even stronger impact on the activity of neural circuits. We have also exposed the effect of nonlinearity of the f-I curve. Our analysis suggests that in a circuit with multiple populations, the most interesting circuit behavior is found when spontaneous baseline activity is close to threshold since in that regime responses will change the most with small changes in population rates. These two features significantly broaden the richness of the dynamics of cortical circuits and enhance their usefulness for cognitive and behavioral computations. We conclude that computational models and mathematical analysis are critical to fully understand the dynamics of neural circuits underlying behavior (Gjorgjieva, Drion and Marder, 2016), especially when several types of interneurons are involved as intuition alone may be misleading and provide erroneous predictions on such circuits.

## 4.4 Methods

### 4.4.1 Firing rate based population model

The state of the system is characterized by the rates $r_i$. To model the average rate of each population we use a function of the input $V_i$ as the one introduced in (Abbott and Chance, 2005)

$$r_i = f(V_i) = \frac{V_i - V_{th}}{\tau(V_{th} - V_r)} \frac{1}{1 - e^{-(V_i - V_{th})}} \tag{4.1}$$

where $V_{th} = -50$ mV and $V_r = -60$ mV are the threshold and reset potentials respectively and $\tau$ is the membrane time constant. $V_i$ is the average input to each of the populations and is given by

$$V_i = V_l + \left( \sum_j W_{ij} r_j + I_i + I^i_{bkg} \right) / g^i_l \tag{4.2}$$

where $V_l = -70$ mV is the reversal potential and $g_l$ is the membrane conductance. $W$ is the connectivity matrix and therefore $\sum_j W_{ij} r_j$ is the recurrent local input. $I_i$ is the external input current and $I^i_{bkg}$ is a constant current that is tuned to obtain the desired baseline activity. The rate dynamics are given by

$$\tau_r \frac{dr_i}{dt} = -r_i + f(V_i) \tag{4.3}$$

where $\tau_r = 2$ ms (Gerstner, 2000). Since the parameters of the f-I curve are population dependent (see **Table 4.2**), different populations will have different rates for the same input. The nonlinearity of the f-I curve has very important consequences. Namely, for low input $f(V_i)$ is almost flat, and therefore changes in the input will have almost no effect on the rate. By contrast, for strong input $f(V_i)$ tends asymptotically to a straight line with slope $\frac{1}{\tau_i (V_{th} - V_r)}$ and changes in the input will elicit a large change in the rate. As we will show later, this feature is key to reproduce the response reversal observed in the experiments.

The connectivity matrix $W$ used in the simulations is generated by rejection sampling,

i.e. by generating random matrices that have the microcircuit structure given in **Fig. 4.1a** and selecting the ones that produce the desired responses. The simulations of **Fig. 4.1**, **Fig. 4.2**, **Fig. 4.3** where done with the connectivity matrix given in **Table 4.1**.

| | | from | | | |
|---|---|---|---|---|---|
| | | E | PV | SST | VIP |
| to | E | 3.36 | -1.84 | -3.23 | 0 |
| | PV | 1.96 | -3.63 | -2.93 | 0 |
| | SST | 2.87 | 0 | 0 | -1.04 |
| | VIP | 1.9 | 0 | -1.17 | 0 |

Table 4.1: Connectivity matrix.

| | E | PV | SST | VIP |
|---|---|---|---|---|
| $g_l$ | 6.25 nS | 10 nS | 5 nS | 5 nS |
| $\tau$ | 28 ms | 8 ms | 16 ms | 16 ms |

Table 4.2: Population dependent parameters.

Behavioral state is modelled with a constant top-down modulatory current of 10 pA that targets VIP cells. The constant background inputs $I^i_{bkg}$ are set so that in the absence of the top-down modulatory current, the E, PV, SST and VIP populations will have spontaneous average rates of 1, 10, 3 and 2 Hz respectively for the low baseline activity scenario and 30, 50, 30 and 20 Hz for the high baseline activity.

### 4.4.2 Response matrix and response reversal

In order to characterize the response of a population to external excitatory input to the network we calculate how its rate will change for a small change in external input. We focus on stationary states $r_i = f(V_i)$. If we apply a small perturbation to the external input $\delta I_i$, the network will reach a new stationary state

$$r_i + \delta r_i = f(V_i + \delta V_i) = f(V_i) + f'(V_i)\delta V_i + O(\delta V_i^2) \tag{4.4}$$

where $f'(V_i)$ is the derivative of $f$ with respect to $V$ and

$$\delta V_i = \left( \sum_j W_{ij} \delta r_j + \delta I_i \right) / g_l^i. \tag{4.5}$$

Since $r_i = f(V_i)$, when we linearize $f$ around $V$ and ignore terms of order $\delta V^2$ and higher we obtain the following self-consistent equation

$$\delta r_i = f'(V_i) \left( \sum_j W_{ij} \delta r_j + \delta I_i \right) / g_l^i. \tag{4.6}$$

We define the entries of response matrix as the derivative $M_{ij} = \frac{\partial r_i}{\partial I_j}$, which can be obtained from the limit $\delta I_j \to 0$ in the system of equations given by (4.6) and in matrix form can be written as

$$M = (D - W)^{-1} \tag{4.7}$$

where $D$ is a diagonal matrix with entries $D_{ii} = g_{l,i} / f'(V_i)$. As it was explained in the results section, the nonlinear behavior of the terms $D_{ii}$ is essential to explain the response reversal regime. $D_{ii}$ becomes arbitrarily large as $V_i \to -\infty$ and decreases monotonically to $d_i^\infty = \tau_i (V_{th} - V_r)/g_l$ when $V_i \to \infty$.

In **Table 4.3** we give the explicit formulas to all the entries of the response matrix in terms of the entries of the connectivity matrix $W$ and $D$ (we denote $w = |W|$, $d_i = D_{ii}$ and $C = \det(D - W)^{-1}$). Note that, because of the complex interactions in the network, the

sign of $M_{ij}$ is never determined exclusively by that of $W_{ij}$. An interesting observation is that diagonal terms in the response matrix $M_{ii}$ only depend on the corresponding direct connections $W_{ii}$ through the determinant $C$.

| |
|---|
| $M_{EE} = C(w_{PP} + d_P)(d_S d_V - w_{SV} w_{VS})$<br>$M_{PE} = C(w_{PE}(d_S d_V - w_{SV} w_{VS}) - w_{PS}(w_{SE} d_V - w_{SV} w_{VE}))$<br>$M_{SE} = C(w_{PP} + d_P)(w_{SE} d_V - w_{SV} w_{VE})$<br>$M_{VE} = C(w_{PP} + d_P)(w_{VE} d_S - w_{SE} w_{VS})$ |
| $M_{EP} = -C w_{EP}(d_S d_V - w_{SV} w_{VS})$<br>$M_{PP} = -C((w_{EE} - d_E)(d_S d_V - w_{SV} w_{VS}) + w_{ES}(w_{SE} d_V - w_{SV} w_{VE}))$<br>$M_{SP} = -C w_{EP}(w_{SE} d_V - w_{SV} w_{VE})$<br>$M_{VP} = -C w_{EP}(w_{VE} d_S - w_{SE} w_{VS})$ |
| $M_{ES} = -C d_V(w_{ES}(w_{PP} + d_P) - w_{EP} w_{PS})$<br>$M_{PS} = -C d_V(w_{ES} w_{PE} - (w_{EE} - d_E) w_{PS})$<br>$M_{SS} = -C d_V((w_{EE} - d_E)(w_{PP} + d_P) - w_{EP} w_{PE})$<br>$M_{VS} = -C(w_{VE}(w_{ES}(w_{PP} + d_P) - w_{EP} w_{PS}) + w_{VS}((w_{EE} - d_E)(w_{PP} + d_P) - w_{EP} w_{PE}))$ |
| $M_{EV} = C w_{SV}(w_{ES}(w_{PP} + d_P) - w_{EP} w_{PS})$<br>$M_{PV} = C w_{SV}(w_{ES} w_{PE} - (w_{EE} - d_E) w_{PS})$<br>$M_{SV} = C w_{SV}((w_{EE} - d_E)(w_{PP} + d_P) - w_{EP} w_{PE})$<br>$M_{VV} = C(w_{ES}(w_{ES}(w_{PP} + d_P) - w_{EP} w_{PS}) - d_S((w_{EE} - d_E)(w_{PP} + d_P) - w_{EP} w_{PE}))$ |

Table 4.3: Entries of the response matrix.

### 4.4.3 Random network model

We consider a network with 800 E units, 100 PV units, 50 SST units and 50 VIP units. Each unit within a population has the same f-I curve with the parameters in **Table 4.2**. The probabilities $p_{ij}$ of a connection from each unit in population $j$ to each unit in population $i$ are estimated from data (Pfeffer et al., 2013; Jiang et al., 2015) and are given in **Table 4.4**.

The strengths of the connections are rescaled so that the average input of a unit in population $j$ from all units in population $i$ is $W_{ij}$ as given in **Table 4.1**. Top-down modulatory current and background input is identical to all units within the same population.

| | | from | | | |
|---|---|---|---|---|---|
| | | E | PV | SST | VIP |
| to | E | 0.02 | 1 | 1 | 0 |
| | PV | 0.01 | 1 | 0.85 | 0 |
| | SST | 0.01 | 0 | 0 | -0.55 |
| | VIP | 0.01 | 0 | 0.5 | 0 |

Table 4.4: Connection probabilities for the random network model.

### 4.4.4 Mouse V1 model

In the simulations of V1 activity we use the connectivity matrix given in **Table 4.5**.

| | | from | | | |
|---|---|---|---|---|---|
| | | E | PV | SST | VIP |
| to | E | 3.30 | -3.48 | -2.98 | 0 |
| | PV | 1.73 | -4.25 | -1.07 | 0 |
| | SST | 3.50 | 0 | 0 | -4.51 |
| | VIP | 0.53 | 0 | -0.13 | 0 |

Table 4.5: Connectivity matrix for the mouse V1 model.

We model visual input with an external excitatory current that targets E and SST cells. In the experiments in (Dipoppa et al., 2016; Pakan et al., 2016) the authors consider three levels of visual stimulation which are: darkness, gray screen and grating. To model darkness condition we assume a total absence of visual stimulation (therefore $I_E = 0$ pA, $I_S = 0$ pA). For gray screen we use a small input current to the excitatory population ($I_E = 50$ pA, $I_S = 0$ pA). Finally to model different grating diameters the value of the input is a sigmoid function of the grating diameter $\theta$:

$$I_i(\theta) = \frac{a_i}{1 + e^{-\theta/b_i + 5}} \tag{4.8}$$

where $b_E = 2$, $b_S = 6$, $a_E = 100$ pA, $a_S = 20$ pA. With this parameters E cells receive center input (input saturates for diameters $\sim 20$ deg) and SST cells receive surround input (input to SST saturates for diameters of $\sim 60$ deg) (Dipoppa et al., 2016).

To demonstrate that our results do hold for a wide range of connectivity matrices and do not have to be fine tuned, we simulate several different connectivity matrices that produce the same qualitative behavior. We also make perturbations of these matrices by multiplying each entry by a random variable uniformly distributed in the interval $[0.9, 1.1]$. This amounts to randomly modifying each connection within $\pm 10\%$ of its original value (see **Fig. 4.9a**).

In the alternative models of **Fig. 4.9b** where visual stimulus input also targets PV cells, we use $I_P = 0$ pA for darkness, $I_P = 10$ pA for gray screen and $b_P = 2$, $a_P = 20$ pA for gratings.

# Chapter 5

# Recurrent neural networks trained for cognitive tasks

## 5.1 Introduction

Computations in the brain are carried out by populations of interconnected neurons. While single-neuron responses can reveal a great deal about the neural mechanisms underlying various sensory, motor, and cognitive functions, neural mechanisms often involve the coordinated activity of many neurons whose complex individual dynamics are not easily explained by tuning to experimental parameters (Barak, Tsodyks and Romo, 2010; Rigotti et al., 2010 2013; Yuste, 2015). A growing recognition of the importance of studying population-level responses is reflected in the increasing number of studies that use large datasets of simultaneously or sequentially recorded neurons to infer neural circuit mechanisms (Mante et al., 2013; Churchland et al., 2012; Carnevale et al., 2015; Sussillo et al., 2015; Siegel, Buschman and Miller, 2015). At the same time, the novel challenges posed by high-dimensional neural data have led to the development of new meth-

ods for analyzing and modeling such data (Stevenson and Kording, 2011; Cunningham and Yu, 2014; Gao and Ganguli, 2015).

One approach that has emerged as a promising tool for identifying the dynamical and computational mechanisms embedded in large neural populations is to study model recurrent neural networks (RNNs) whose connection weights have been optimized to perform the same tasks as recorded animals (Mante et al., 2013; Carnevale et al., 2015). In Mante et al. (2013), for example, the "trained" network was analyzed to reveal a previously unknown selection mechanism for context-dependent integration of sensory stimuli that was consistent with data obtained from behaving monkeys. RNNs of rate units, which describe biological circuits as a set of firing rates (nonlinearities) interacting through synapses (connection weights) (**Fig. 5.1**), interpolate between biophysically detailed spiking-neuron models and the wider class of continuous-time dynamical systems: the units of an RNN can be interpreted as the temporal or ensemble average of one or more co-tuned spiking neurons (Gerstner and Kistler, 2002), while any nonlinear dynamical system can be approximated by an RNN with a sufficient number of units (Funahashi and Nakamura, 1993; Siegelmann and Sontag, 1995). The optimization of network parameters typically specifies the desired output but not the manner in which to achieve this output, i.e., the *what* but not the *how*. Trained RNNs therefore serve as a source of candidate hypotheses about circuit mechanisms and a testing ground for data analyses that link neural computation to behavior. Complete access to the activity and connectivity of the circuit, and the ability to manipulate them in arbitrary ways, make trained networks a convenient proxy for biological circuits and a valuable platform for theoretical investigation (Yamins et al., 2014; Sussillo and Barak, 2013; Gao and Ganguli, 2015).

For many tasks of interest, however, training can result in multiple networks that

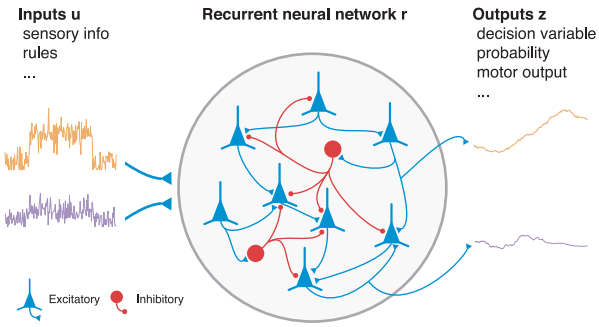


Figure 5.1: **Recurrent neural network (RNN).** A trained RNN of excitatory and inhibitory rate units $\mathbf{r}(t)$ receives time-varying inputs $\mathbf{u}(t)$ and produces the desired time-varying outputs $\mathbf{z}(t)$. Inputs encode task-relevant sensory information or internal rules, while outputs indicate a decision in the form of an abstract decision variable, probability distribution, or direct motor output. Only the recurrent units have their own dynamics: inputs are considered to be given and the outputs are read out from the recurrent units. Each unit of an RNN can be interpreted as the temporally smoothed firing rate of a single neuron or the spatial average of a group of similarly tuned neurons.

achieve the same behavioral performance but differ substantially in their connectivity and dynamics. As highlighted in recent work (Sussillo et al., 2015), the particular solution that is discovered by the training algorithm depends strongly on the set of constraints and "regularizations" used in the optimization process, so that training RNNs to perform a task is not entirely unbiased with respect to the *how*. Indeed, for the purposes of modeling animal tasks in systems neuroscience the question is no longer whether an RNN can be trained to perform the task—the answer appears to be yes in a wide range of settings—but what architectures and regularizations lead to network activity that is most similar to neural recordings obtained from behaving animals.

Answering this question is essential if RNNs are to provide insights into the operation of the brain at the level of neural circuits (Zipser and Andersen, 1988), and extends the classical connectionist approach (Rumelhart, Hinton and Williams, 1985; Cohen, Dunbar and McClelland, 1990). Doing so requires a simple and flexible framework for the exploratory training of RNNs to investigate the effects of different constraints on network properties, particularly those constraints that render the RNNs more biologically plausi-

ble. For instance, many RNNs studied to date have "firing rates" that are both positive and negative. More fundamentally, existing networks do not satisfy Dale's principle (Eccles, Fatt and Koketsu, 1954), the basic and ubiquitous observation that neurons in the mammalian cortex have purely excitatory or inhibitory effects on other neurons. The analogous constraint that all connection weights from a given unit must have the same sign can have a profound effect on the types of dynamics, such as non-normality (Murphy and Miller, 2009), that operate in the circuit. Moreover, connections from excitatory and inhibitory neurons exhibit different levels of sparseness and specificity, with non-random features in the distribution of connection patterns among neurons both within local circuits (Markram et al., 2004; Song et al., 2005; Pfeffer et al., 2013; Potjans and Diesmann, 2014; Jiang et al., 2015) and among cortical areas (Ercsey-Ravasz et al., 2013; Markov et al., 2014; Song, Kennedy and Wang, 2014). Notably, long-range projections between areas are primarily excitatory. Such details must be included in a satisfactory model of local and large-scale cortical computation.

We address this challenge by describing flexible, gradient descent-based training of excitatory-inhibitory RNNs that can incorporate a variety of biological knowledge, particularly of local and large-scale connectivity in the brain. Several different methods have previously been used to train RNNs for cognitive tasks in neuroscience, including first-order reduced and controlled error (FORCE) (Sussillo and Abbott, 2009; Laje and Buonomano, 2013; Carnevale et al., 2015) and Hessian-free (HF) (Martens and Sutskever, 2011; Mante et al., 2013; Barak et al., 2013). Here we use minibatch stochastic gradient descent (SGD) with the modifications described in Pascanu, Mikolov and Bengio (2012), which remove the major difficulties associated with pure gradient descent training of RNNs. SGD is conceptually simple without sacrificing performance (Bengio, Boulanger-

Lewandowski and Pascanu, 2013; Hardt, Recht and Singer, 2015) and is particularly advantageous in the present context for the following reasons: Unlike FORCE and like HF, SGD allows us to more easily formulate the problem of training an RNN as one of minimizing an objective function that can be modified to induce different types of solutions (Sussillo et al., 2015). Meanwhile, like FORCE and unlike HF, for many tasks SGD can update parameters on a trial-by-trial basis, i.e., in an "online" fashion. This opens up the possibility of exploring across-trial effects that cannot be studied when large numbers of trials are required for each iteration of learning, as in the HF algorithm. Although none of the learning methods discussed here can at present be considered biological, recent work also suggests that spike-timing dependent plasticity (STDP) (Dan and Poo, 2006), which is believed to be a basic rule governing synaptic weight changes in the brain, may correspond to a form of SGD (Bengio, Lee, Bornschein, Mesnard and Lin, 2015; Bengio, Mesnard, Fischer, Zhang and Wu, 2015). However, the focus of our approach will be on the results, not the mechanism, of learning.

We provide an implementation of this framework based on the Python machine learning library Theano (Bergstra et al., 2010), whose automatic differentiation capabilities facilitate modifications and extensions. Theano also simplifies the use of Graphics Processing Units (GPUs) when available to speed up computations. The implementation was designed to minimize the overhead for each new task by only requiring a specification of the network structure and correct input-output relationship to be learned. It also streamlines the testing and analysis of the resulting networks by using the same (customizable) specification for both training and testing. We demonstrate the application of this framework to well-known experimental paradigms that illustrate the diversity of tasks and details that can be modeled: perceptual decision-making, context-dependent integration, mul-

tisensory integration, parametric working memory, and eye-movement sequence generation. Using the resulting networks we perform both single-neuron and population-level analyses associated with the corresponding experimental paradigm. Our results show that trained RNNs provide a unified setting in which diverse computations and mechanisms can be studied, laying the foundation for more neuroscientists to harness trained RNNs in their own investigations of the neural basis of cognition.

## 5.2 Materials and Methods

In this section we first define the RNNs used in this work, show how constraints can be introduced, then describe training the networks using a modified form of stochastic gradient descent (SGD).

### 5.2.1 Recurrent neural networks

RNNs receive a set of $N_{\text{in}}$ time-varying inputs $\mathbf{u}(t)$ and produce $N_{\text{out}}$ outputs $\mathbf{z}(t)$, where inputs encode task-relevant sensory information and outputs typically represent a decision variable or probability distribution (**Fig. 5.1**). Outputs can also relate to the direct motor effector, such as eye position, by which an animal indicates its decision in the behavioral paradigm. We consider RNNs whose $N$ firing rates $\mathbf{r}(t)$ are related to their corresponding currents $\mathbf{x}(t)$ by the threshold (rectified) linear "$f$-$I$ curve" $[x]_+ = \max(x, 0)$, which maps arbitrary input currents to positive firing rates: $x$ if $x > 0$ and 0 otherwise.

The RNNs are described by the equations

$$\tau\dot{\mathbf{x}} = -\mathbf{x} + W^{\text{rec}}\mathbf{r} + W^{\text{in}}\mathbf{u} + \sqrt{2\tau\sigma^2_{\text{rec}}}\,\boldsymbol{\xi}, \tag{5.1}$$

$$\mathbf{r} = [\mathbf{x}]_+, \tag{5.2}$$

$$\mathbf{z} = W^{\text{out}}\mathbf{r}, \tag{5.3}$$

or, more explicitly,

$$\tau\frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} W^{\text{rec}}_{ij} r_j + \sum_{k=1}^{N_{\text{in}}} W^{\text{in}}_{ik} u_k + \sqrt{2\tau\sigma^2_{\text{rec}}}\,\xi_i, \tag{5.4}$$

$$r_i = [x_i]_+, \tag{5.5}$$

$$z_\ell = \sum_{i=1}^{N} W^{\text{out}}_{\ell i} r_i \tag{5.6}$$

for $i = 1,\ldots,N$ and $\ell = 1,\ldots,N_{\text{out}}$. In these equations $\tau$ is the time constant of the network units, $W^{\text{in}}$ is an $N \times N_{\text{in}}$ matrix of connection weights from the inputs to network units, $W^{\text{rec}}$ is an $N \times N$ matrix of recurrent connection weights between network units, $W^{\text{out}}$ is an $N_{\text{out}} \times N$ matrix of connection weights from the network units to the outputs, and $\boldsymbol{\xi}$ are $N$ independent Gaussian white noise processes with zero mean and unit variance that represent noise intrinsic to the network. It is worth noting that if for some $\ell = 1,\ldots,N'$, $N' \leq N$, the output weights $W^{\text{out}}_{\ell i} = \delta_{\ell i}$ where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, then the readout is the same as a subset of the network firing rates. This is useful in situations where the aim is to fix a subset of the units to experimentally recorded firing rates.

Without the rectification nonlinearity $[x]_+$ (in which case $\mathbf{r} = \mathbf{x}$), Eqs 5.1-5.3 would describe a linear system whose dynamics is completely determined by $W^{\text{rec}}$. Thus, one way to understand the effect of rectification is to consider a linear dynamical system whose

coupling matrix $W^{\text{rec}}$ at any given time includes only those columns that correspond to "active" units with positive summed current $x_i$ (and hence positive firing rate $r_i$) (Hahnloser, 1998). This toggles the network between different linear maps, thereby endowing the network with the capacity for more complex computations than would be possible with a single linear network (Pascanu, Montufar and Bengio, 2013; Montufar et al., 2014). As a convenient baseline, the recurrent noise in Eq 5.1 has been scaled so that in the corresponding linear network without rectification and when $W^{\text{rec}} = W^{\text{in}} = 0$,

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + \sqrt{2\tau\sigma^2_{\text{rec}}}\,\boldsymbol{\xi}, \tag{5.7}$$

each unit is an Ornstein-Uhlenbeck process with variance $\sigma^2_{\text{rec}}$.

In practice, the continuous-time dynamics in Eqs 5.1-5.3 are discretized to Euler form (which we indicate by writing time as a subscript, $X_t = X(t \cdot \Delta t)$ for a time-dependent variable $X$) in time steps of size $\Delta t$ as (Gillespie, 1996)

$$\mathbf{x}_t = (1-\alpha)\mathbf{x}_{t-1} + \alpha(W^{\text{rec}}\mathbf{r}_{t-1} + W^{\text{in}}\mathbf{u}_t) + \sqrt{2\alpha\sigma^2_{\text{rec}}}\,\mathbf{N}(0,1), \tag{5.8}$$

$$\mathbf{r}_t = [\mathbf{x}_t]_+, \tag{5.9}$$

$$\mathbf{z}_t = W^{\text{out}}\mathbf{r}_t, \tag{5.10}$$

where $\alpha = \Delta t/\tau$ and $\mathbf{N}(0,1)$ are normally distributed random numbers with zero mean and unit variance, sampled independently at every time step. In this formulation, the discrete-time RNNs used in machine learning applications correspond to $\alpha = 1$ or $\Delta t = \tau$. In our network, $\Delta t < \tau$. To minimize computational effort we train the network with a value of $\Delta t$ that is small enough such that the same network behavior is recovered in the continuous limit of $\Delta t \to 0$. We chose $\Delta t = 0.2\tau$, therefore $\alpha = 0.2$.

Although the details of the inputs to the network are specific to each task, it is convenient to represent all inputs as a rectified sum of baseline $\mathbf{u}^0$, task-dependent signal $\mathbf{u}^{\text{task}}(t)$, and Gaussian white noise $\boldsymbol{\xi}$:

$$\mathbf{u}(t) = \left[\mathbf{u}^0 + \mathbf{u}^{\text{task}}(t) + \sqrt{2\tau\sigma_{\text{in}}^2}\,\boldsymbol{\xi}\right]_+ \tag{5.11}$$

in the continuous description, and

$$\mathbf{u}_t = \left[\mathbf{u}^0 + \mathbf{u}_t^{\text{task}} + \frac{1}{\alpha}\sqrt{2\alpha\sigma_{\text{in}}^2}\,\mathbf{N}(0,1)\right]_+ \tag{5.12}$$

in the discrete-time description. Motivated by the interpretation that the network under study is only one part of a larger circuit, the baseline and noise terms in the inputs can together be considered the spontaneous firing rate of "upstream" units that project to the network.

We note that in Eq 5.1 the external "sensory" noise ultimately combines with the intrinsic noise, with the difference that input noise can be shared between many units in the network while the recurrent noise is private to each unit. There are many cases where the external and internal noise trade off in their effect on the network, for instance on its psychometric performance in a perceptual decision-making task. However, the two sources of noise can be biologically and conceptually quite different (Brunton, Botvinick and Brody, 2013), and for this reason it is helpful to separate the two types of noise in our formulation.

Finally, in many cases (the exception being networks that are run continuously without reset) it is convenient to optimize the initial condition $\mathbf{x}_0 = \mathbf{x}(0)$ at time $t = 0$ along with the network weights. This merely selects a suitable starting point for each run, re-

ducing the time it takes for the network to relax to its spontaneous state in the absence of inputs. It has little effect on the robustness of the network due to the recurrent noise used both during and after training; in particular, the network state at the time of stimulus onset is highly variable across trials.

### 5.2.2 RNNs with separate excitatory and inhibitory populations

A basic and ubiquitous observation in the mammalian cortex, known in the more general case as Dale's principle (Eccles, Fatt and Koketsu, 1954), is that cortical neurons have either purely excitatory or inhibitory effects on postsynaptic neurons. Moreover, excitatory neurons outnumber inhibitory neurons by a ratio of roughly 4 to 1. In a rate model with positive firing rates such as the one given by Eqs 5.1-5.3, a connection from unit $j$ to unit $i$ is "excitatory" if $W^{\text{rec}}_{ij} > 0$ and "inhibitory" if $W^{\text{rec}}_{ij} < 0$. A *unit* $j$ is excitatory if all of its projections on other units are zero or excitatory, i.e., if $W^{\text{rec}}_{ij} \geq 0$ for all $i$; similarly, unit $j$ is inhibitory if $W^{\text{rec}}_{ij} \leq 0$ for all $i$. In the case where the outputs are considered to be units in a downstream network, consistency requires that for all $\ell$ the readout weights satisfy $W^{\text{out}}_{\ell j} \geq 0$ and $W^{\text{out}}_{\ell j} \leq 0$ for excitatory and inhibitory units $j$, respectively. Since long-range projections in the mammalian cortex are exclusively excitatory, for most networks we limit readout to the excitatory units. It is also natural in most cases to assume that inputs to the network are long-range inputs from an upstream circuit, and we assume all elements of the input weight matrix $W^{\text{in}}$ are non-negative. For consistency with the following, we indicate this as $W^{\text{in}} = W^{\text{in},+}$. Once again, this is only meaningful if the inputs themselves are always non-negative, motivating the rectification of inputs in Eq 5.11.

In order to train RNNs that satisfy the above constraints, we parametrize the recurrent

weight matrix $W^{\text{rec}}$ as the product of a non-negative matrix $W^{\text{rec},+}$ and a diagonal matrix $D$ of 1's and $-1$'s, $W^{\text{rec}} = W^{\text{rec},+}D$. For example, consider a network containing 4 excitatory units and 1 inhibitory unit; the excitatory/inhibitory signature of the network is then $D = \text{diag}(1,1,1,1,-1)$ (a matrix with the specified entries on the diagonal and zeros everywhere else), and the full recurrent weight matrix has the form

$$\underbrace{\begin{pmatrix} & + & + & + & - \\ + & & + & + & - \\ + & + & & + & - \\ + & + & + & & - \\ + & + & + & + & \end{pmatrix}}_{W^{\text{rec}}} = \underbrace{\begin{pmatrix} & + & + & + & + \\ + & & + & + & + \\ + & + & & + & + \\ + & + & + & & + \\ + & + & + & + & \end{pmatrix}}_{W^{\text{rec},+}} \underbrace{\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & -1 \end{pmatrix}}_{D}, \tag{5.13}$$

where absent matrix elements indicate zeros. Although an individual unit in an RNN does not necessarily represent a single neuron, we typically fix the self-connections represented by the diagonal elements of $W^{\text{rec}}$ to be zero, see below. Similarly, if the readout from the network is considered to be long-range projections to a downstream network, then the output weights are parametrized as $W^{\text{out}} = W^{\text{out},+}D$.

During training, the positivity of $W^{\text{in},+}$, $W^{\text{rec},+}$, and $W^{\text{out},+}$ can be enforced in several ways, including rectification $[W]_+$ and the absolute value function $|W|$. Here we use rectification.

### 5.2.3 Specifying the pattern of connectivity

In addition to dividing units into separate excitatory and inhibitory populations, we can also constrain their pattern of connectivity. This can range from simple constraints such

as the absence of self-connections to more complex structures derived from biology. Local cortical circuits have distance (Levy and Reyes, 2012), layer (Thomson et al., 2002; Binzegger, Douglas and Martin, 2004; Potjans and Diesmann, 2014), and cell-type (Markram et al., 2004; Wang et al., 2004; Pfeffer et al., 2013; Jiang et al., 2015) dependent patterns of connectivity and different overall levels of sparseness for excitatory to excitatory, inhibitory to excitatory, excitatory to inhibitory, and inhibitory to inhibitory connections (Fino and Yuste, 2011; Karnani, Agetsuma and Yuste, 2014). Although the density of connections in a trained network can be either fixed (hard constraint) or induced through regularization (soft constraint) (see Eq 5.28), here we focus on the former to address the more general problem of imposing known biological structure on trained networks. For instance, in models of large-scale, distributed computation in the brain we can consider multiple cortical "areas" characterized by local inhibition within areas and long-range excitation between areas. These long-range connections can be distributed according to a highly complex topology (Ercsey-Ravasz et al., 2013; Markov et al., 2014; Song, Kennedy and Wang, 2014). It is also desirable when testing specific hypotheses about circuit structure to fix a subset of the connection weights to predefined values while leaving others as "plastic," modifiable by training.

A simple way to impose hard constraints on the connectivity is to parametrize the weight matrices using masks. As an example, suppose we would like to train a subset of the excitatory weights and also fix two of the inhibitory weights to $w_1$ and $w_2$ so that they

are not modified during training. We can implement this by writing

$$W^{\text{rec},+} = \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}}_{M^{\text{rec}}} \odot \underbrace{\begin{pmatrix} \cdot & + & + & + & \cdot \\ + & \cdot & + & \cdot & + \\ + & + & \cdot & + & + \\ \cdot & + & + & \cdot & \cdot \\ + & + & + & + & \cdot \end{pmatrix}}_{W^{\text{rec,plastic},+}} + \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 & w_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{W^{\text{rec,fixed},+}}, \tag{5.14}$$

where $\odot$ denotes the element-wise multiplication of two matrices (not standard matrix multiplication). Here $W^{\text{rec,plastic},+}$ is obtained by rectifying the (unconstrained) trained weights $W^{\text{rec,plastic}}$, so that $W^{\text{rec,plastic},+} = [W^{\text{rec,plastic}}]_+$, while $W^{\text{rec,fixed},+}$ is a matrix of fixed weights. The elements that are marked with a dot are irrelevant and play no role in the network's dynamics. Eq 5.14 has the effect of optimizing only those elements which are nonzero in the multiplying mask $M^{\text{rec}}$, which ensures that the weights corresponding to zeros do not contribute. Some elements, for instance the inhibitory weights $w_1$ and $w_2$ in Eq. 5.14, remain fixed at their specified values throughout training. Explicitly, the full weight matrix of the RNN is related to the underlying trained weight matrix $W^{\text{rec,plastic}}$ by (cf. Eq 5.13)

$$W^{\text{rec}} = (M^{\text{rec}} \odot [W^{\text{rec,plastic}}]_+ + W^{\text{rec,fixed},+})D, \tag{5.15}$$

and similarly for the input and output weights.

### 5.2.4 Initialization

In networks that do not contain separate excitatory and inhibitory populations, it is convenient to initialize the recurrent weight matrix as $W^{\text{rec}} = \rho W_0^{\text{rec}}$, where $W_0^{\text{rec}}$ is formed

by setting a fraction $p$, $0 < p \leq 1$, of elements to nonzero values drawn from a Gaussian distribution with mean 0 and variance $(pN)^{-1}$, and the remaining fraction $1-p$ to zero (Sussillo and Abbott, 2009). This can be understood as first generating a random matrix $W_0^{\text{rec}}$, then multiplying by $\rho/\rho_0$ where $\rho_0 = 1$ is the spectral radius of $W_0^{\text{rec}}$ and $\rho$ is the desired spectral radius of the initial weight matrix. Here the spectral radius is the largest absolute value of the eigenvalues.

To initialize an excitatory-inhibitory network with an arbitrary pattern of connections, we similarly first generate a matrix $W_0^{\text{rec}}$ and let $W^{\text{rec}} = (\rho/\rho_0)W_0^{\text{rec}}$ where $\rho_0$ is the spectral radius of $W_0^{\text{rec}}$. Unlike in the case of random Gaussian matrices, the (asymptotically) exact spectral radius is usually unknown and must be computed numerically. Moreover, since the signs of the matrix elements are determined by the excitatory or inhibitory nature of the units, it is more natural to use a distribution over positive numbers to first generate $W_0^{\text{rec},+}$ (Eq 5.13). Many distributions, including the uniform and log-normal distributions, can be used; inspired by previous work (Festa, Hennequin and Lengyel, 2014), here we use the gamma distribution to initialize the recurrent weight matrix $W_0^{\text{rec},+}$. The means $\mu_E$ (excitatory) and $\mu_I$ (inhibitory) of the gamma distributions are chosen to balance the excitatory and inhibitory inputs to each unit (Rajan and Abbott, 2006), i.e., $\sum_{j\in\text{exc}} |\mu_j| = \sum_{j\in\text{inh}} |\mu_j|$, with the overall mean set by the imposed spectral radius $\rho$. We did not use the "initialization trick" of Le, Jaitly and Hinton (2015), as this requires the existence of self-connections.

For the input weight matrix $W_0^{\text{in},+}$ and output weight matrix $W_0^{\text{out},+}$, we initialize with small positive numbers drawn from a uniform distribution.

### 5.2.5 Training RNNs with gradient descent

To train an RNN, we assume that at each time step (or subset of time steps) there is a correct set of target outputs $\mathbf{z}_t^{\text{target}}$ that depend on the current and previous history of inputs $\mathbf{u}_{t'}$ for $t' \leq t$, i.e., we only consider tasks that can be translated into a "supervised" form. The goal is then to find network parameters, which we collectively denote as $\boldsymbol{\theta}$, that minimize the difference between the correct output and the actual output of the network. More generally, we minimize an objective function $\mathscr{E}(\boldsymbol{\theta})$ that includes not only this error but other terms such as an $L_1$-regularization term (for encouraging sparse weights or activation patterns) that influence the types of solutions found by the training algorithm. We begin with the case where the objective function depends only on the error; one possibility for the *loss* $\mathscr{L}(\boldsymbol{\theta})$ that measures the difference between the correct and actual outputs is the squared sum of differences averaged over $N_{\text{trials}}$ trials, $N_{\text{out}}$ outputs, and $N_{\text{time}}$ time points:

$$\mathscr{E} = \frac{1}{N_{\text{trials}}} \sum_{n=1}^{N_{\text{trials}}} \mathscr{L}_n, \tag{5.16}$$

$$\mathscr{L}_n = \frac{1}{N_{\text{out}} N_{\text{time}}} \sum_{\ell=1}^{N_{\text{out}}} \sum_{t=1}^{N_{\text{time}}} M_{t\ell}^{\text{error}} \Big[ (\mathbf{z}_t)_\ell - (\mathbf{z}_t^{\text{target}})_\ell \Big]^2. \tag{5.17}$$

For each trial $n$ in Eq 5.17, $(\mathbf{z}_t)_\ell$ is the $\ell$-th output, at time $t$, of the discretized network in Eq 5.10. The error mask $M^{\text{error}}$ is a matrix of ones and zeros that determines whether the error in output $\ell$ at time $t$ should be taken into account. In many decision-making tasks, for example, this allows us to train networks by specifying only the final, but not the intermediate, time course for the outputs.

In gradient descent training the parameters of the network are updated iteratively according to (for more sophisticated forms of gradient descent see, e.g., Sutskever et al.

(2013))

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)} + \boldsymbol{\delta\theta}^{(i-1)}, \tag{5.18}$$

where $i$ denotes the iteration. The parameter change, $\boldsymbol{\delta\theta}$, is taken to be proportional to the negative gradient of the objective function with respect to the network parameters as

$$\boldsymbol{\delta\theta}^{(i-1)} = -\eta \nabla \mathscr{E}^{(i-1)}, \tag{5.19}$$

where $\eta$ is the *learning rate* and $\nabla \mathscr{E}^{(i-1)} = \nabla \mathscr{E}(\boldsymbol{\theta}^{(i-1)})$ is the value of the gradient evaluated on the parameters from iteration $i-1$. Importantly, the required gradient can be computed efficiently by backpropagation through time (BPTT) (Rumelhart, Hinton and Williams, 1985) and *automatically* by the Python machine library Theano (Bergstra et al., 2010). In component form the parameter update at iteration $i$ is given by

$$\theta_k^{(i)} = \theta_k^{(i-1)} - \eta \Big( \frac{\partial \mathscr{E}}{\partial \theta_k} \Big)^{(i-1)}, \tag{5.20}$$

where $k$ runs over all the parameters of the network that are being optimized. Eqs 5.18 and 5.19 are motivated by the observation that, for a small change $\boldsymbol{\delta\theta}$ in the value of the parameters, the corresponding change in the value of the objective function is given by

$$\mathscr{E}(\boldsymbol{\theta} + \boldsymbol{\delta\theta}) - \mathscr{E}(\boldsymbol{\theta}) \simeq \nabla \mathscr{E} \cdot \boldsymbol{\delta\theta} = |\nabla \mathscr{E}||\boldsymbol{\delta\theta}| \cos\phi, \tag{5.21}$$

where $|\cdot|$ denotes the norm of a vector and $\phi$ is the angle between $\nabla \mathscr{E}$ and $\boldsymbol{\delta\theta}$. This change is most negative when $\phi = 180°$, i.e., when the change in parameters is in the opposite direction of the gradient. "Minibatch stochastic" refers to the fact that the gradient of the objective function $\mathscr{E}(\boldsymbol{\theta})$ is only *approximated* by evaluating $\mathscr{E}(\boldsymbol{\theta})$ over a rela-

tively small number of trials (in particular, smaller than or comparable to the number of trial conditions) rather than using many trials to obtain the "true" gradient. Intuitively, this improves convergence to a satisfactory solution when the objective function is a highly complicated function of the parameters by stochastically sampling the gradient and thereby escaping saddle points (Dauphin et al., 2014) or poor local minima, while still performing an averaged form of gradient descent over many stochastic updates.

Even so, SGD with the objective function given in Eqs 5.16 and 5.17 often fails to converge to a solution when the network must learn dependencies between distant time points (Bengio, Simard and Frasconi, 1994). To remedy this problem, which is due to some gradient components being too large (*exploding* gradients) and some gradient components being too small (*vanishing* gradients), we follow Pascanu, Mikolov and Bengio (2012) in making two modifications. First, the exploding gradient problem is addressed by simply "clipping" the gradient when its norm exceeds a maximum $G$: instead of Eq 5.19 for the direction and size of the update, we use

$$\boldsymbol{\delta\theta}^{(i-1)} = \begin{cases} -\eta\nabla\mathscr{E}^{(i-1)} \times \dfrac{G}{|\nabla\mathscr{E}^{(i-1)}|} & \text{if } |\nabla\mathscr{E}^{(i-1)}| > G, \\ -\eta\nabla\mathscr{E}^{(i-1)} & \text{otherwise.} \end{cases} \tag{5.22}$$

Second, the vanishing gradient problem is addressed by modifying the objective function with the addition of a regularization term:

$$\mathscr{E} = \frac{1}{N_{\text{trials}}} \sum_{n=1}^{N_{\text{trials}}} (\mathscr{L}_n + \lambda_\Omega \Omega_n), \tag{5.23}$$

$$\Omega_n = \sum_{t=1}^{N_{\text{time}}} \left( \frac{\left| \frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_{t-1}} \right|^2}{\left| \frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_t} \right|^2} - 1 \right)^2 . \tag{5.24}$$

In Eq 5.23 the multiplier $\lambda_\Omega$ determines the effect of the regularization term $\Omega_n$, with no

effect for $\lambda_\Omega = 0$. In Eq 5.24, the first term in parentheses is the ratio between the squared norms of two vectors, which we would like to be close to 1. The somewhat opaque (row) vector expression in the numerator can be unpacked as (cf. Eq 5.8)

$$\left(\frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_t}\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_{t-1}}\right)_j = \sum_{k=1}^{N}\frac{\partial \mathscr{L}_n}{\partial (\mathbf{x}_t)_k}\frac{\partial (\mathbf{x}_t)_k}{\partial (\mathbf{x}_{t-1})_j} \tag{5.25}$$

$$= \left[(1-\alpha)\frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_t} + \alpha\left(\frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_t}W^{\text{rec}}\right)\odot(\mathbf{r}'_{t-1})\right]_j. \tag{5.26}$$

Here each component $r'(x_t)$ of $\mathbf{r}'(\mathbf{x}_t)$ is the derivative of the $f$-$I$ curve, i.e., 1 if $x > 0$ and 0 otherwise in the case of rectification, and $\odot$ denotes element-wise multiplication of two vectors. For consistency in notation we treat $\mathbf{r}'_{t-1}$ here as a row vector. One subtlety in the implementation of this term is that, for computational efficiency, only the "immediate" derivative of $\Omega_n$ with respect to the network parameters is used, i.e., with $\mathbf{x}_t$ and $\partial \mathscr{L}_n/\partial \mathbf{x}_t$ treated as constant (Pascanu, Mikolov and Bengio, 2012). The relevant network parameters in this case are the elements of the trained weight matrix $W^{\text{rec,plastic}}$, which is related to $W^{\text{rec}}$ through Eq. 5.15.

The role of the regularization term $\Omega_n$ is to preserve the size of the gradients as errors are backpropagated through time. This is accomplished by preserving the norm of $\partial \mathbf{x}_t/\partial \mathbf{x}_{t-1}$, which propagates errors in time (Pascanu, Mikolov and Bengio, 2012), along $\partial \mathscr{L}_n/\partial \mathbf{x}_t$, which is the direction in which the change in the objective function is greatest with respect to $\mathbf{x}_t$. More intuitively, the impact of the regularization term on network dynamics can be understood by noting that if $\partial \mathbf{x}_t/\partial \mathbf{x}_{t'}$ is small for some $t' < t$ then, by definition, $\mathbf{x}_t$ does not depend on small changes in $\mathbf{x}_{t'}$, which may occur when $\mathbf{x}$ is close to an attractor. Preserving the norm of $\partial \mathbf{x}_t/\partial \mathbf{x}_{t-1}$ through time therefore encourages the network to remain at the boundaries between basins of attraction and thus encourages

longer computation times. For instance, this results in perceptual decision networks that can integrate their inputs for a long period of time, before converging to one of the choice attractors. We note that, although the numerator and denominator in Eq 5.24 appear, by the chain rule, to preserve the ratio of $\partial \mathscr{L}_n / \partial \mathbf{x}_{t-1}$ to $\partial \mathscr{L}_n / \partial \mathbf{x}_t$, this is only approximately true. Specifically,

$$\frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_{t-1}} = \frac{\partial \mathscr{L}_n}{\partial \mathbf{x}_{t-1}} - \frac{\partial \mathscr{L}_{n,t-1}}{\partial \mathbf{x}_{t-1}}, \tag{5.27}$$

because $\mathscr{L}_{n,t-1}$, the component of $\mathscr{L}_n$ from time $t-1$, depends on $\mathbf{x}_{t-1}$ but not on $\mathbf{x}_t$.

Finally, additional regularization terms may be included to change either the dynamics or the connectivity. For instance, there are two ways of obtaining sparse recurrent connectivity. First, we can impose a hard constraint that fixes a chosen subset of weights to be nonzero and modifiable by the optimization algorithm as described above. Second, we may apply a soft constraint by adding the sum of the $L_1$-norms of the weights to the objective function:

$$\mathscr{E} = \frac{1}{N_{\text{trials}}} \sum_{n=1}^{N_{\text{trials}}} (\mathscr{L}_n + \lambda_\Omega \Omega_n) + \frac{\lambda_1^{\text{rec}}}{N^2} \sum_{j,k=1}^{N} |W_{jk}^{\text{rec}}|. \tag{5.28}$$

In addition, we may choose to encourage solutions with small firing rates through regularization of the $L_2$-norms of the firing rates Sussillo et al. (2015):

$$\mathscr{E} = \frac{1}{N_{\text{trials}}} \sum_{n=1}^{N_{\text{trials}}} (\mathscr{L}_n + \lambda_\Omega \Omega_n + \lambda_2^{\text{fr}} R_n^{\text{fr}}) + \frac{\lambda_1^{\text{rec}}}{N^2} \sum_{j,k=1}^{N} |W_{jk}^{\text{rec}}|, \tag{5.29}$$

$$R_n^{\text{fr}} = \frac{1}{N N_{\text{time}}} \sum_{j=1}^{N} \sum_{t=1}^{N_{\text{time}}} (\mathbf{r}_t)_j^2 \tag{5.30}$$

where $(\mathbf{r}_t)_j$ is the firing rate of the $j$-th unit at time $t$ on each trial. Again, we gain flexibility in defining more complex regularization terms because Theano computes the nec-

essary gradients using BPTT. Although BPTT is simply a specialized chain rule for neural networks, automatic differentiation frees us from implementing new gradients each time the objective function is changed. This greatly facilitates the exploration of soft constraints such as those considered in Sussillo et al. (2015).

### 5.2.6 Training protocol

To demonstrate the robustness of the training method, we used many of the same parameters to train all tasks (**Table 5.1**). In particular, the learning rate $\eta$, maximum gradient norm $G$, and the strength $\lambda_\Omega$ of the vanishing-gradient regularization term were kept constant for all networks. We also successfully trained networks with values for $G$ and $\lambda_\Omega$ that were larger than the default values given in **Table 5.1**. When one or two parameters were modified to illustrate a particular training procedure, they are noted in the task descriptions. For instance, the number of trials used for each parameter update (gradient batch size) was the same in all networks except for the context-dependent integration task (to account for the large number of conditions) and sequence execution task (because of online training, where the number of trials is one). As a simple safeguard against extreme fine-tuning, we removed all weights below a threshold, $w_{\min}$, after training. We also note that, unlike in previous work (e.g., Mante et al. (2013)), we used the same level of stimulus and noise for both training and testing.

Code for generating the figures in this work are available from `https://github.com/xjwanglab/py`
The distribution includes code for training the networks, running trials, performing analyses, and creating the figures.

| Parameter | Symbol | Default value |
|---|---|---|
| Learning rate | $\eta$ | 0.01 |
| Maximum gradient norm | $G$ | 1 |
| Multiplier for vanishing-gradient regularization $\Omega$ | $\lambda_\Omega$ | 2 |
| Unit time constant | $\tau$ | 100 ms |
| Time step (training) | $\Delta t$ | $\tau/5$ |
| Time step (testing) | $\Delta t$ | 0.5 ms |
| Initial spectral radius of recurrent weight matrix | $\rho$ | 1.5 |
| Gradient minibatch size | $N_{\text{trials}}$ | 20 |
| Baseline input | $u^0$ | 0.2 |
| Standard deviation for input noise | $\sigma_{\text{in}}$ | 0.01 |
| Standard deviation for recurrent noise | $\sigma_{\text{rec}}$ | 0.15 |
| Minimum weight threshold after training | $w_{\text{min}}$ | $10^{-4}$ |

Table 5.1: **Parameters for stochastic gradient descent (SGD) training of recurrent neural networks (RNNs).** Unless noted otherwise in the task description, networks were trained and run with the parameters listed here.

## 5.3 Results

In this section we present the results of applying the training framework to well-known experimental paradigms in systems neuroscience: perceptual decision-making (Newsome, Britten and Movshon, 1989; Roitman and Shadlen, 2002; Kiani, Hanks and Shadlen, 2008), context-dependent integration (Mante et al., 2013), multisensory integration (Raposo, Kaufman and Churchland, 2014), parametric working memory (Romo et al., 1999; Barak et al., 2013), and eye-movement sequence generation (Averbeck and Lee, 2007). In addition to establishing the relative ease of obtaining networks that perform the selected tasks, we show several single-neuron and population analyses associated with each paradigm. These analyses demonstrate that trained networks exhibit many, though not yet all, features observed in recorded neurons, and the study of these networks therefore has the potential to yield insights into biological neural circuits. A summary of the tasks can be found in (**Table 5.2**).

The tasks presented in this section represent only a small sample of the diversity of

tasks used in neuroscience. In addition, we have chosen—in most cases arbitrarily—a simple set of constraints that do not necessarily reflect the full biological reality. Nevertheless, our work provides the foundation for further exploration of the constraints, regularizations, and network architectures required to achieve the greatest correspondence between trained RNNs and biological neural networks.

### 5.3.1 Perceptual decision-making task

Many experimental paradigms in neuroscience require subjects to integrate noisy sensory stimuli in order to choose between two actions (**Fig. 5.1**). Here we present networks trained to perform two variants of perceptual decision-making inspired by the two common variants of the random dot motion discrimination task (Newsome, Britten and Movshon, 1989; Roitman and Shadlen, 2002; Kiani, Hanks and Shadlen, 2008). For both versions, the network has 100 units (80 excitatory and 20 inhibitory) and receives two noisy inputs, one indicating evidence for choice 1 and the other for choice 2, and must decide which is larger. Importantly, the network is not explicitly told to integrate—it is instead only required to "make a decision" following the offset of stimulus by holding a high value in the output corresponding to the higher input, and a low value in the other.

In the variable stimulus-duration version of the task (**Fig. 5.2a**), stimulus durations are drawn randomly from a truncated exponential distribution (we note that this is often called the "fixed-duration" version because the experimentalist sets the reaction time, in contrast to the "reaction-time" version in which the subject chooses when to respond). This minimizes the network's ability to anticipate the end of the stimulus and therefore encourages the network to continue integrating information as long as the stimulus is present (Kiani, Hanks and Shadlen, 2008). In the reaction-time version (**Fig. 5.2b**), the

| Task | Network inputs | Outputs | Features |
|---|---|---|---|
| Perceptual decision making (variable stimulus duration, VS; reaction-time, RT) Roitman and Shadlen (2002) | Motion 1/2 Start of stimulus | Choice 1/2 | Psychometric curves (VS, RT) Percent correct as a function of stimulus duration (VS) Reaction-time as a function of coherence, distribution (RT) Coherence-dependent firing rates (VS) Convergence of firing rates aligned to decision time (RT) |
| Perceptual decision making (fixed stimulus duration) | Motion 1/2 | Choice 1/2 | Psychometric curves No Dale's principle vs. Dale's principle Dense vs. constrained initial connectivity |
| Context-dependent integration Mante et al. (2013) | Motion 1/2 Color 1/2 Motion/Color context | Choice 1/2 | Psychometric curves, gating State-space analysis Mixed selectivity of single-unit responses Distribution of regression coefficients |
| Context-dependent integration | Same | Same | Two areas |
| Multisensory integration Raposo, Kaufman and Churchland (2014) | Pos./Neg. tuned visual/auditory Start of stimulus | Choice high-/low | Psychometric curves with multisensory enhancement Heterogeneous selectivity in single-unit responses |
| Parametric working memory Romo et al. (1999) | Pos./Neg. tuned frequency | Choice $f_1 > f_2$ Choice $f_1 < f_2$ | Heterogeneous tuning Correlation of tuning across population Change of tuning during trial |
| Sequence execution Averbeck and Lee (2007) | Targets (9) Sequence (8) | Eye position $(x, y)$ | Continuous trials Online learning State-space analysis: hierarchical decision making |

Table 5.2: **Summary of tasks.** In the multisensory integration and parametric working memory tasks, networks receive both positively (pos.; increasing function) and negatively (neg.; decreasing function) tuned versions of the same input.

network must respond soon after the onset of an ongoing stimulus. To control the speed-accuracy tradeoff, the target outputs during training did not require the network to commit to a decision immediately but instead after a short delay (Roitman and Shadlen, 2002); the delay determines the cost incurred for answering early but incorrectly versus correctly but at a later time.

All trials begin with a "fixation" period during which both outputs must maintain a low value, requiring the network to react only to the stimulus. The fixation can be enforced during training in several ways, including a variable fixation period whose duration is drawn from another truncated exponential distribution, or by introducing "catch trials" when no stimuli are presented. For simplicity, here we used a small proportion of catch trials mixed into the training, together with an additional, unambiguous start cue that signals the onset of stimulus.

Networks trained for both versions of the task show comparable performance in their psychometric functions (**Fig. 5.2c, 5.2d**), which are the percentage of trials on which the network chose choice 1 as a function of the signed coherence. Coherence is a measure of the difference between evidence for choice 1 and evidence for choice 2, and positive coherence indicates evidence for choice 1 and negative for choice 2. In experiments with monkeys the signs correspond to inside and outside, respectively, the receptive field of the recorded neuron; although we do not show it here, this can be explicitly modeled by combining the present task with the model of "eye position" used in the sequence execution task (below). We emphasize that, unlike in the usual machine learning setting, our objective is not to achieve "perfect" performance. Instead, the networks were trained to an overall performance level of approximately 85% across all nonzero coherences to match the smooth psychometric profiles observed in behaving monkeys. We note that

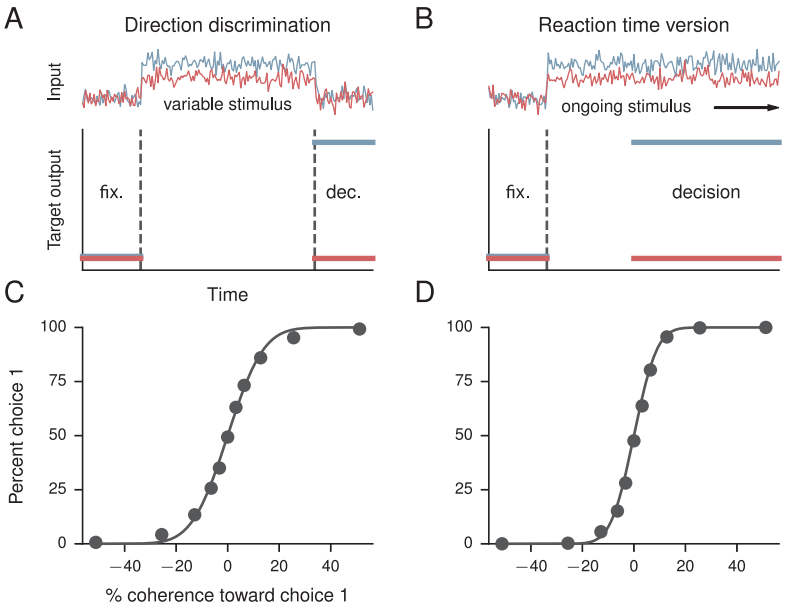


Figure 5.2: **Perceptual decision-making task.** (**a**) Inputs (upper) and target outputs (lower) for a perceptual decision-making task with variable stimulus duration, which we refer to as VS here. The choice 1 output must hold low during fixation (fix.), then high during the decision (dec.) period if the choice 1 input is larger than choice 2 input, low otherwise, and similarly for the choice 2 output. There are no constraints on output during the stimulus period. (**b**) Inputs and target outputs for the reaction-time version of the integration task, which we refer to as RT. Here the outputs are encouraged to respond after a short delay following the onset of stimulus. The reaction time is defined as the time it takes for the outputs to reach a threshold. (**c**) Psychometric function for the VS version, showing the percentage of trials on which the network chose choice 1 as a function of the signed coherence. Coherence is a measure of the difference between evidence for choice 1 and evidence for choice 2, and positive coherence indicates evidence for choice 1 and negative for choice 2. Solid line is a fit to a cumulative Gaussian distribution. (**d**) Psychometric function for the RT version.

this implies that some networks exhibit a slight bias toward choice 1 or choice 2, as is the case with animal subjects unless care is taken to eliminate the bias through adjustment of the stimuli. Together with the input noise, the recurrent noise enables the network to smoothly interpolate between low-coherence choice 1 and low-coherence choice 2 trials, so that the network chooses choice 1 on approximately half the zero-coherence trials when there is no mean difference between the two inputs. Recurrent noise also forces the network to learn more robust solutions than would be the case without.

For the variable stimulus duration version of the decision-making task, we computed the percentage of correct responses as a function of the stimulus duration for different coherences (**Fig. 5.3a**), showing that for easy, high-coherence trials the duration of the stimulus period only weakly affects performance (Kiani, Hanks and Shadlen, 2008). In contrast, for difficult, low-coherence trials the network can improve its performance by integrating for a longer period of time. **Fig. 5.3b** shows the activity of an example unit (selective for choice 1) across all correct trials, averaged within conditions after aligning to the onset of the stimulus. The activity shows a clear tuning of the unit to different signed coherences.

For the reaction-time version of the task, we defined a threshold for the output (here arbitrarily taken to be 1, slightly less than the target of 1.2 during training) that constituted a "decision." The time it takes to reach this threshold is called the *reaction time*, and **Fig. 5.3c** shows this reaction time as a function of coherence for correct trials, while the inset shows the distribution of reaction times on correct trials. In the case of the reaction-time version of the task, it is interesting to consider the activity of single units aligned to the decision time in each trial, which shows that the firing rate of the unit converges to a similar value for all positive coherences (**Fig. 5.3d**) (Roitman and Shadlen, 2002). This

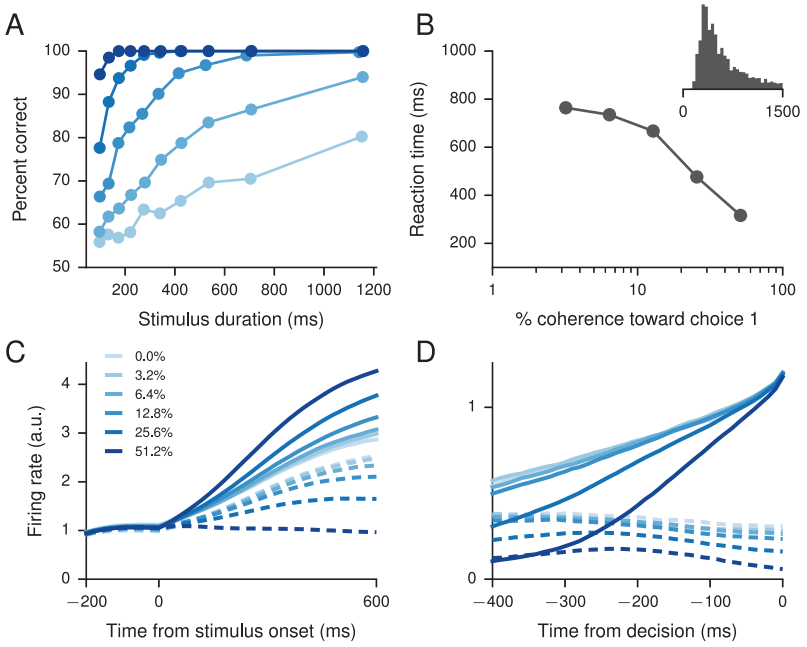


Figure 5.3: **Perceptual decision-making task analysis.** (**a**) Percentage of correct responses as a function of stimulus duration in the VS version, for each nonzero coherence level. (**b**) Reaction time for correct trials in the RT version as a function of coherence. Inset: Distribution of reaction times on correct trials. (**c**) Example activity of a single unit in the VS version across all correct trials, averaged within conditions after aligning to the onset of the stimulus. Solid (dashed) lines denote positive (negative) coherence. (**d**) Example activity of a single unit in the RT version, averaged within conditions and across all correct trials aligned to the reaction time.

is a nontrivial observation in both experiment (Roitman and Shadlen, 2002) and model, as the decision threshold is only imposed on the outputs and not on the recurrent units themselves.

To illustrate the effect of constraints on connectivity structure—but not on performance—we also trained three networks for the fixed stimulus-duration version of the task shown in **Fig. 5.2a**. For these networks we did not use a start cue. In the first network, no constraints were imposed on the connection weights except for the absence of self-connections (**Fig. 5.4a**). The second network was required to satisfy Dale's principle, with a 4-to-1 ratio of the number of excitatory to inhibitory units, and purely excitatory inputs and outputs (**Fig. 5.4b**). The third network was similar, but with the additional constraint that the inputs that signal evidence for choice 1 and choice 2 project to distinct groups of recurrent units and decisions are read out from the same group of excitatory units (**Fig. 5.4c**). The two groups of excitatory units send zero excitatory projections to each other, communicating instead only through the inhibitory units and excitatory units that receive no inputs.

In all three cases, a clear structure could be discerned in the connectivity of the trained network by sorting the units by their selectivity index

$$d' = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}, \tag{5.31}$$

where $\mu_1, \sigma_1^2$ are the mean and variance of the unit's activity, during the stimulus period, on trials in which the network chose choice 1, and similarly for $\mu_2, \sigma_2^2$ for choice 2. For the network without separate excitatory and inhibitory units (**Fig. 5.4a**), clustering manifests in the form of strong excitation among units with similar $d'$ and strong inhibition

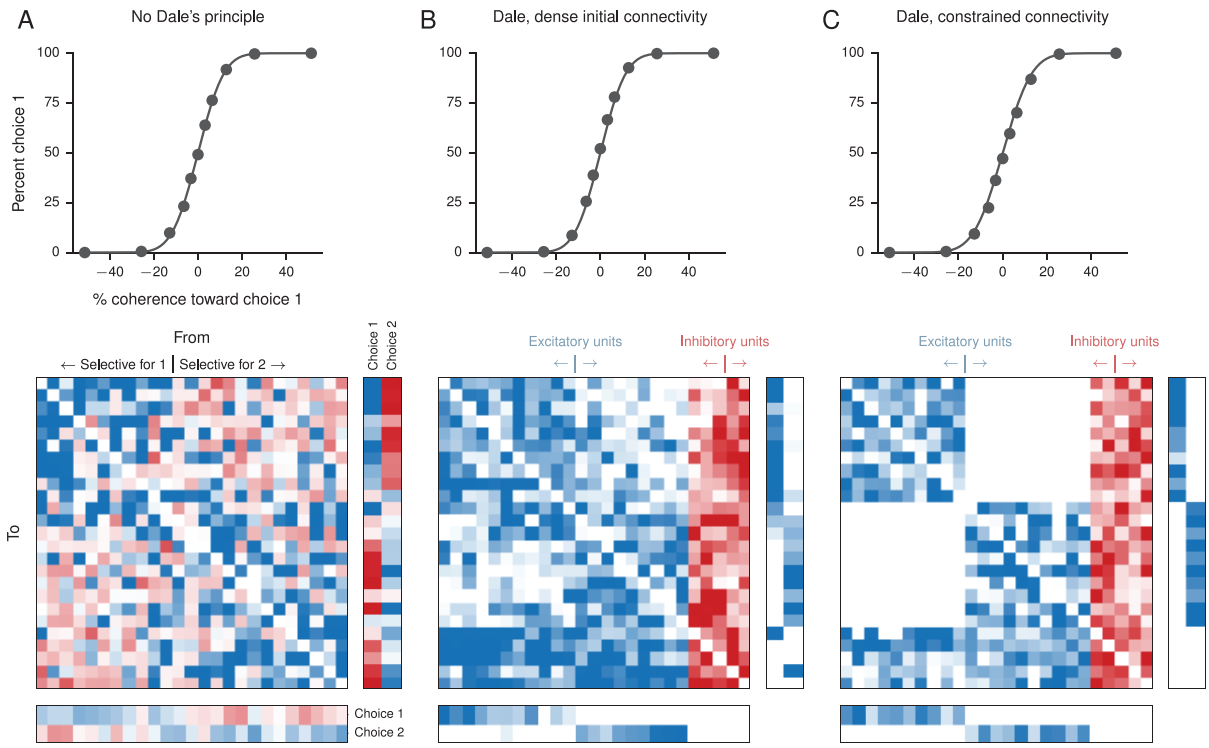


Figure 5.4: **Perceptual decision-making networks with different constraints.** (**a**) Psychometric function (percent choice 1 as a function of signed coherence) and connection weights (input, upper-right; recurrent, upper-left; and output, lower) for a network in which all weights may be positive or negative, trained for a perceptual decision-making task. Connections go from columns ("pre-synaptic") to rows ("post-synaptic"), with blue representing positive weights and red negative weights. Different color scales (arbitrary units) were used for the input, recurrent, and output matrices but are consistent across the three networks shown. In the psychometric function, solid lines are fits to a cumulative Gaussian distribution. In this and the networks in B and C, self-connections were not allowed. In each case 100 units were trained, but only the 25 units with the largest absolute selectivity index (Eq 5.31) are shown, ordered from most selective for choice 1 (large positive) to most selective for choice 2 (large negative). (**b**) A network trained for the same task as in A but with the constraint that excitatory units may only project positive weights and inhibitory units may only project negative weights. All input weights were constrained to be excitatory, and the readout weights, considered to be "long-range," were nonzero only for excitatory units. All connections except self-connections were allowed, but training resulted in a strongly clustered pattern of connectivity. Units are again sorted by selectivity but separately for excitatory and inhibitory units (20 excitatory, 5 inhibitory). (**c**) Same as B but with the additional constraint that excitatory recurrent units receiving input for choice 1 and excitatory recurrent units receiving input for choice 2 do not project to one another, and each group sends output to the corresponding choice.

between units with different $d'$. The learned input weights also excite one population and inhibit the other. In the case of the network with separate excitatory and inhibitory populations (**Fig. 5.4b**), units with different $d'$ interact primarily through inhibitory units (Wong and Wang, 2006). Importantly, despite the fact that the recurrent weight matrix was initialized with dense, all-to-all connectivity, the two populations send fewer excitatory projections to each other after training. Similarly, despite the fact that the input weights initially send evidence for both choices to the two populations, after training the two groups receive evidence for different choices. Output weights also became segregated after training. In the third network this structure was imposed from the start, confirming that such a network could learn to perform the task (**Fig. 5.4c**).

### 5.3.2 Context-dependent integration task

In this section and the next we show networks trained for experimental paradigms in which making a correct decision requires integrating two separate sources of information. We first present a task inspired by the context-dependent integration task of Mante et al. (2013), in which a "context" cue indicates that one type of stimulus (the motion or color of the presented dots) should be integrated and the other completely ignored to make the optimal decision.

A network trained for the context-dependent integration task is able to integrate the relevant input while ignoring the irrelevant input. This is reflected in the psychometric functions, the percentage of trials on which the network chose choice 1 as a function of the signed motion and color coherences (**Fig. 5.5a**). The network contains a total of 150 units, 120 of which are excitatory and 30 inhibitory. The training protocol was very similar to the (fixed-duration) single-stimulus decision-making task except for the presence of

two independent stimuli and a set of context inputs that indicate the relevant stimulus. Because of the large number of conditions, we increased the number of trials for each gradient update to 50.

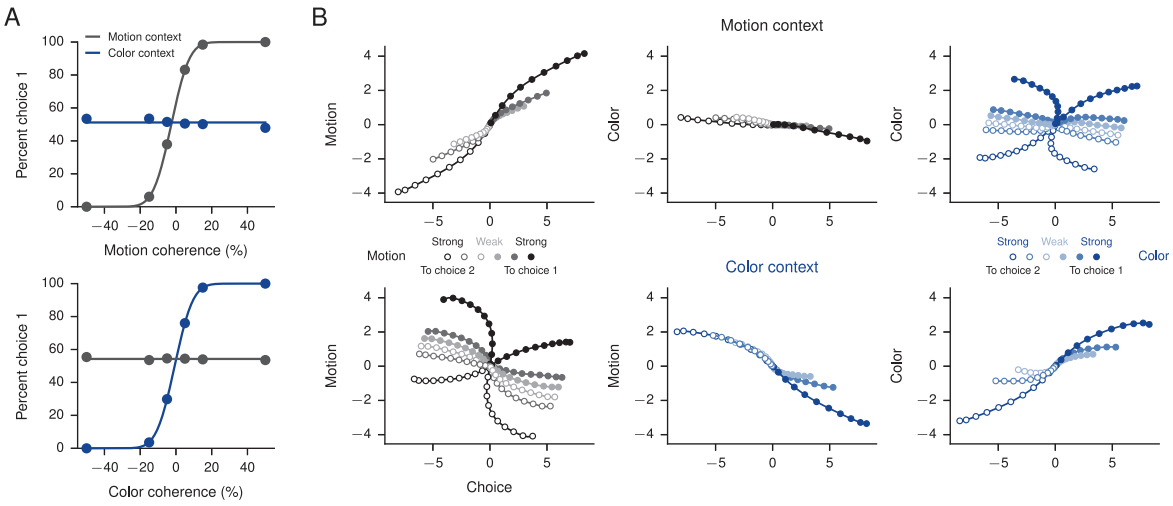


Figure 5.5: **Context-dependent integration task.** (**a**) Psychometric function, showing the percentage of trials on which the network chose choice 1 as a function of the signed motion (upper) and signed color (lower) coherence in motion-context (black) and color-context (blue) trials. (**b**) Average population responses in state space during the stimulus period, projected to the 3-dimensional subspace capturing variance due to choice, motion, and color as in Mante et al. (2013). Only correct trials were included. The task-related axes were obtained through a linear regression analysis. Note that "choice" here has a unit-specific meaning that depends on the preferred choice of the unit as determined by the selectivity index (Eq 5.31). For both motion (black) and color (blue), coherences increase from light to dark. Upper plots show trials during the motion context, and lower plots show trials during the color context.

Previously, population responses in the monkey prefrontal cortex were studied by representing them as trajectories in neural state space (Mante et al., 2013). This was done by using linear regression to define the four orthogonal, task-related axes of choice, motion, color, and context. The projection of the population responses onto these axes reveals how the different task variables are reflected in the neural activity. **Fig. 5.5b** shows the results of repeating this analysis (Mante et al., 2013) with the trained network during the stimulus period. The regression coefficients (**Fig. 5.6b**) reveal additional relationships between the task variables, which in turn reflect the mixed selectivity of individual

units to different task parameters as shown by sorting and averaging trials according to different criteria (**Fig. 5.6a**).

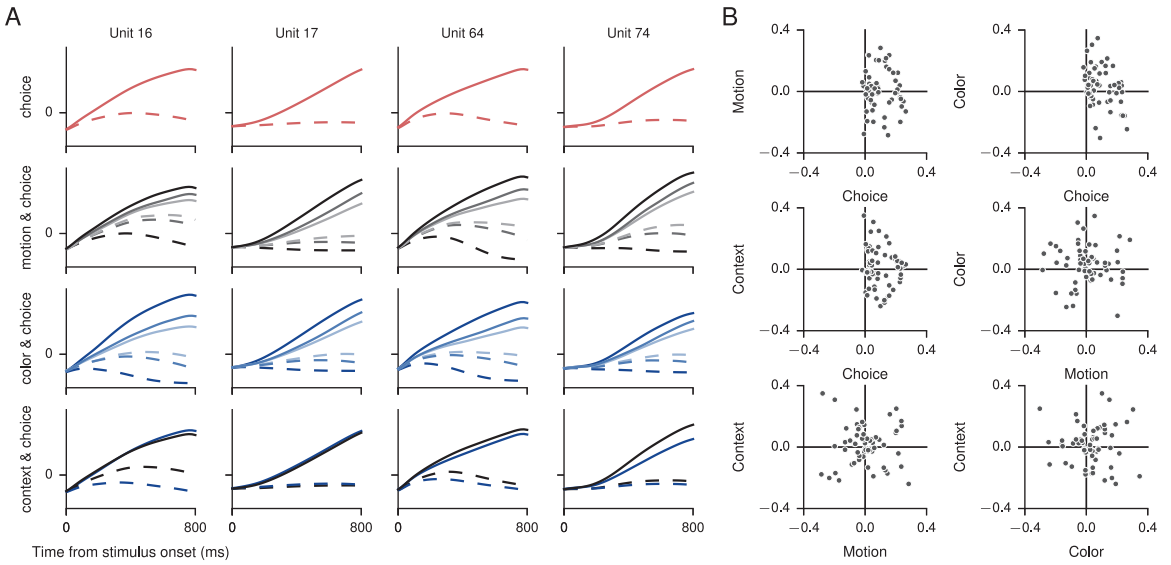


Figure 5.6: **Context-dependent integration task analysis.** (**a**) Normalized responses of four recurrent units during the stimulus period show mixed representation of task variables. Solid lines indicate the preferred choice and dashed lines the nonpreferred choice of each unit. (**b**) Denoised regression coefficients from the linear regression analysis. By definition, the coefficients for choice are almost exclusively positive.

As a proof of principle, we trained an additional network that could perform the same task but consisted of separate "areas," with one area receiving inputs and the other sending outputs (**Fig. 5.7b**), which can be compared to the unstructured connectivity of the original network (**Fig. 5.7a**). Here each area is conceived of as a cortical area containing a group of inhibitory units that only project locally to excitatory and inhibitory units in the same area. Thus there are no interareal connections originating from inhibitory units. The "sensory" area that receives inputs sends dense, "long-range" excitatory feedforward connections to the "motor" area from which outputs are read out, and receives "sparse" (connection probability 0.2) excitatory feedback projections from the motor area. This example illustrates the promise of using RNNs to explore how large-scale function may arise in the brain.

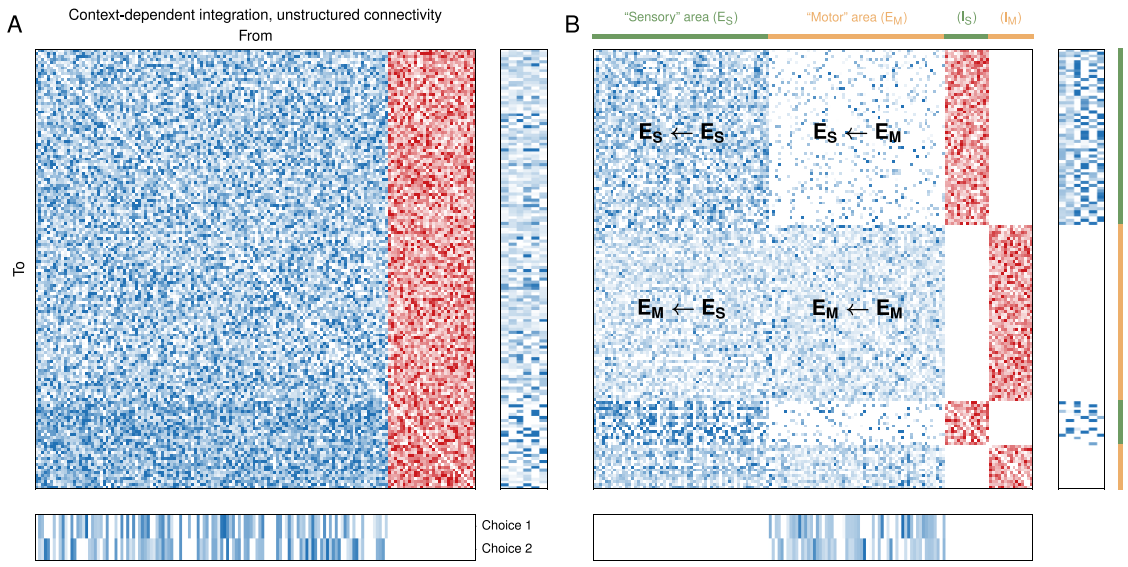


Figure 5.7: **Constraining the connectivity.** Connectivity after training for the context-dependent integration task (**Fig. 5.5**), when the connection matrix is (**a**) unstructured and (**b**) structured. Both networks consist of 150 units (120 excitatory, 30 inhibitory). In B the units are divided into two equal-sized "areas," each with a local population of inhibitory units ($I_S$ and $I_M$) that only project to units in the same area. The "sensory" area (green) receives excitatory inputs and sends dense, "long-range" excitatory feedforward connections $E_M \leftarrow E_S$ to the "motor" area (orange) from which the outputs are read out. The sensory area receives sparse excitatory feedback projections $E_S \leftarrow E_M$ from the motor area.

### 5.3.3 Multisensory integration task

The multisensory integration task of Raposo, Kaufman and Churchland (2014) also presents the animal—rats, in this case—with two sources of information. In contrast to the previous task, however, in the multisensory integration task it is advantageous for the animal to integrate both sources of information when they are available. Specifically, visual flashes and auditory clicks were presented at rates between 9 events/sec and 16 events/sec, and the animal was required to determine whether the inputs were below or above the threshold of 12.5 events/sec. When both visual and auditory inputs were present, they were congruent (presented at the same rate). A network trained for this task is also given one or more congruent inputs, and can improve its performance by combining both inputs when they are available (**Fig. 5.8a,b**). The network contains 150 units, 120 of which are excitatory and 30 inhibitory. A third of the units in the network (both excitatory and inhibitory) received only visual input, another third only auditory input, and the remaining third did not receive any input. Outputs were read out from the entire excitatory population.

The training was again mostly similar to the (fixed-duration) single-stimulus perceptual decision-making task, except for the presence of two congruent inputs on multisensory trials. However, in the present task the network must determine whether the given input is larger or smaller than an arbitrary strength, in contrast to the previous integration tasks where two inputs are compared to each other. As a result, giving the network both positively tuned (increasing function of event rate) and negatively tuned (decreasing function of event rate) inputs (Miller et al., 2003) greatly improved training. Although gradient-descent training can find a solution when the inputs are purely positively tuned, this results in much longer training times and more idiosyncratic unit activities. This il-

lustrates that, while RNN training methods are powerful, they must be supplemented with knowledge gained from experiments and previous modeling studies. As in experimentally recorded neurons, the units of the network exhibit heterogeneous responses, with some units showing selectivity to choice, others to modality, and still others showing mixed selectivity (**Fig. 5.8c**).

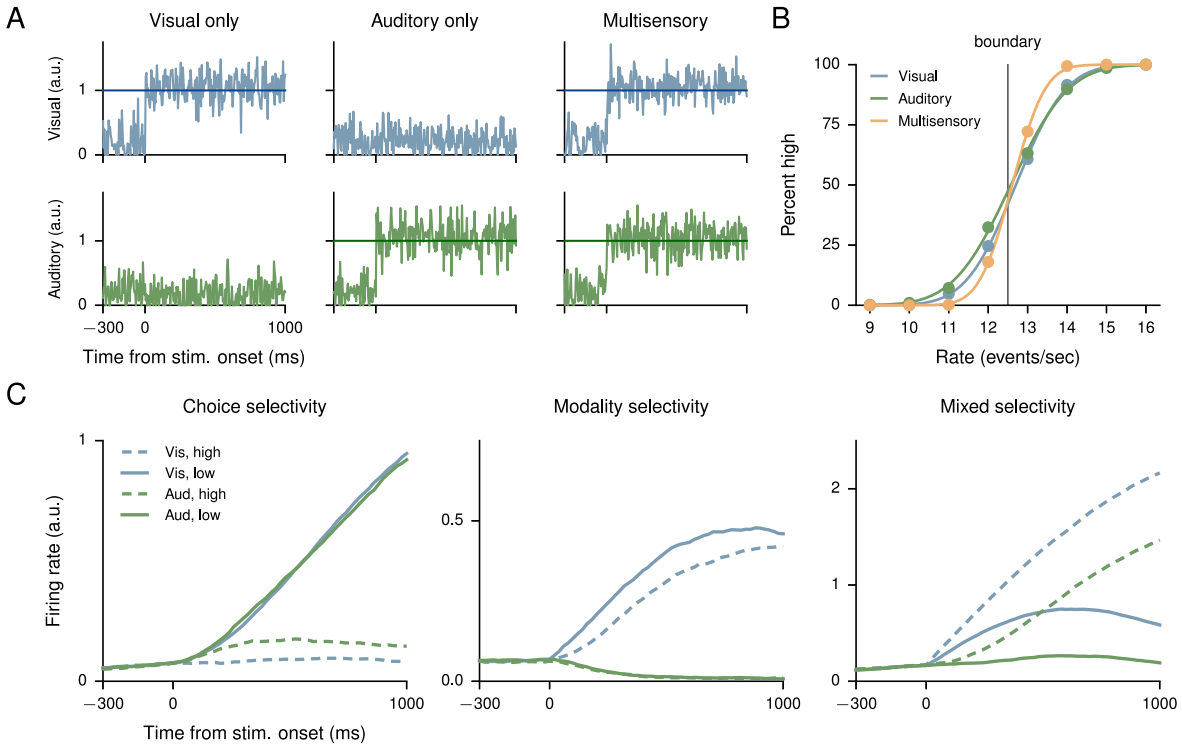


Figure 5.8: **Multisensory integration task.** (**a**) Example inputs for visual only (left), auditory only (middle), and multisensory (both visual and auditory, right) trials. Network units receive both positively tuned (increasing function of event rate) and negatively tuned (decreasing function of event rate) inputs; panels here show positively tuned input corresponding to a rate of 13 events/sec, just above the discrimination boundary. As in the single-stimulus perceptual decision-making task, the outputs of the network were required to hold low during "fixation" (before stimulus onset), then the output corresponding to a high rate was required to hold high if the input was above the decision boundary and low otherwise, and vice versa for the output corresponding to a low rate. (**b**) Psychometric functions (percentage of choice high as a function of the event rate) for visual, auditory, and multisensory trials show multisensory enhancement. (**c**) Sorted activity on visual only and auditory only trials for three units selective for choice (high vs. low, left), modality (visual vs. auditory, middle), and both (right).

The context-dependent and multisensory integration tasks represent the two end-

cases of when two separate sources of information are available for making a decision. It is of great interest for future inquiry how the *same* network or set of networks may switch from completely ignoring one input to using both inputs to make the optimal decision depending on the task.

### 5.3.4 Parametric working memory task

One of the most important—and therefore one of the most widely studied—cognitive functions is working memory, the ability to maintain and manipulate information for several seconds during the planning and execution of a task (Wang, 1999; Barak and Tsodyks, 2014). Working memory has notably been studied in the context of both oculomotor parametric working memory (Funahashi, Bruce and Goldman-Rakic, 1989) and vibrotactile frequency discrimination (Romo et al., 1999; Barak, Tsodyks and Romo, 2010), and here we trained a network to perform a task based on the frequency discrimination task. In this task, two temporally separated stimuli, represented by constant inputs whose magnitudes are proportional to the frequency (**Fig. 5.9a**), are presented and the network must determine which of the two is of higher frequency. This requires the network to remember the frequency of the first input $f_1$ throughout the 3-second delay period in order to compare to the second input $f_2$ at the end of the delay period. The network contains a total of 500 units (400 excitatory, 100 inhibitory), with a connection probability of 0.1 from excitatory units to all other units and 0.5 from inhibitory units to all other units; these connection probabilities are consistent with what is known for local microcircuits in cortex (Fino and Yuste, 2011; Karnani, Agetsuma and Yuste, 2014). During training only, the delay was varied by uniformly sampling from the range 2.5–3.5 seconds. As in the multisensory integration task, because the network must compare a single input

against itself (rather than comparing two simultaneous inputs to each other), it is helpful for the network to receive both positively tuned and negatively tuned inputs.

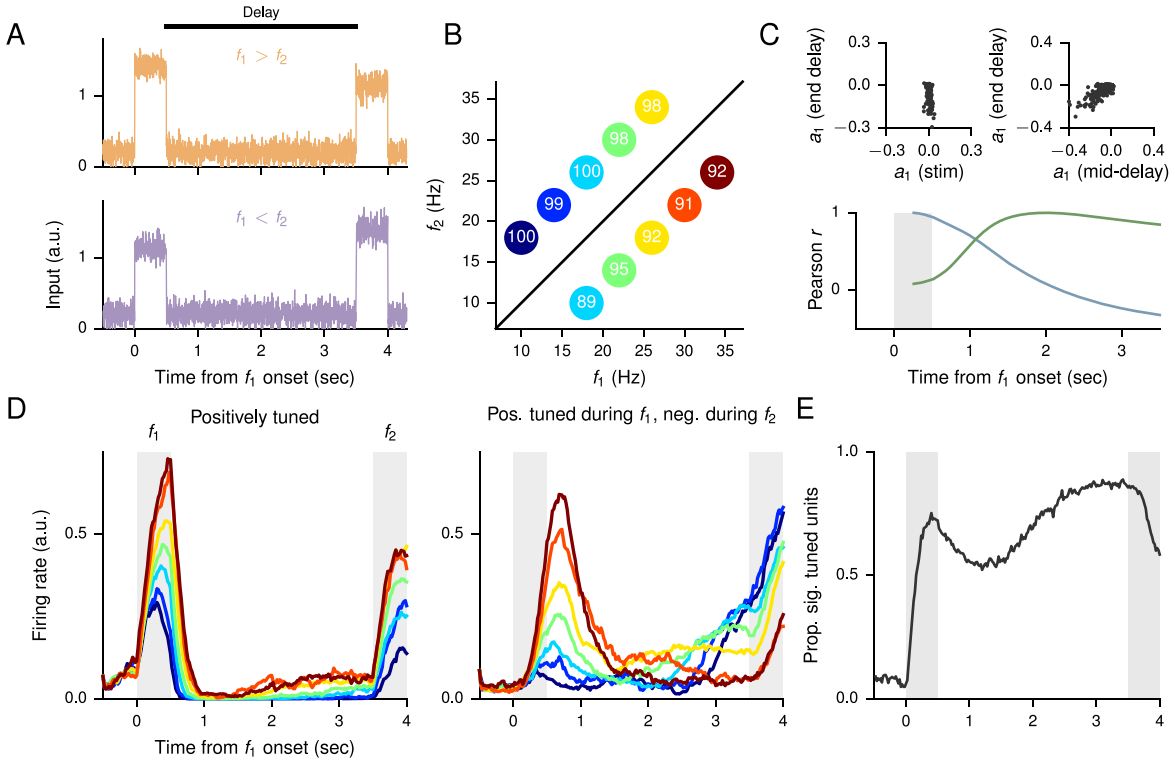


Figure 5.9: **Parametric working memory task.** (**a**) Sample positively tuned inputs, showing the case where $f_1 > f_2$ (upper) and $f_1 < f_2$ (lower). Recurrent units also receive corresponding negatively tuned inputs. (**b**) Percentage of correct responses for different combinations of $f_1$ and $f_2$. This plot also defines the colors used for each condition, labeled by $f_1$, in the remainder of the figure. Due to the overlap in the values of $f_1$, there are 7 distinct colors representing 10 trial conditions. (**c**) Lower: Correlation of the tuning $a_1$ (see text) at different time points to the tuning in the middle of the first stimulus period (blue) and middle of the delay period (green). Upper: The tuning at the end of delay vs. middle of the first stimulus (left) and the end of delay vs. middle of the delay (right). (**d**) Single-unit activity for a unit that is positively tuned for $f_1$ during both stimulus periods (left), and for a unit that is positively tuned during the first stimulus period but negatively tuned during the second stimulus period (right). (**e**) Proportion of significantly tuned units based on a simple linear regression of the firing rates as a function of $f_1$ at each time point.

The network's performance on each condition is shown in **Fig. 5.9b**. Based on the experimental results, we trained the network until the lowest percentage of correct responses in any condition exceeded 85%; for most conditions the performance is much higher (Barak et al., 2013). Several different types of behavior are observed in the unit

activities. For instance, some units are positively tuned for the frequency $f_1$ during both stimulus periods (**Fig. 5.9d, left**). Other units are positively tuned for $f_1$ during the first stimulus period but negatively tuned during the second (**Fig. 5.9d, right**); the switch can occur at various times during the delay. Following Barak et al. (2013), we performed a simple linear analysis of the tuning properties of units at different times by fitting the firing rate at each time point to the form $r(t) = a_0(t) + a_1(t)f_1$. The results are presented in **Fig. 5.9c**, which shows the correlation of $a_1$ between different time points across the population, and **Fig. 5.9e**, which shows the percentage of significantly tuned (two-sided $p$-value $< 0.05$) units at different times. The latter shows trends similar to those observed in monkeys.

### 5.3.5 Eye-movement sequence execution task

An experimental paradigm that is qualitatively very different from the previous examples involves the memorized execution of a sequence of motor movements, and is inspired by the task of Averbeck and Lee (2007). An important difference from a modeling point of view in this case is that, unlike in previous tasks where we interpreted the outputs as representing a decision variable between two choices, here we interpret the network's two outputs to be the $x$ and $y$-coordinates corresponding to the monkey's eye position on the screen. After maintaining fixation on the central dot for 1 second, the task is to execute a sequence of three eye movements and hold for 500 ms each (**Fig. 5.10a**). For each movement, two targets are presented as inputs to indicate the possible moves in addition to the current dot; although the targets could be presented in a more realistic manner—in a tuning curve-representation, for instance—here we use the simple encoding in which each input corresponds to a potential target location. Throughout the trial, an additional

input is given that indicates which sequence, out of a total of 8, is being executed (**Fig. 5.10b**).

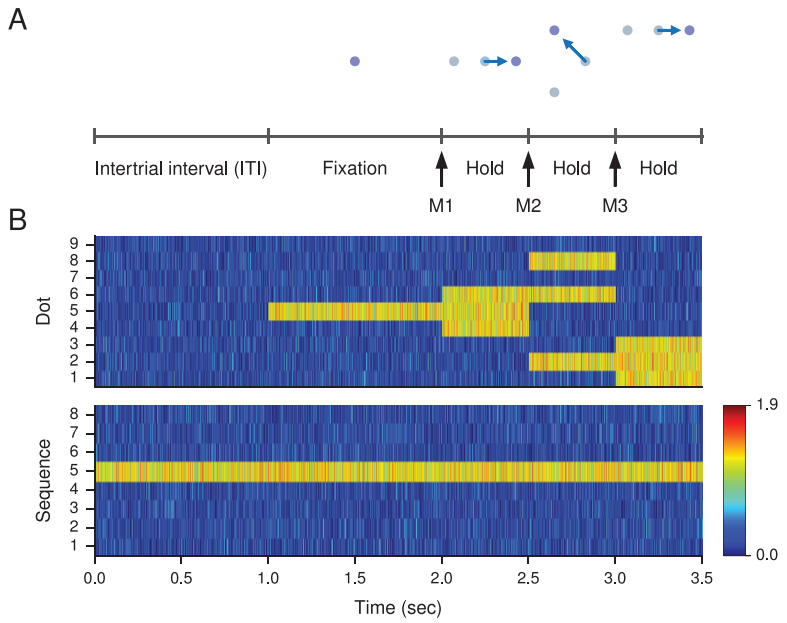


Figure 5.10: **Eye-movement sequence execution task.** (**a**) Task structure (for Sequence 5) and (**b**) sample inputs to the network. During the intertrial interval (ITI) the network receives only the input indicating the current sequence to be executed. Fixation is indicated by the presence of a fixation input, which is (the central) one of 9 possible dot positions on the screen. During each movement, the current dot plus two possible target dots appear.

For this task we trained a 200-unit (160 excitatory, 40 inhibitory) network on a trial-by-trial basis, i.e., the network parameters were updated after each trial. This corresponds to setting the gradient minibatch size to 1. Moreover, the network was run "continuously," without resetting the initial conditions for each trial (**Fig. 5.11**). During the intertrial interval (ITI), the network returns its eye position to the central fixation point from its location at the end of the third movement, so that the eye position is in the correct position for the start of the next fixation period. This occurs even though the target outputs given to the network during training did not specify the behavior of the outputs during the ITI, which is interesting for future investigation of such networks' ability to learn tasks with minimal supervision.

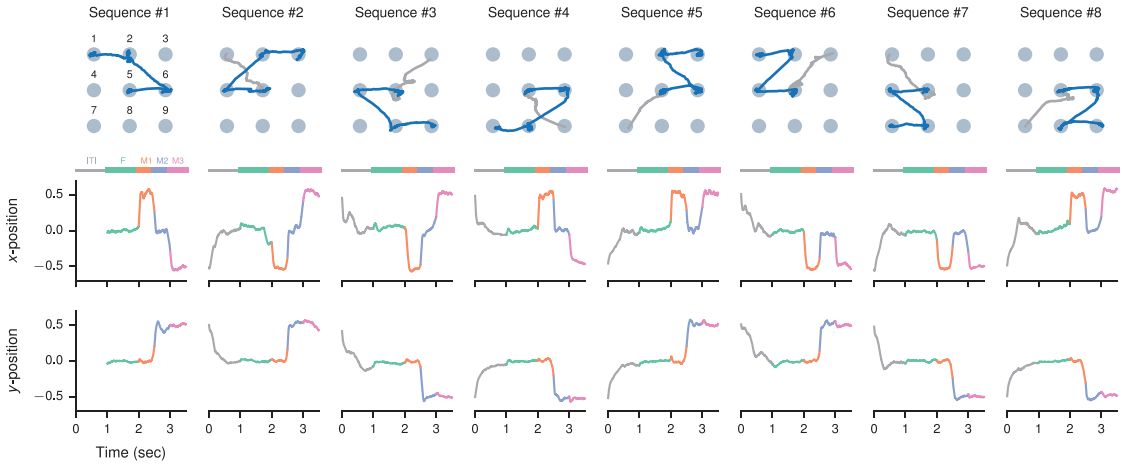


Figure 5.11: **Example run of the network.** Example run in which the network continuously executes each of the 8 sequences once in a particular order; the network can execute the sequences in any order. Each sequence is separated by a 1-second ITI during which the eye position returns from the final dot in the previous trial to the central fixation dot. Upper: Eye position in "screen" coordinates. Lower: $x$ and $y$-positions of the network's outputs indicating a point on the screen. Note the continuity of dynamics across trials.

During training, each sequence appeared once in a block of 8 randomly permuted trials. Here we used a time constant of $\tau = 50$ ms to allow faster transitions between dots. For this task only, we used a smaller recurrent noise of $\sigma_{\text{rec}} = 0.01$ because the output values were required to be more precise than in previous tasks, and did not limit readout to excitatory units to allow for negative coordinates. We note that, in the original task of Averbeck and Lee (2007) the monkey was also required to infer the sequence it had to execute in a block of trials, but we did not implement this aspect of the task. Instead, the sequence was explicitly indicated by a separate set of inputs.

Because the sequence of movements are organized hierarchically—for instance, the first movement must decide between going left and going right, the next movement must decide between going up and going down, and so forth—we expect a hierarchical trajectory in state space. This is confirmed by performing a principal components analysis and projecting the network's dynamics onto the first two principal components (PCs) computed across all conditions (**Fig. 5.12**).

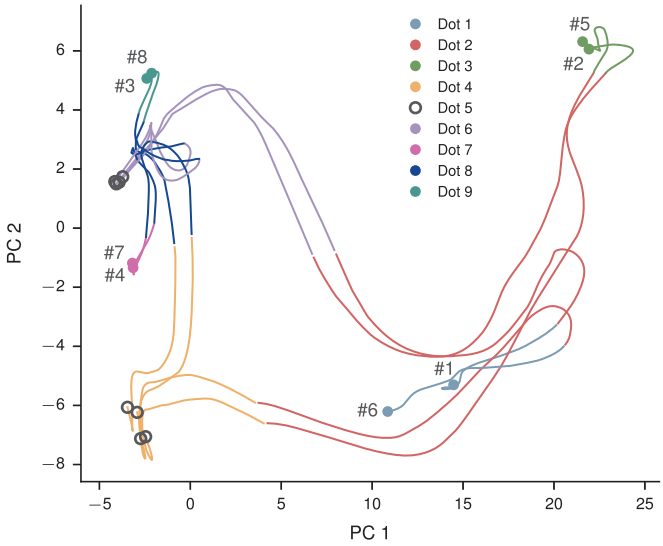


Figure 5.12: **State-space trajectories.** State-space trajectories during the three movements M1, M2, and M3 for each sequence, projected on the first two principal components (PCs) (71% variance explained, note the different axis scales). The network was run with zero noise to obtain the plotted trajectories. The hierarchical organization of the sequence of movements is reflected in the splitting off of state-space trajectories. Note that all sequences start at fixation, or dot 5 (black), and are clustered here into two groups depending on the first move in the sequence.

## 5.4 Discussion

In this work we have described a framework for gradient descent-based training of excitatory-inhibitory RNNs, and demonstrated the application of this framework to tasks inspired by well-known experimental paradigms in systems neuroscience.

Unlike in machine learning applications, our aim in training RNNs is not simply to maximize the network's performance, but to train networks so that their performance matches that of behaving animals while both network activity and architecture are as close to biology as possible. We have therefore placed great emphasis on the ability to easily explore different sets of constraints and regularizations, focusing in particular on "hard" constraints informed by biology. The incorporation of separate excitatory and inhibitory populations and the ability to constrain their connectivity is an important step in this direction, and is the main contribution of this work.

The framework described in this work for training RNNs differs from previous studies (Mante et al., 2013; Sussillo et al., 2015) in several other ways. In this work we use threshold (rectified) linear units for the activation function of the units. Biological neurons rarely operate in the saturated firing-rate regime, and the use of an unbounded nonlinearity obviates the need for regularization terms that prevent units from saturating (Sussillo et al., 2015). Despite the absence of an upper bound, all firing rates nevertheless remained within a reasonable range. We also favor first-order SGD optimization over second-order HF methods. This is partly because of SGD's widely acknowledged effectiveness in current approaches to machine learning, but also because gradient descent, unlike HF, allows for trial-by-trial learning and may ultimately be more easily related to synaptic learning rules in the brain (Bengio, Lee, Bornschein, Mesnard and Lin, 2015; Bengio, Mesnard, Fischer, Zhang and Wu, 2015).

Eqs 5.1-5.3 are a special case of the more general set of equations for RNNs, which in turn represent only one of many possible RNN architectures. For instance, machine learning applications typically employ a type of RNN known as Long Short-Term Memory (LSTM), which uses multiplicative gates to facilitate learning of long-term dependencies and currently represents one of the most powerful methods for solving sequence-related problems (Hochreiter and Schmidhuber, 1997). For reasons of biological interpretation, in our implementation we only consider generalizations that retain the "traditional" RNN architecture given by Eqs 5.1-5.3. These generalizations include additive bias terms in recurrent and output units (corresponding to variable thresholds), different time constants for each unit (e.g., faster inhibitory units), correlated noise (Renart et al., 2010), and other types of nonlinearities besides simple rectification (e.g., supralinear (Rubin, Van Hooser and Miller, 2015) or saturating $f$-$I$ curves) for either recurrent units or outputs. We found

that biases, though not used for the networks in this work, can improve training in some situations by endowing the network with greater flexibility. The choice of output nonlinearity can be particularly relevant when considering the precise meaning of the outputs, such as whether the outputs are considered a decision variable, probability distribution, or eye position. Probability output models are useful, for instance, when the animal's confidence about its decision is of interest in addition to its actual decision.

Several works (Mante et al., 2013; Barak et al., 2013; Carnevale et al., 2015) have now demonstrated the value of trained RNNs in revealing circuit mechanisms embedded in large neural populations. In addition to the pioneering work on uncovering a previously unknown selection mechanism for context-dependent integration of sensory inputs in Mante et al. (2013), work reported in Carnevale et al. (2015) used trained RNNs to reveal possible dynamical implementations of response criterion modulation in a perceptual detection task under temporal uncertainty. Yet, more recent methods for training networks have not been widely available or easily accessible to the neuroscience community. We have endeavored to change this by providing an easy-to-use but flexible implementation of our framework that facilitates further modifications and extensions. For the tasks featured in this work, the amount of time needed for training was relatively short and largely consistent across different initializations (Fig. 5.13), and could be made even shorter for exploratory training by reducing the network size and noise level. Although further improvements can be made, our results already demonstrate that exploratory network training can be a practical and useful tool for neuroscientists. Moreover, while the present learning rule is not biologically plausible, it is of interest whether the behavioral trajectory of learning can be made similar to that of animals learning the same tasks. In particular, the question of how many trials are needed to learn a given task in model

RNNs and animals merits further investigation.

Many interesting and challenging questions remain. Although RNNs of rate units often provide a valuable starting point for investigating both the dynamical and neural computational mechanisms underlying cognitive functions, they will not always be the most appropriate level of description for biological neural circuits. In this work we have not addressed the question of how the firing rate description given by RNN training can be properly mapped to the more realistic case of spiking neurons, and indeed it is not completely clear, at present, how spiking neurons may be directly trained for general tasks using this type of approach. In this work we have only addressed tasks that could be formulated in the language of supervised learning, i.e., the correct outputs were explicitly given for each set of inputs. Combining RNN training with reinforcement learning methods (Sutton and Barto, 1998; Bakker, 2002; Roelfsema and van Ooyen, 2005) will be essential to bringing network training closer to the reward-based manner in which animals are trained. Despite limitations, particularly on the range of tasks that can be learned, progress on training spiking neurons with STDP-type rules and reinforcement learning is promising (Izhikevich, 2007; Potjans, Morrison and Diesmann, 2009; Neymotin et al., 2013), and future work will incorporate such advances. Other physiologically relevant phenomena such as bursting, adaptation, and oscillations are currently not captured by our framework, but can be incorporated in the future; adaptation, for example, can be included in phenomenological form appropriate to a rate model (Benda and Herz, 2003; Engel and Wang, 2011).

We have also not addressed what computational advantages are conferred, for example, by the existence of separate excitatory and inhibitory populations, instead taking it as a biological fact that must be included in models of animal cognition. Indeed, al-

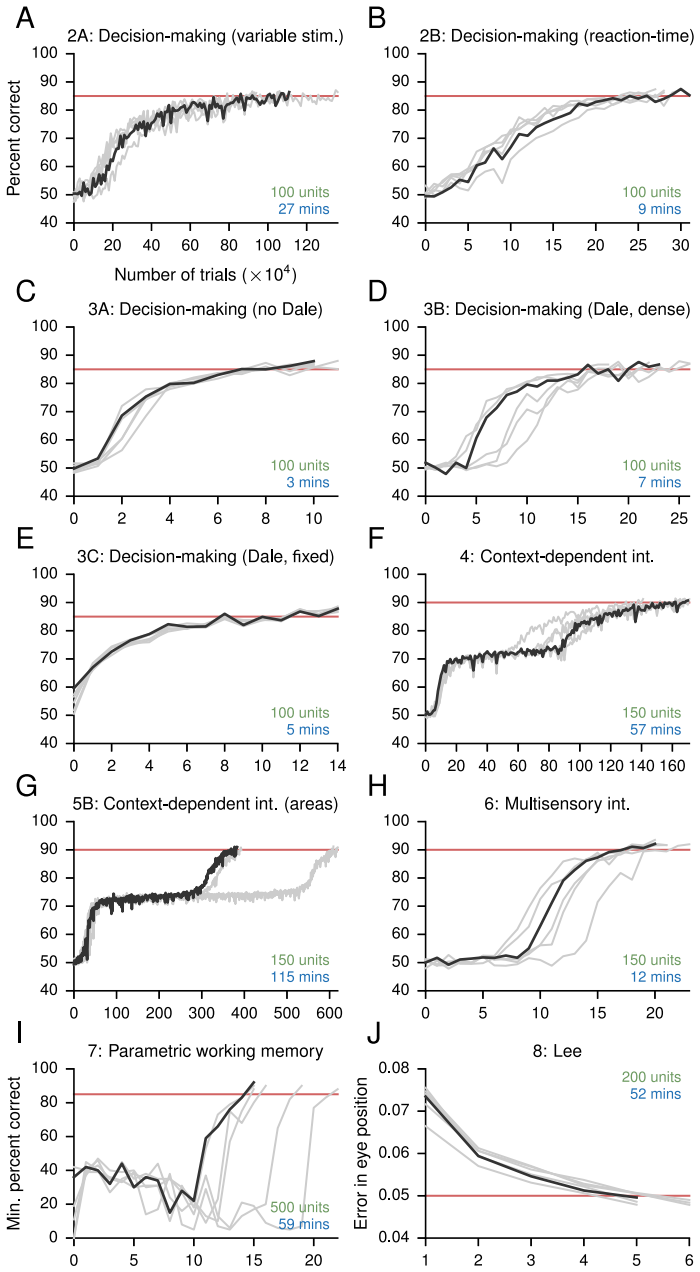


Figure 5.13: **Estimated performance during training for networks in the Results.** (**a**)-(**i**) Percentage of correct responses. (**j**) Error in eye position. For each network the relevant figure in the main text and a brief description are given. Black lines are for the networks shown in the main text, while gray lines show the performance for 5 additional networks trained for the same tasks but using different initial weights. Red lines indicate the target performance; training terminated when the mean performance on several (usually 5) evaluations of the validation dataset exceeded the target performance. In I the target performance indicates the minimum, rather than mean, percentage of correct responses across conditions. The number of recurrent units (green) is indicated for each network. The number of minutes (in "real-time") needed for training (blue) are estimates for a MacBook Pro running OS X Yosemite 10.10.4, with a 2.8 GHz Intel Core i7 CPU and 16 GB 1600 MHz DDR3 memory. GPUs were not used in the training of these networks.

though our discussion has focused on the distinction between excitatory and inhibitory neurons, the functional role of inhibitory units may only become apparent when the full diversity of excitatory and inhibitory neuronal morphology and physiology, their layer and type-specific distribution and connectivity (Thomson et al., 2002; Binzegger, Douglas and Martin, 2004), and domain-specific (e.g., dendritic versus somatic) targeting of excitatory pyramidal cells by interneurons (Markram et al., 2004; Wang et al., 2004; Pfeffer et al., 2013; Jiang et al., 2015) in the brain are taken into account. Some of these phenomena can already be implemented in the framework by fixing the pattern of connectivity between groups of units (corresponding, for example, to different layers in a cortical column), and future work will explore the implications of such structure on network dynamics.

Finally, although we have performed a few basic analyses of the resulting networks, we have not addressed the detailed mechanisms by which networks accomplish their tasks. In this regard, although state-space analyses of fixed and "slow" points (Sussillo and Barak, 2013) are illuminating they do not yet explain how the network's connectivity, combined with the nonlinear activation functions, lead to the observed neural trajectories. Discovering general methods for the systematic analysis of trained networks remains one of the most important areas of inquiry if RNNs are to provide useful insights into the operation of biological neural circuits. As a platform for theoretical investigation, trained RNNs offer a unified setting in which diverse cognitive computations and mechanisms can be studied. Our results provide a valuable foundation for tackling this challenge by facilitating the generation of candidate networks to study, and represent a fruitful interaction between modern machine learning and neuroscience.

# Chapter 6

# Task representations in networks trained for many cognitive tasks

## 6.1 Introduction

The prefrontal cortex is important for numerous cognitive functions (Fuster, 2015; Miller and Cohen, 2001; Wang, 2013), partly because of its central role in task representation (Wallis, Anderson and Miller, 2001; Sakai, 2008; Cole et al., 2011; Tschentscher, Mitchell and Duncan, 2017). Electrophysiological experiments using behaving animals reported prefrontal neurons that are either selective for different aspects of a given task (Hanes, Patterson and Schall, 1998; Padoa-Schioppa and Assad, 2006) or functionally mixed (Rigotti et al., 2013; Mante et al., 2013). Much less is known about functional specialization of task representations at the neuronal level. Imagine a single-neuron recording that could be carried out with animals switching between many different tasks. Is each task supported by a "private" set of neurons, or does each task involve every neuron in the network, or somewhere in between? If two tasks require a common underlying cognitive

process, such as working memory or decision making, what would be the relationship between their neural representations? In other words, what would be the "neural relationship" between this pair of tasks? Would the two tasks utilize a shared neural substrate?

Humans readily learn to perform many cognitive tasks in a short time. By following verbal instructions such as "Release the lever only if the second item is not the same as the first," humans can perform a novel task without any training at all (Cole et al., 2011). A cognitive task is typically composed of elementary sensory, cognitive, and motor processes (Sakai, 2008). Performing a task without training requires composing elementary processes that are already learned into temporal sequences that enable correct performance on the new task. This property, called "compositionality," has been proposed as a fundamental principle underlying flexible cognitive control (Cole, Laurent and Stocco, 2013). Indeed, human studies have suggested that the representation of complex cognitive tasks in the lateral prefrontal cortex is compositional (Cole et al., 2011; Reverberi, Görgen and Haynes, 2012). However, these tasks involved verbal instructions; it is unknown whether non-verbal tasks commonly used in animal physiological experiments also display compositionality and whether relatively simple neural network models are sufficient to support compositional task structures.

These questions remain difficult to address with conventional experimental and modeling approaches. Experiments with laboratory animals have so far been largely limited to a single task at a time; on the other hand, human imaging studies lack the spatial resolution to address questions at the single neuron level. Therefore, the lack of neural recordings from animals performing many different tasks leaves unanswered important questions regarding how a single network represents and supports distinct tasks. Theo-

retically, designing a single neural circuit model capable of multiple tasks is challenging and virtually nonexistent. To tackle these problems, we took the approach of training recurrent neural networks (RNNs) (Mante et al., 2013; Zipser and Andersen, 1988; Song, Yang and Wang, 2016; Carnevale et al., 2015; Rajan, Harvey and Tank, 2016; Chaisangmongkon et al., 2017; Eliasmith et al., 2012). In this work, we trained a single RNN to perform 20 cognitive tasks. We found that after training the emerging task representations are organized in the form of clustering of recurrent units. Our network also makes numerous testable predictions regarding the neural relationship between pairs of cognitive tasks. Surprisingly, we found that compositionality of task representations emerges from training in our network model, which can be instructed to perform new tasks without further training. Our work provides a framework for investigating neural representations of task structures and neural relationships between tasks.

## 6.2 Results

### 6.2.1 Training neural networks for many cognitive tasks

To study how various cognitive tasks might be represented in a single neural circuit, we trained a recurrent neural network model (**Fig. 6.2**) to perform 20 tasks, most of which are commonly used in neurophysiological studies of nonhuman animals and crucial to our understanding of the neural mechanisms of cognition. The chosen set of tasks includes variants of memory-guided response (Funahashi, Bruce and Goldman-Rakic, 1989), simple perceptual decision making (Gold and Shadlen, 2007), context-dependent decision-making (Mante et al., 2013; Siegel, Buschman and Miller, 2015), multi-sensory integration (Raposo, Kaufman and Churchland, 2014), parametric working memory (Romo et al.,

| Task name | Abbreviation | Task family | Reference |
|---|---|---|---|
| Go | Go | Go | N/A |
| Reaction-time go | RT Go | Go | N/A |
| Delayed go | Dly Go | Go | (Funahashi, Bruce and Goldman-Rakic, 1989) |
| Anti-response | Anti | Anti | (Munoz and Everling, 2004) |
| Reaction-time anti-response | RT Anti | Anti | (Munoz and Everling, 2004) |
| Delayed anti-response | Dly Anti | Anti | (Munoz and Everling, 2004) |
| Decision making 1 | DM 1 | DM | (Gold and Shadlen, 2007) |
| Decision making 2 | DM 2 | DM | (Gold and Shadlen, 2007) |
| Context-dependent decision making 1 | Ctx DM 1 | DM | (Mante et al., 2013) |
| Context-dependent decision making 2 | Ctx DM 2 | DM | (Mante et al., 2013) |
| Multi-sensory decision making | MultSen DM | DM | (Raposo, Kaufman and Churchland, 2014) |
| Delayed decision making 1 | Dly DM 1 | Dly DM | (Romo et al., 1999) |
| Delayed decision making 2 | Dly DM 2 | Dly DM | (Romo et al., 1999) |
| Context-dependent delayed decision making 1 | Ctx Dly DM 1 | Dly DM | N/A |
| Context-dependent delayed decision making 2 | Ctx Dly DM 2 | Dly DM | N/A |
| Multi-sensory delayed decision making | MultSen Dly DM | Dly DM | N/A |
| Delayed match-to-sample | DMS | Matching | (Miller, Erickson and Desimone, 1996) |
| Delayed non-match-to-sample | DNMS | Matching | (Miller, Erickson and Desimone, 1996) |
| Delayed match-to-category | DMC | Matching | (Freedman and Assad, 2016) |
| Delayed non-match-to-category | DNMC | Matching | (Freedman and Assad, 2016) |

Table 6.1: Names and abbreviations of all tasks trained in the networks. Most of the trained tasks are derived from archetypal cognitive tasks used in non-human animal experiments. We grouped our tasks into five task families. We are not aware of experimental studies that investigated the Ctx Dly DM 1, Ctx Dly DM 2, or MultSen Dly DM tasks in non-human animals.

1999), inhibitory control (e.g., in anti-saccade) (Munoz and Everling, 2004), delayed match-to-sample (Miller, Erickson and Desimone, 1996), and delayed match-to-category (Freedman and Assad, 2016) tasks (**Table 1**, **Fig. 6.1**).

The recurrent network model emulates a “cognitive-type" cortical circuit such as the prefrontal cortex (Wang, 2013), which receives converging inputs from multiple sensory pathways and projects to downstream motor areas. We designed our network architecture to be general enough for all the tasks mentioned above, but otherwise as simple as possible to facilitate analysis. For every task, the network receives noisy inputs of three types: fixation, stimulus, and rule (**Fig. 6.2**). The fixation input indicates whether the network should "fixate" or respond (e.g. "saccade"). Thus the decrease in the fixation input provides a "go signal" to the network. The stimulus inputs consist of two modalities, each represented by a ring of input units that encodes a one-dimensional circular vari-

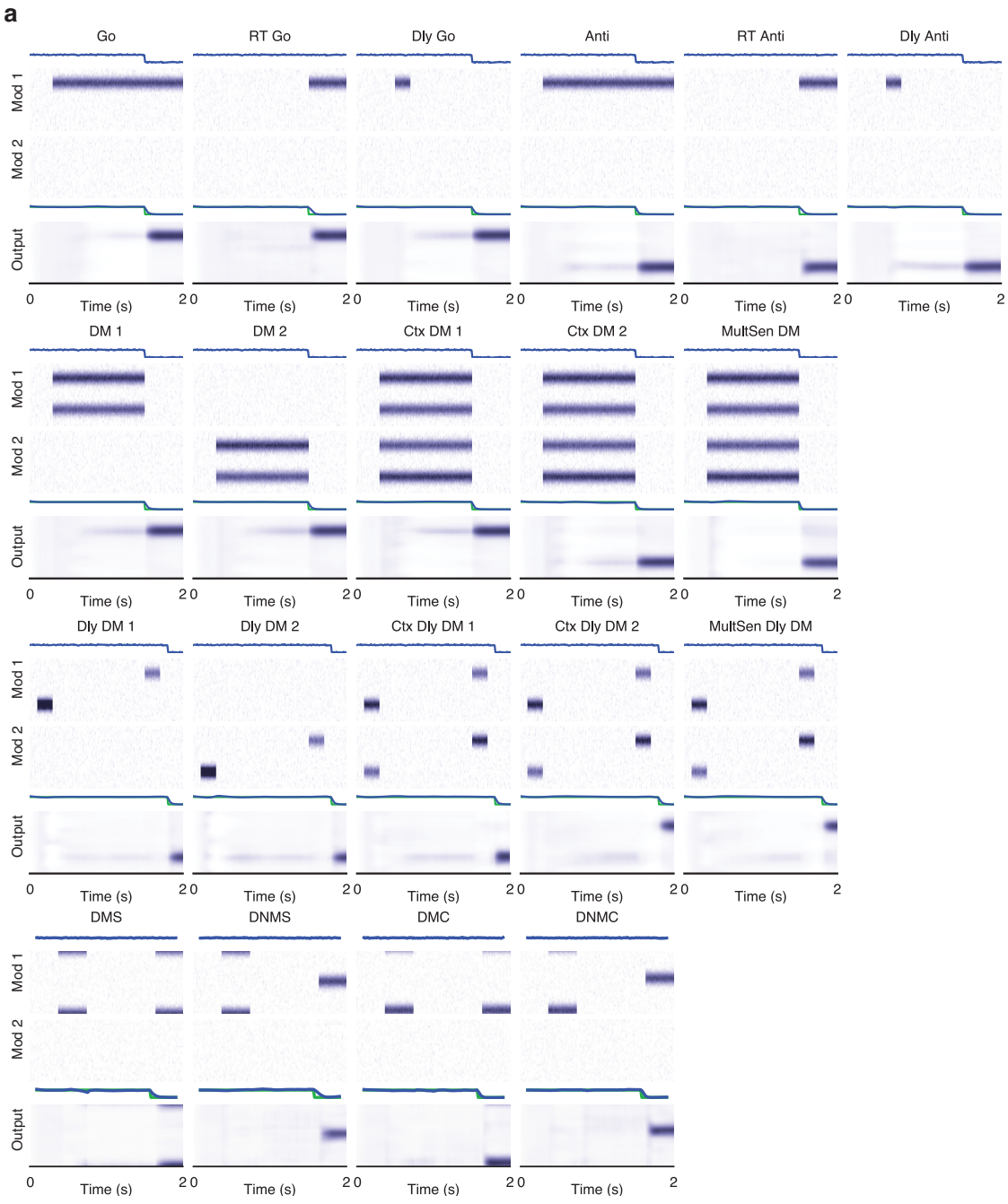


Figure 6.1: Sample trials from the 20 tasks trained. (**a**) Convention is the same as **Fig. 6.2**. Output activities are obtained from a sample network after training.

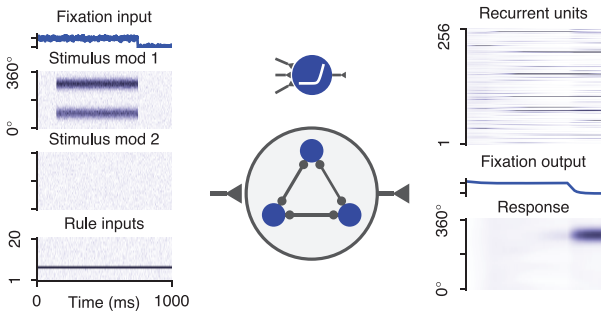


Figure 6.2: A recurrent neural network model. A recurrent neural network (middle) described by rate units receives inputs (left) encoding a fixation cue, stimuli from two modalities, and a rule signal (which instructs the system which task to perform in a given trial). The network has 256 recurrent units (top right), and it projects to a fixation output unit (which should be active when a motor response is unwarranted) and a population of units selective for response directions (right). All units in the recurrent network have non-negative firing rates. All connection weights and biases are modifiable by training using a supervised learning protocol.

able such as motion direction or color on a color wheel (Chaisangmongkon et al., 2017). A single rule input unit is activated in each trial, instructing the network on which task it is currently supposed to perform. The network projects to a fixation output unit and a group of motor units encoding the response direction. To mimic biological neurons, all units in our recurrent network receive private noise and have non-negative activities, imposed by a realistic neuronal input-output function (Abbott and Chance, 2005).

Before training, a network is incapable of performing any task. It is trained with supervised learning (Mante et al., 2013; Song, Yang and Wang, 2016), that modifies all connection weights (input, recurrent, and output) to minimize the difference between the network output and a desired (target) output. All tasks were randomly interleaved during training (at the end we will present results from sequential training). Below we show results obtained from networks of 256 recurrent units, and results are robust with respect to the exact network size. After training, a single network model achieved high behavioral performance across all tasks (**Fig. 6.3**). Furthermore, by conducting a battery of psychometric tests, we demonstrate that the network displays behavioral features con-

sistent with animal studies (**Fig. 6.5**). For instance, in perceptual decision-making tasks, the network achieves better performance with higher coherence and longer duration of the stimulus (**Fig. 6.4a**) (Gold and Shadlen, 2007), and it combines information from different sources to form decisions (**Fig. 6.4b**) (Raposo, Kaufman and Churchland, 2014). In working memory tasks, the network can maintain information throughout a delay period of up to five seconds (Funahashi, Bruce and Goldman-Rakic, 1989; Romo et al., 1999; Fuster, 2015).

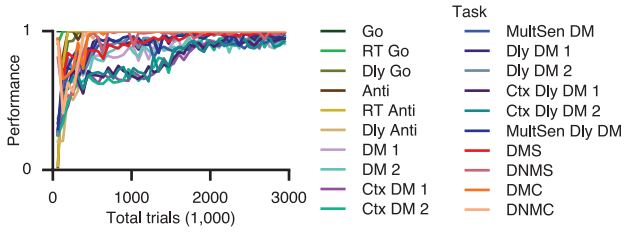


Figure 6.3: The network successfully learned to perform 20 tasks.

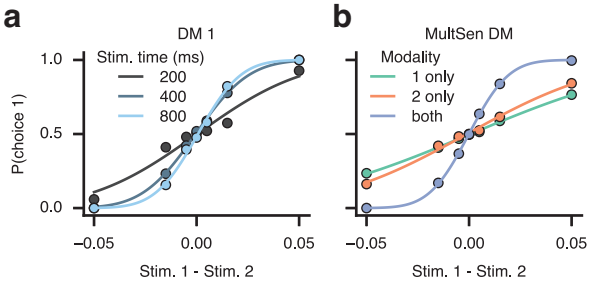


Figure 6.4: Psychometric curves in two decision making (DM) tasks. (**a**) Perceptual decision-making relies on temporal integration of information, as the network performance improves when the noisy stimulus is presented for a longer time. (**b**) In a multisensory integration task, the trained network combines information from two modalities to improve performance (compared with performance when information is only provided by a single modality).

### 6.2.2 Dissecting the circuit for the family of Anti tasks

For trained neural networks to be useful model systems for neuroscience, it is critical that we attempt to understand the circuit mechanism underlying the network computation (Sussillo and Barak, 2013). Here we demonstrate how a trained network could be

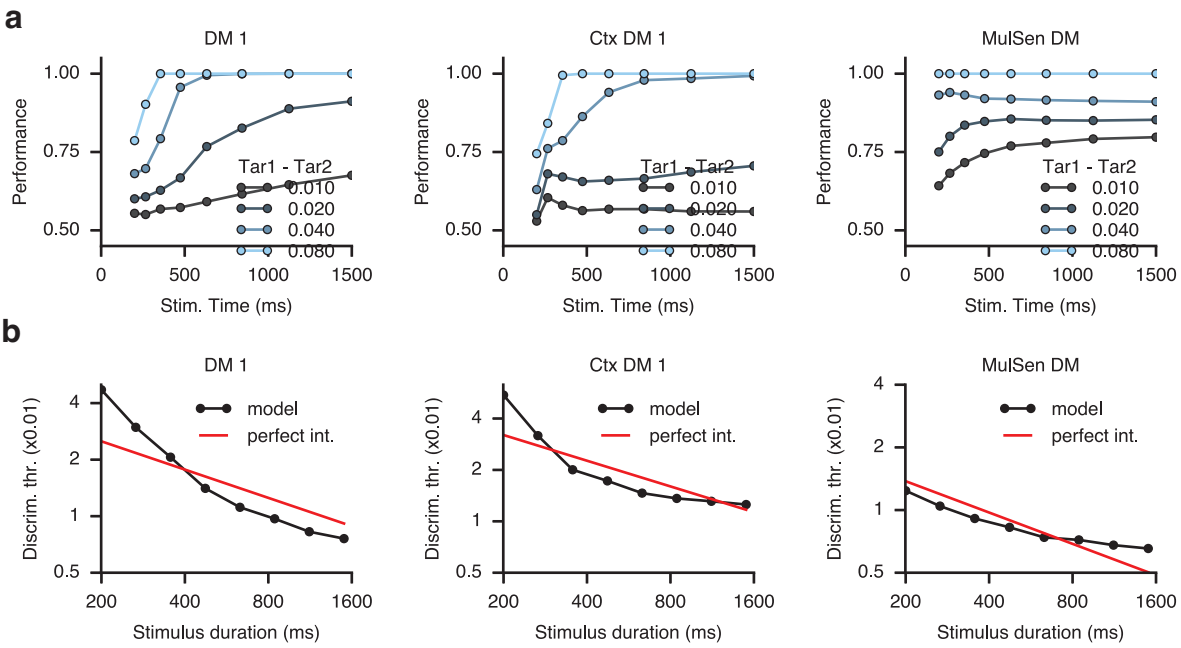


Figure 6.5: Psychometric tests for a range of tasks. (**a**) Decision making performances improve with longer stimulus presentation time and stronger stimulus coherence. (**b**) Discrimination thresholds decrease with longer stimulus presentation time. The discrimination thresholds are estimated by fitting cumulative Weibull functions.

dissected and analyzed in a sample family of cognitive tasks. Anti-response tasks are important tools to investigate voluntary action and inhibitory control (Munoz and Everling, 2004). These tasks require an anti-response, in the opposite direction from the more common pro-response towards a stimulus' location. Our set of tasks includes three tasks from the Anti task family (**Table 1**): the Anti, RT Anti, and Dly Anti tasks. We found that a subgroup of units emerged in a trained network, which we call Anti units (**Fig. 6.6a**). These units are primarily selective to stimuli in the Anti family of tasks. Inactivating or "lesioning" all Anti units at once resulted in a complete failure in performing the family of tasks that require an anti-response, but had essentially no impact on the performance of the other tasks (**Fig. 6.6b**).

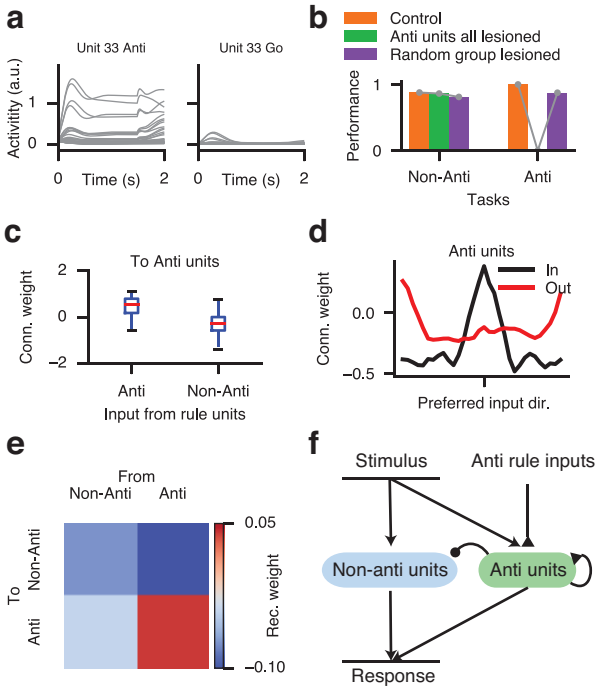


Figure 6.6: Dissecting the circuit for a family of tasks. (**a**) An example Anti unit, which is primarily selective in the Anti-family of tasks. Different traces show neural activities across stimulus conditions within a task. (**b**) After lesioning all Anti units together (green), the network can no longer perform any of the Anti tasks, while performance for other tasks remain intact. Instead, lesioning the same number of randomly selected units had a minor impact on the performance. (**c**) Anti units receive strong positive connections from rule units representing the Anti tasks but negative connections from non-Anti rule units. (**d**) Average connections from input units to Anti units (black) and those onto output units (red) display opposite preferred directions, thereby vector conversion (from pro- to anti-response) is realized. Both input and output connections are sorted by each unit's preferred input direction, defined as the stimulus direction represented by the strongest-projecting input unit. (**e**) Network wiring architecture that emerged from training, in which Anti units excite themselves and strongly inhibit other units. (**f**) Circuit diagram summarizing the neural mechanism of the Anti-family tasks.

Since we have access to all the information of the trained network, we next investigated the connection weights of Anti units to understand their roles. Anti units receive strong positive connection weights from the three rule input units representing Anti tasks (**Fig. 6.6c**), which explained why Anti units are only active during Anti tasks. Next, we studied the connection weights of Anti units with the stimulus-encoding input ring and the response-encoding output ring. For each Anti unit, the preferred input and output directions defined by the input and output connection weights are 180 degrees apart (**Fig.**

**6.6d**). These opposite preferred directions serve as the neural substrate for vector inversion (anti-mapping) required by Anti tasks. Finally, the Anti units strongly inhibit the rest of the recurrent units (Non-Anti units) through recurrent connections (**Fig. 6.6e**), suppressing a pro-response with inhibitory control. Thus, the circuit mechanism underlying Anti tasks in our trained network is delineated: A group of units emerge from training that are specialized for the anti-response process and are essential in every task that requires this process. The Anti rule inputs engage vector-inverting Anti units, which in turn exert inhibitory control over Non-Anti units (**Fig. 6.6f**).

### 6.2.3 Functional clusters encode subsets of tasks

The focus of our analysis was to examine the neural representation of tasks. After training, it is conceivable that each unit of the recurrent network is only selective in one or a few tasks, forming highly-specialized task representations. On the other hand, task representations may be completely mixed, where all units are engaged in every task. We sought to assess where our network lies on the continuum between these two extreme scenarios.

To quantify single-unit task representation, we need a measure of task selectivity that is general enough so it applies to a broad range of tasks, and at the same time simple enough so it can be easily computed. We propose a measure that we call Task Variance (see Online Methods). For each unit, the task variance for a given task is obtained by first computing the variance of neural activities across all possible stimulus conditions at a given time point, then averaging that variance across time (excluding the fixation epoch) (**Fig. 6.7a**). Task variance is agnostic about the task setup and can be easily computed in models and is also applicable to the analysis of experimental data.

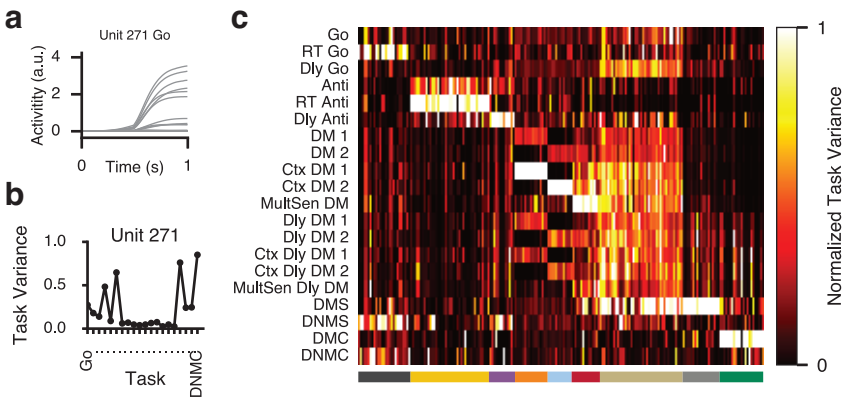


Figure 6.7: The emergence of functionally specialized clusters for task representation. (**a**) Neural activity of a single unit during an example task. Different traces correspond to different stimulus conditions. (**b**) Task variances across all tasks for the same unit. For each unit, task variance measures the variance of activities across all stimulus conditions. (**c**) Task variances across all tasks and active units, normalized by the peak value across tasks for each unit. Units form distinct clusters identified using the K-means clustering method based on normalized task variances. Each cluster is specialized for a subset of tasks. A task can involve units from several clusters. Units are sorted by their cluster membership, indicated by colored lines at the bottom.

By computing the task variance for all trained tasks, we can study how individual units are differentially selective in all the tasks (**Fig. 6.7b**). For better comparison across units, we normalized the task variance of each unit such that the maximum normalized variance over all tasks is one. By analyzing the patterns of normalized task variance for all active units, we found that units are self-organized into distinct clusters through learning (**Fig. 6.7c**, **Fig. 6.8**) (see Methods). We identified about 10 clusters in the network. Each cluster is mainly selective in a specific subset of tasks. To understand the causal role of these clusters, we lesioned each of them while monitoring the change in performance across all 20 tasks (**Fig. 6.9**). We found one cluster that is specialized for the Anti-family tasks, and it consists mainly of Anti units analyzed in **Fig. 6.6**. Another two clusters are specialized for decision-making tasks involving modality 1 and 2 respectively. Furthermore, clusters selective in the parametric working memory tasks (Dly DM task family) are

also selective in the perceptual decision making tasks (DM task family), indicating a common neural substrate for these two cognitive functions in our network (Wang, 2002). We can also study how units are clustered based on epoch variance, a measure that quantifies how selective units are in each task epoch (**Fig. 6.10**). One cluster of units presumably supports response generation, as it is highly selective in the response epoch but not the stimulus epoch. Our results indicate that the network successfully identified common sensory, cognitive, and motor processes underlying subsets of tasks, and through training developed units dedicated to the shared processes rather than the individual tasks.

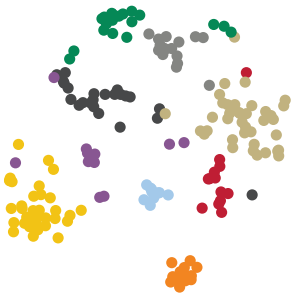

Figure 6.8: Visualization of the task variance map. For each unit, task variances across tasks form a vector that is embedded in the two-dimensional space using t-distributed Stochastic Neighbor Embedding (t-SNE). Units are colored according to their cluster membership.

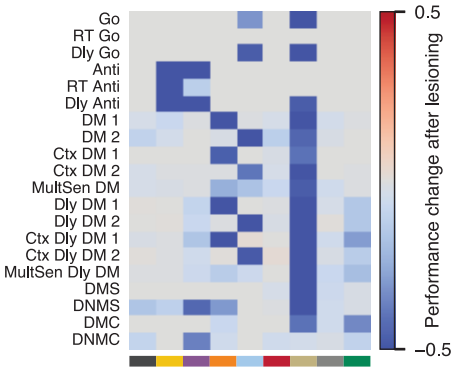


Figure 6.9: Change in performance across all tasks when each cluster of units is lesioned.

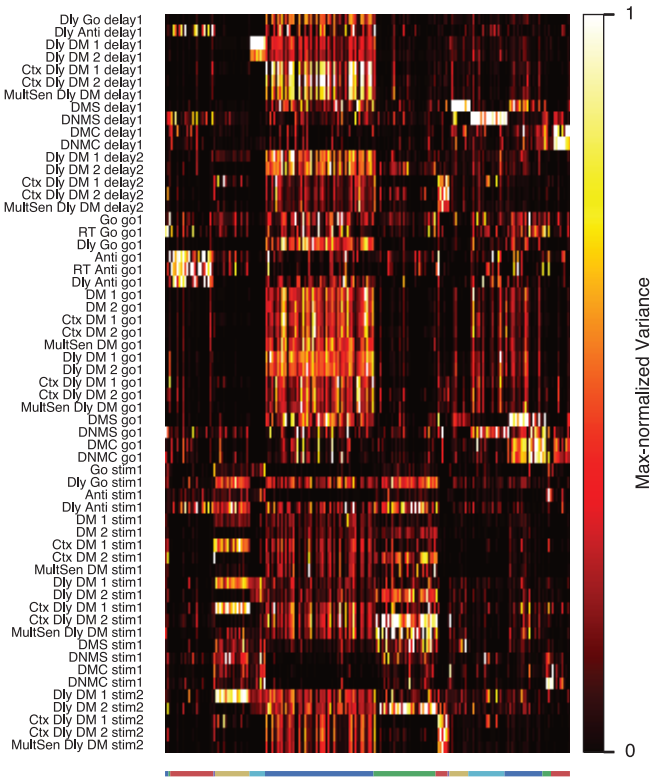


Figure 6.10: Epoch variances across all task epochs and active units. Epoch variance is computed in a similar way to task variance, except that it is computed for individual task epochs instead of tasks. There are clusters of units that are selective in specific epochs.

### 6.2.4 Distinct types of neural relationship between pairs of tasks

The map of normalized task variance in **Fig. 6.7c** allowed us to visualize the whole network across many tasks all at once. However, it is of limited use when we try to compare with experimental data or to read out the neural relationship between a given pair of tasks. To quantify how each unit is selective in one task in comparison to another task, we introduce a simple measure based on task variance: the Fractional Task Variance (FTV). For unit $i$, the fractional task variance with respect to task A and task B is defined as

$$\mathrm{FTV}_i(A,B) = \frac{\mathrm{TV}_i(A) - \mathrm{TV}_i(B)}{\mathrm{TV}_i(A) + \mathrm{TV}_i(B)}, \tag{6.1}$$

where $\mathrm{TV}_i(A)$ and $\mathrm{TV}_i(B)$ are the task variances for tasks A and B respectively. Fractional task variance ranges between $-1$ and $+1$. Having a $\mathrm{FTV}_i(A,B)$ close to +1 (or -1) means that unit $i$ is primarily selective in task A (or B).

For every pair of tasks, we can compute the fractional task variance for all units that are active in at least one of the two tasks. Each distribution of FTVs contains rich information about the single-unit level neural relationship between the pair of tasks. Having 20 tasks provides us with 190 distinct FTV distributions (**Fig. 6.11**), from the shape of which we summarized six typical neural relationships (**Fig. 6.12**).

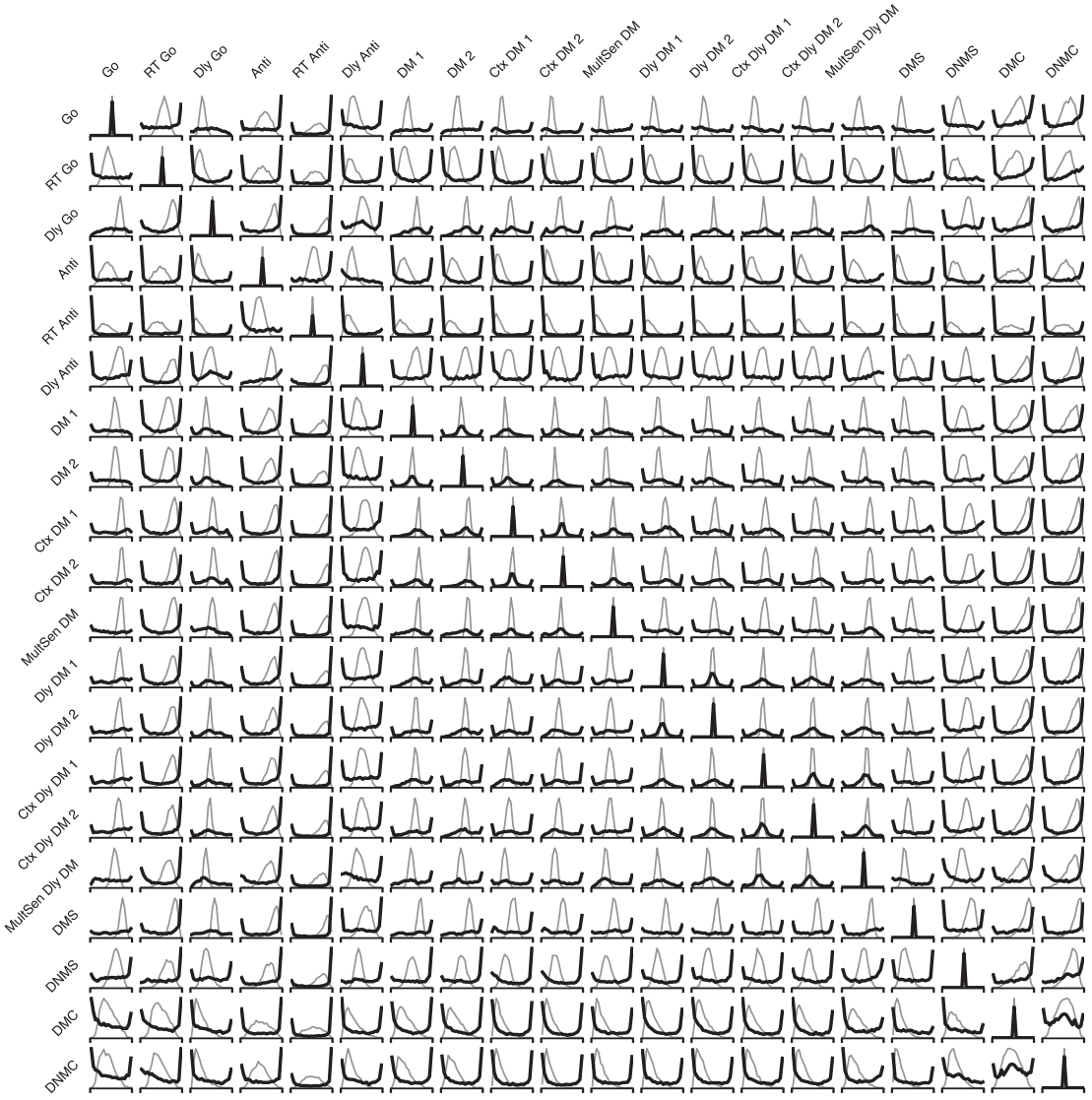


Figure 6.11: Fractional variance distributions for all pairs of tasks. There is a total of 190 unique pairs of tasks from all 20 tasks trained. Each fractional variance distribution (black) shown here is averaged across 20 networks. As a control, we also computed fractional variance distributions (gray) from activities of surrogate units that are generated by randomly mixing activities of the original network units (see Online Methods).

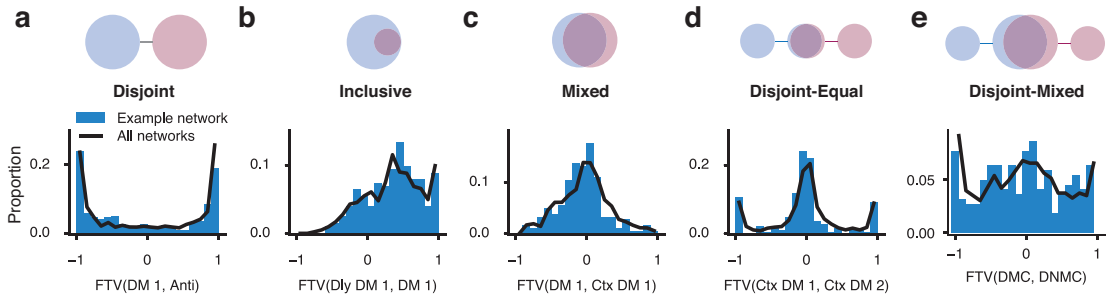


Figure 6.12: A diversity of neural relationships between pairs of tasks. For a pair of tasks, we characterize their neural relationship by the distribution of fractional task variances over all units. We observed five typical relationships: Disjoint (**a**), Inclusive (**b**), Mixed (**c**), Disjoint-Equal (**d**), and Disjoint-Mixed (**e**). Blue: distribution for one example network. Black: averaged distribution over 20 networks.

1. Equal. When the two tasks are the same or very similar, the FTV distribution has a trivial single peak around 0.

2. Disjoint (**Fig. 6.12a**). When two tasks have a disjoint relationship like the Anti task and the DM1 task, the FTV distribution is characterized by two peaks at the two ends and few units in between. There is little overlap between units selective in the two tasks. The shape of the FTV distribution is rather robust across independently trained networks.

3. Inclusive (**Fig. 6.12b**). This relationship is embodied by a strongly skewed FTV distribution, suggesting that one task is neurally a subset of another task. In this case, there are no units that are selective in the DM1 task yet not in the Dly DM 1 task.

4. Mixed (**Fig. 6.12c**). A mixed relationship is characterized by a broad uni-modal FTV distribution centered around 0 with no clear peak at the two ends. This distribution suggests that the two tasks share the same neural circuit, but may utilize them in different ways.

5. Disjoint-Equal (**Fig. 6.12d**). For Ctx DM 1 and 2, the FTV distribution is trimodal, with two peaks at the two ends and an additional peak around 0. This relationship can be considered as a combination of the Disjoint and the Equal relationships. In this sce-

nario, the two tasks each gets a private neural population, while they also share a third population.

6. Disjoint-Mixed (**Fig. 6.12e**). This relationship is a combination of the Disjoint and the Mixed relationships. Many units only participate in one of the two tasks, while the rest of the units are mixed in both tasks.

In summary, we introduced a simple yet informative measure to study the diverse neural relationships between pairs of tasks. We found that these relationships can be categorized into several canonical types. Our results on FTV distributions (**Fig. 6.11**) provide an array of straightforward predictions on pairwise neural relationships between cognitive tasks.

### 6.2.5 Compositional representations of tasks

A cognitive task can, in general, be expressed abstractly as a sequence of sensory, cognitive and motor processes, and cognitive processes may involve a combination of basic functions (such as working memory) required to perform the task. The compositionality of cognitive tasks is natural for human subjects because tasks are instructed with natural languages, which are compositional in nature (Cole, Laurent and Stocco, 2013). For example, the Go task can be instructed as "Saccade to the direction of the stimulus after the fixation cue goes off," while the Dly Go task can be instructed as "Remember the direction of the stimulus, then saccade to that direction after the fixation cue goes off." Therefore, the Dly Go task can be expressed as a composition of the Go task with a particular working memory process. Similarly, the Anti task can be combined with the same working memory process to form the Dly Anti task.

Here we test whether the network developed compositional representations for tasks,

even when it was never explicitly provided with the relationships between tasks. For the sake of simplicity, we studied the representation of each task as a single high-dimensional vector. To compute this "task vector", we averaged neural activities across all possible stimulus conditions within each task and focused on the steady-state response during the stimulus epoch (**Fig. 6.13a**).

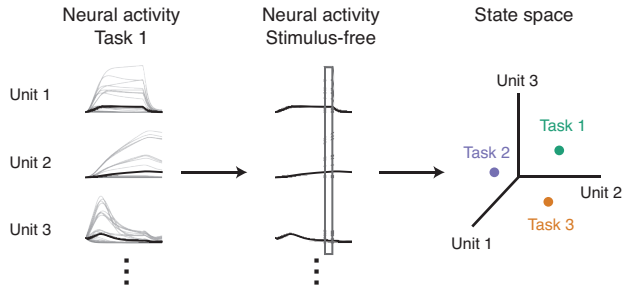


Figure 6.13: Representation of tasks in state space. The representation of each task is the population activity of the recurrent network at the end of the stimulus presentation, averaged across different stimulus conditions.

The neural population state near the end of stimulus presentation is representative of how the network processed the stimulus in a particular task to meet the computational need of subsequent behavioral epochs. Indeed, this idea is confirmed using principal component analysis, which revealed that task vectors in the state space spanned by the top two principal components are distinct for all twenty tasks (**Fig. 6.14**).

When plotting the task vectors representing the Go, Dly Go, Anti, and Dly Anti tasks, we found that the vector pointing from the Go vector towards the Dly Go vector is very similar to the vector pointing from the Anti vector to the Dly Anti vector (**Fig. 6.15a**). This finding is surprisingly robust and becomes even more apparent when we combined results from many networks (**Fig. 6.15b**). The Go-to-Dly Go vector and the Anti-to-Dly Anti vector presumably reflect the cognitive process of working memory. Similar findings are made with another set of tasks. The vector pointing from the Ctx DM 1 task to the Ctx DM 2 task is similar to the vector pointing from the Ctx Dly DM 1 task to the Ctx Dly DM 2 task

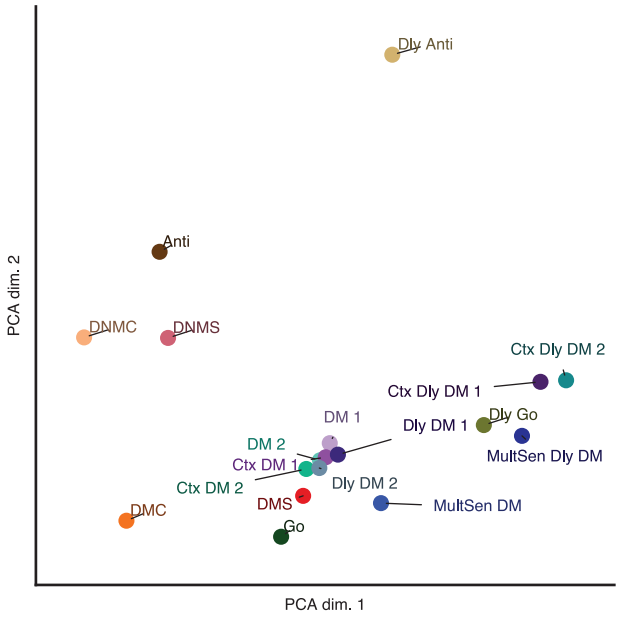


Figure 6.14: Representation of all tasks in state space. The representation of each task is computed the same way as in **Fig. 6.13**. Here showing the representation of all tasks in the top two principal components. RT Go and RT Anti tasks are not shown here because there is no well-defined stimulus epoch in these tasks.

(**Fig. 6.16**). The Ctx DM 1-to-Ctx DM 2 vector reflect the difference between the gating modality 1 and the gating modality 2 processes. These results suggest that sensory, cognitive, and motor processes can be represented as vectors in the task space. Therefore, the representation of a task can potentially be expressed as a linear summation of vectors representing the underlying sensory, cognitive, and motor processes. This finding is reminiscent of previous work showing that neural networks can represent words and phrases compositionally (Mikolov et al., 2013).

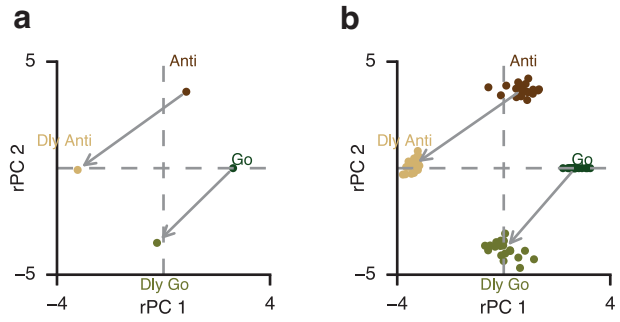


Figure 6.15: Compositional representation of tasks in state space. (**a**) Representations of the Go, Dly Go, Anti, Dly Anti tasks in the space spanned by the top two principal components (PCs) for a sample network. For better comparison across networks, the top two PCs are rotated and reflected (rPCs) to form the two axes (see Online Methods). (**b**) The same analysis as in (**a**) is performed for 20 networks, and the results are overlaid.

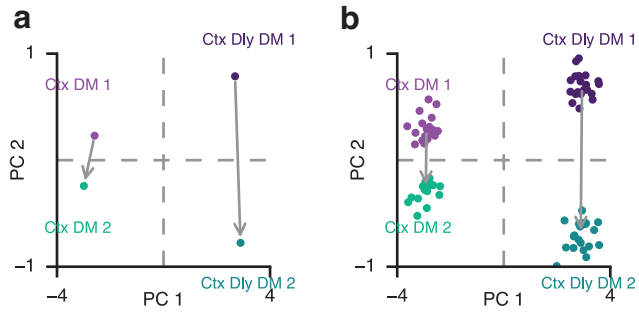


Figure 6.16: Compositional representation of tasks in state space. (**a**) Representations of the Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM 2 tasks in the top two PCs for a sample network. (**b**) The same analysis as in (**a**) for 20 networks.

### 6.2.6 Performing tasks with composition of rule inputs

We showed that the representation of tasks could be compositional in principle. However, it is unclear whether in our network this principle of compositionality can be extended from representing to performing tasks. The network is normally instructed which task to perform by activation of the corresponding rule input unit. What would the network do in response to a compositional rule signal as a combination of several activated and deactivated rule units? We tested whether the network can perform tasks by receiving composite rule inputs (**Fig. 6.17**).

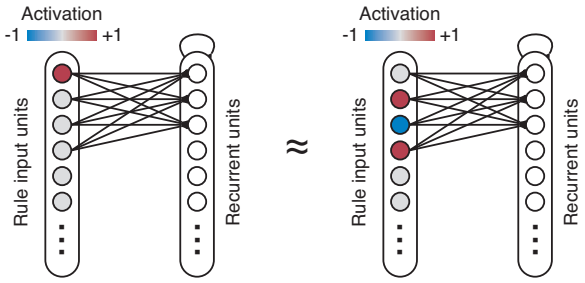


Figure 6.17: Performing tasks with algebraically composite rule inputs. During training, a task is always instructed by activation of the corresponding rule input unit (left). After training, the network can potentially perform a task by activation or deactivation of a set of rule input units meant for other tasks (right).

Consider the same two sets of tasks as in **Fig. 6.15** and **Fig. 6.16**. The network can perform the Dly Anti task almost perfectly when provided with the particular combination of rule inputs: Anti + (Dly Go - Go) (**Fig. 6.18a**). In contrast, the network fails to perform the Dly Anti task when provided with several other combinations of rule inputs (**Fig. 6.18a**). Similarly, the network can perform the Ctx Dly DM 1 task best when provided the composite rule inputs of Ctx Dly DM 2 + (Ctx DM 1 - Ctx DM 2) (**Fig. 6.18b**). In accordance with these results, we found that connection weights from individual rule input units to recurrent units also display a compositional structure (**Fig. 6.19**). Together, these results further confirmed that our network learned the implicit compositional re-

lationship between tasks. In such a network, learning a new task may not require any modification to the recurrent connections. Instead, it only requires learning the appropriate combination of rule inputs that control the information flow within the network (Miller and Cohen, 2001).

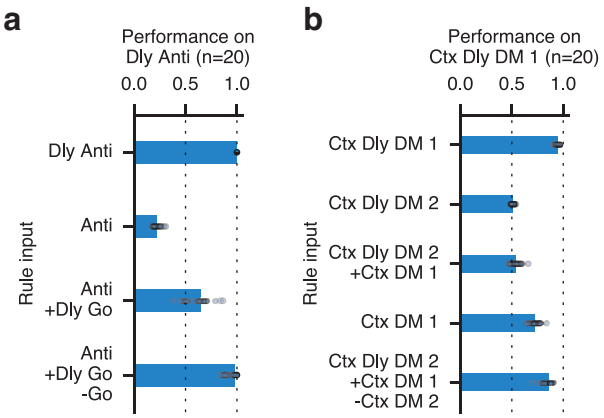


Figure 6.18: Performing tasks with algebraically composite rule inputs. (**a**) The network can perform the Dly Anti task well if given the Dly Anti rule input or the Anti + (Dly Go - Go) rule input. The network fails to perform the Dly Anti task when provided other combinations of rule inputs. (**b**) Similarly, the network can perform the Ctx Dly DM 1 task when provided with the Ctx Dly DM 2 + (Ctx DM 1 - Ctx DM 2) rule input. Circles represent results of individual networks, while bars represent median performances of 20 networks.

### 6.2.7 Continual training of many cognitive tasks

In humans and other animals, the performance of a well-trained task can be retained, even without re-training, for months or even years. However, when using traditional network training techniques, artificial neural networks rapidly forget previously learned tasks after being exposed to new tasks. This failure of retaining memories during sequential training of tasks, termed "catastrophic forgetting," is inevitable when using common network architectures and training methods (Kirkpatrick et al., 2017; Zenke, Poole and Ganguli, 2017). Network parameters (such as connection weights) optimal for a new task can be destructive for old tasks (**Fig. 6.20**). Recent work proposed several contin-

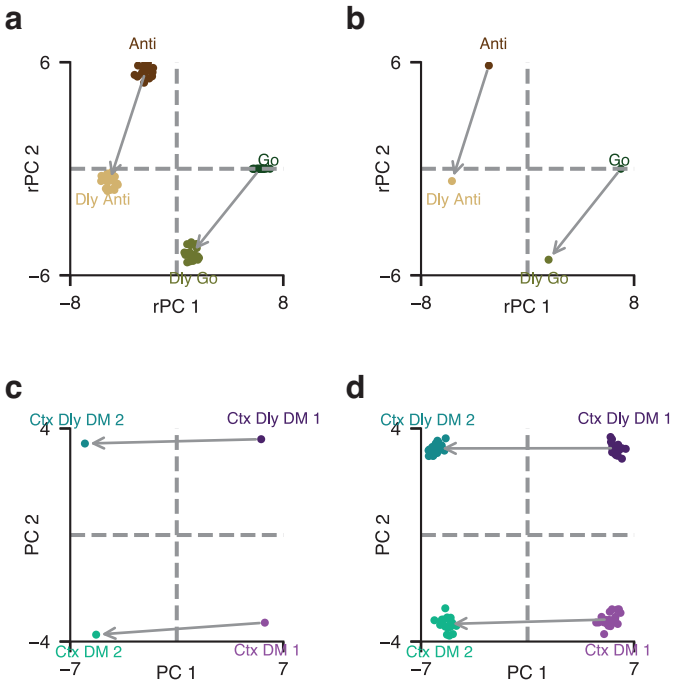


Figure 6.19: Connection weights of rule inputs in state space. (**a**) Connection weights from rule input units representing Go, Dly Go, Anti, Dly Anti tasks visualized in the space spanned by the top two principal components (PCs) for a sample network. Similar to 6.15, the top two PCs are rotated and reflected (rPCs) to form the two axes. (**b**) The same analysis as in (**a**) is performed for 20 networks, and the results are overlaid. (**c**) Connection weights from rule input units representing Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM 2 tasks visualized in the top two PCs for a sample network. (**d**) The same analysis as in (**c**) for 20 networks.

ual learning methods to battle catastrophic forgetting (Benna and Fusi, 2016; Kirkpatrick et al., 2017; Zenke, Poole and Ganguli, 2017). These methods typically involve selective protection of connection weights that are deemed important for previously learned tasks.

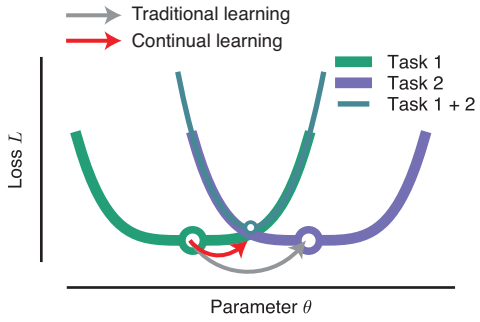


Figure 6.20: Schematics of continual learning. The network learns to perform a task by modifying parameters to minimize the loss function for this task. When a network is trained on two tasks sequentially with traditional learning techniques (gray arrow), training for the second task can easily result in the failure of performing the first task, because the minima (circle) of the loss functions of tasks 1 (green) and 2 (purple) are far apart. Continual learning techniques can protect previously-learned tasks by preventing large changes of important network parameters (red arrow). Arrows show changes of an example parameter $\theta$ when task 2 is trained after task 1 is already learned.

By employing one such techniques (Zenke, Poole and Ganguli, 2017), we were able to radically improve the performance of networks that are sequentially trained on a set of cognitive tasks (**Fig. 6.21**). For example, the network can retain high performance in a working memory task after successfully learning ten additional tasks (**Fig. 6.22**). Interestingly, even though asked to retain structures for old tasks, our networks trained with the continual learning technique are overall better at learning new tasks. In comparison to networks trained with traditional techniques, the continual learning network learned faster and better at all tasks from the Dly DM family. We speculate that the network trained with the traditional technique developed a structure overly-specialized for previously trained tasks; in contrast, the continual learning network maintains a structure flexible enough for learning new tasks.

Finally, we analyzed the FTV distributions for three example pairs of tasks in the

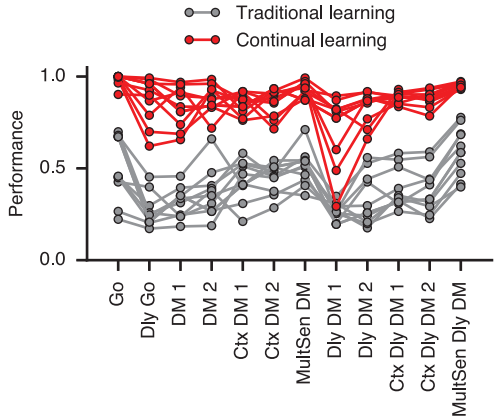


Figure 6.21: Final performance across all trained tasks. Final performance across all trained tasks with traditional (gray) or continual (red) learning techniques. Only 12 tasks are trained due to difficulty of learning more tasks even with continual learning techniques. Lines represent results of individual networks.

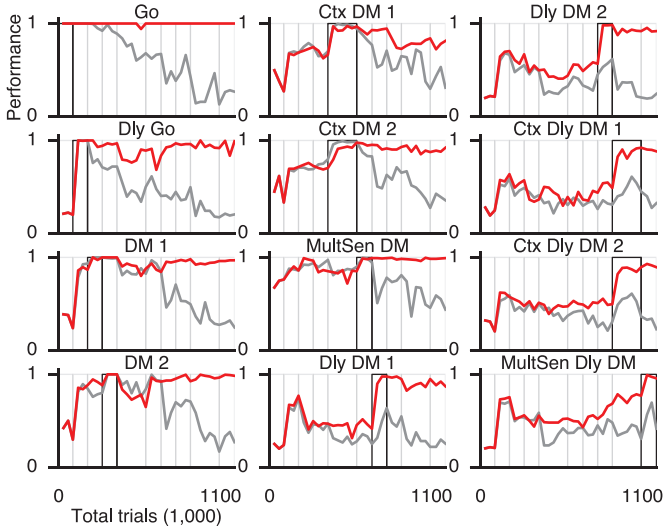


Figure 6.22: Sequential training of cognitive tasks. Performance of all tasks during sequential training of one network with traditional (gray) or continual (red) learning techniques. For each task, the black box indicates the period in which this task is trained.

continual learning networks (**Fig. 6.23**). The shapes of these FTV distributions can be markedly different from the corresponding ones of the interleaved-training networks (**Fig. 6.12**). It is possible that this result depends on factors in the continual learning, such as the order of individual tasks used during training, more careful comparisons are needed in future studies. Nevertheless, our findings suggest that sequential training of tasks could drastically shape neural network representations.

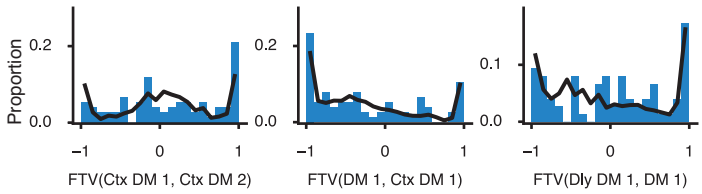


Figure 6.23: Fractional variance distributions for three pairs of tasks. Blue: distribution for one sample network. Black: averaged distribution over five networks.

## 6.3 Discussion

Higher cortical areas, especially the lateral prefrontal cortex, are remarkably versatile for their engagement in a wide gamut of cognitive functions. Here we investigated how multiple cognitive tasks are represented in a single recurrent neural network model. First, we demonstrated how the trained neural network could be dissected and understood for a family of tasks. Next, we identified clusters of units that are each specialized for a subset of tasks. Each cluster potentially represents a particular sequence of the sensori-motor events and a subset of cognitive (such as working memory, categorization, decision-making, inhibitory control) processes that are the building blocks for flexible behavior. We proposed a measure, the fractional task variance, that probes the neural relationship between a pair of tasks at the single-neuron level. This measure allowed us to summarize

six distinct and typical kinds of neural relationships in our network. This measure can be readily applied to firing activity of single units recorded from animals performing two or more tasks. Surprisingly, we found that the representation of tasks in our network is compositional, a critical feature for cognitive flexibility. By virtue of the compositionality, a task can be correctly instructed by composing instructions for other tasks. Finally, using a recently proposed continual learning technique, we can train the network to learn many tasks sequentially.

Monkeys, and in some cases rodents, can be trained to alternate between two tasks (Mante et al., 2013; Munoz and Everling, 2004; Siegel, Buschman and Miller, 2015; Mirabella et al., 2007; Wimmer et al., 2015). Single-unit recordings from these experiments can potentially be analyzed to compute the fraction task variance distributions. Theoretical studies argued that for maximum cognitive flexibility, prefrontal neurons should be selective to mixtures of multiple task variables (Rigotti et al., 2010). Mixed selectivity neurons are indeed ubiquitous within the prefrontal cortex (Rigotti et al., 2013). We showed that most units in our network are strongly selective to rules (**Fig. 6.7**). Meanwhile, these units are selective to other aspects of tasks (otherwise their task variances would be zero). For example, the Anti units (**Fig. 6.6**) are highly activated only during the Anti tasks and when the stimulus is in their preferred directions. Therefore, units in our network display strong nonlinear mixed selectivity, as found in neurons of the prefrontal cortex (Rigotti et al., 2013). Conceptually, this work extends the notion of mixed selectivity from within a single task to across multiple tasks.

Multiple cognitive tasks are more common in human imaging studies. In a series of experiments, Cole and colleagues trained humans to perform 64 cognitive tasks following compositional rule instructions (Cole et al., 2011 2013). They trained linear classifiers to

decode rules from prefrontal neural activity patterns. These classifiers can significantly generalize to novel tasks (Cole et al., 2011), consistent with a compositional neural representation of rules. Although trained with discrete rule instructions, our network shows a clear compositional structure in its representations, using the population activity at a single time point (near the end of stimulus presentation). Temporal dynamics in neural circuits are ubiquitous during cognitive tasks (Machens, Romo and Brody, 2010) and are potentially critical for cognitive computations (Chaisangmongkon et al., 2017), so the study of steady-state responses here is merely a first step towards understanding the dynamical representation of tasks. Cole et al. found that humans can rapidly adapt to new tasks by adjusting the functional connectivity patterns of parietal-frontal flexible hubs (Cole et al., 2013). In the future, graph-theoretic analysis can be used to test whether our trained network developed flexible hubs that coordinate information flow across the network. There exists a structural hierarchy within the human prefrontal cortex, with more abstract cognitive processes being represented in the more anterior areas (Koechlin, Ody and Kouneiher, 2003; Badre, 2008). It is unclear if our trained network developed hierarchical representations of cognitive processes or tasks. If it did, a subset of units should represent more abstract aspects of the tasks, while other units represent the concrete, sensorimotor aspects. This question is hard to address for now because the 20 tasks we chose are not organized in a clearly hierarchical way (Koechlin, Ody and Kouneiher, 2003).

Training artificial neural networks for multiple tasks has a long history in the field of machine learning (Caruana, 1997). However, it has mainly been used as a method to improve training and generalization. There were few studies on the representation of task structure or task set in trained networks. Modern artificial neural networks are capable of

highly complex tasks, such as playing Atari games (Mnih et al., 2015), which likely involve a range of cognitive skills. However, in contrast to cognitive tasks that are specifically designed to shed light on neural mechanisms of cognition, complex real-life tasks remain challenging to analyze. In principle, we can strike a balance between the two approaches by designing a set of tasks that are complex enough, yet still amenable to analysis. The ability to "open the box" and elucidate the inner working of the network after training is crucial for understanding neural mechanisms of cognition in Neuroscience.

Like other works on trained neural networks (Mante et al., 2013; Zipser and Andersen, 1988; Song, Yang and Wang, 2016; Carnevale et al., 2015; Rajan, Harvey and Tank, 2016; Chaisangmongkon et al., 2017; Yamins et al., 2014; Eliasmith et al., 2012), the machine learning protocol we used is not validated biologically. Besides, our RNN consists of a single neural population, in contrast to the brain system where a number of interacting brain regions are engaged in a cognitive task (Siegel, Buschman and Miller, 2015; Christophel et al., 2017). Although our neural network model developed functionally specialized clusters of units through training, it is unclear how to map them onto different brain areas. Furthermore, in our network, a rule input is explicitly provided throughout the trial, therefore there is no need for the network to hold the "task set" internally by virtue of persistent activity (Wallis, Anderson and Miller, 2001; Sakai, 2008). This, however, can be remedied by providing the rule cue only at the beginning of each trial which would encourage the network to internally sustain the task set. We can even ask the network to figure out a task rule by trial-and-error (Song, Yang and Wang, 2017). In spite of potential caveats, training these recurrent networks takes only hours, offering an efficient computational platform to test hypotheses about neural representations and mechanisms that could guide experiments and data analysis. Furthermore, this approach can

yield new conceptual insights, as shown here by the finding of compositional task representation. Future progress in this direction, at the interface between neuroscience and artificial intelligence, will advance our understanding of flexible behavior in many cognitive tasks.

## 6.4 Methods

### 6.4.1 Network structure

The recurrent neural networks shown in the main text all contain $N_{\text{rec}} = 256$ units. The results are largely insensitive to the network size. Similar results were obtained in networks of sizes between 128 and 512 units (the range we tested). The network is a time-discretized recurrent neural network with positive activity (Song, Yang and Wang, 2016). Before time-discretization, the network activity $\mathbf{r}$ follows a continuous dynamical equation

$$\tau \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t} = -\mathbf{r} + f(W^{\text{rec}}\mathbf{r} + W^{\text{in}}\mathbf{u} + \mathbf{b}) + \sqrt{2\tau\sigma_{\text{rec}}^2}\xi. \tag{6.2}$$

In this equation, $\mathbf{u}$ is the input to the network, $\mathbf{b}$ is the bias or background input, $\tau = 100$ms is the neuronal time constant, $f(\cdot)$ is the neuronal nonlinearity that keeps the unit activity non-negative, $\xi$ are $N_{\text{rec}}$ independent Gaussian white noise processes with zero mean and unit variance, and $\sigma_{\text{rec}} = 0.05$ is the strength of the noise. In particular, we use a standard softplus function

$$f(x) = \log(1 + \exp(x)), \tag{6.3}$$

which after re-parameterization is very similar to a neuronal nonlinearity, i.e., f-I curve, commonly used in previous neural circuit modelings (Abbott and Chance, 2005). A set of output units $\mathbf{z}$ read out nonlinearly from the network,

$$\mathbf{z} = g(W^{\text{out}}\mathbf{r}), \tag{6.4}$$

where $g(x) = 1/(1+\exp(-x))$ is the logistic function, bounding output activities between 0 and 1. $W^{\text{in}}, W^{\text{rec}}, W^{\text{out}}$ are the input, recurrent, and output connection matrices respectively.

After using the first-order Euler approximation with a time discretization step $\Delta t$, we have

$$\mathbf{r}_t = (1-\alpha)\mathbf{r}_{t-1} + \alpha \cdot f(W^{\text{rec}}\mathbf{r}_{t-1} + W^{\text{in}}\mathbf{u}_t + \mathbf{b}) + \sqrt{2\alpha\sigma^2_{\text{rec}}}\mathbf{N}(0,1). \tag{6.5}$$

Here $\alpha \equiv \Delta t/\tau$, and $\mathbf{N}(0,1)$ stands for the standard normal distribution. We imposed no constraint on the sign or the structure of the weight matrices $W^{\text{in}}, W^{\text{rec}}, W^{\text{out}}$. The network and the training are implemented in TensorFlow (Abadi et al., 2015).

The network receives four types of noisy inputs,

$$\mathbf{u} = (u_{\text{fix}}, \mathbf{u}_{\text{mod1}}, \mathbf{u}_{\text{mod2}}, \mathbf{u}_{\text{rule}}) + \mathbf{u}_{\text{noise}}. \tag{6.6}$$

$$\mathbf{u}_{\text{noise}} = \sqrt{\frac{2}{\alpha}}\sigma_{\text{in}}\mathbf{N}(0,1). \tag{6.7}$$

Here the input noise strength $\mathbf{N}(0,1) = 0.01$. The fixation input $u_{\text{fix}}$ is typically at the high value of 1 when the network should fixate. The fixation input goes to zero when the network is required to respond. The stimulus inputs $\mathbf{u}_{\text{mod1}}$ and $\mathbf{u}_{\text{mod2}}$ comprise of two "rings" of units, each representing a one-dimensional circular variable described by the degree around a circle. Each ring contains 32 units, whose preferred directions are uniformly spaced from 0 to $2\pi$. For unit $i$ with a preferred direction $\theta_i$, its activity for a stimulus presented at direction $\psi$ is

$$u_i = \gamma \cdot 0.8 \exp\left[-\frac{1}{2}(\frac{8|\psi - \psi_i|}{\pi})^2\right], \tag{6.8}$$

where $\gamma$ is the strength of the stimulus. For multiple stimuli, input activities are added together. The network also receives a set of rule inputs $\mathbf{u}_{\text{rule}}$ that encode which task the network is supposed to perform on each trial. Normally, $\mathbf{u}_{\text{rule}}$ is a one-hot vector. That means the rule input unit corresponding to the current task is activated at 1, while other rule input units remain at 0. Therefore the number of rule input units equals to the number of tasks trained. For compositional rule inputs (**Fig. 6.17**), the activation of rule input units can be an arbitrary pattern. For example, for the combined rule input Anti + (Dly Go - Go), the activities of the rule input units corresponding to the Go, Dly Go, and Anti tasks are -1, +1, and +1 respectively. In total there are $N_{\text{in}} = 1+32\times 2+20 = 85$ input units.

The network projects to an output ring $\mathbf{z}_{\text{out}}$, which also contains 32 units. The output ring units encode the response directions using similar tuning curves to the ones used for the input rings. In addition, the network projects to a fixation output unit $\mathbf{z}_{\text{fix}}$, which should be at the high activity value of 1 before the response and at 0 once a response is generated. In total there are $N_{\text{out}} = 1+32 = 33$ output units.

We lesion a network unit by setting to zero its projection weights to all recurrent and output units.

### 6.4.2 Tasks and performances

Here we first describe the common setup for the 20 tasks trained. Deviations from the common setup will be described below individually. The rule input unit corresponding to the current task will be activated throughout the whole trial. The network receives a fixation input, which is activated from the beginning of the trial. When the fixation input is on, the network should fixate by having the fixation output unit at a high activity $\hat{\mathbf{z}}_{\text{fix}} = 0.85$. The offset of the fixation input usually indicates the onset of the response or

go epoch, when the network needs to report the response direction through activities of the output ring. During the response epoch, the fixation output unit has a target output of $\hat{\mathbf{z}}_{\text{fix}} = 0.05$. For a target response direction $\psi$, the target output activity of an output unit $i$ is

$$\hat{\mathbf{z}}_i = 0.8 \exp\left[-\frac{1}{2}\left(\frac{8|\psi - \psi_i|}{\pi}\right)^2\right] + 0.05, \tag{6.9}$$

where $\psi_i$ is the preferred response direction of unit $i$. When no response is required, the target output activity is fixed at $\hat{\mathbf{z}}_i = 0.05$. The network also receives one or two stimuli. Each stimulus contains information from modality 1, 2, or both.

A trial is considered correct only if the network correctly maintained fixation and responded to the correct direction. The response direction of the network is read out using a population vector method. The decoded response direction is considered correct if it is within 36 degree of the target direction. If the activity of the fixation output falls below 0.5, the network is considered to have broken fixation.

The discrimination thresholds $a$ in **Fig. 6.5** are obtained by fitting Weibull functions to performances $p$ as a function of coherences $c$ at a fixed stimulus duration,

$$p = 1 - 0.5 \exp\left(-(c/a)^b\right). \tag{6.10}$$

Each task can be separated into distinct epochs. Fixation (fix) epoch is the period before any stimulus is shown. It is followed by the stimulus epoch 1 (stim1). If there are two stimuli separated in time, then the period between the two stimuli is the delay epoch, and the second stimulus is shown in the stimulus epoch 2 (stim2). The period when the network should respond is the go epoch. The duration of the fixation, stim1, delay1,

stim2, and go epochs are $T_{\text{fix}}, T_{\text{stim1}}, T_{\text{delay1}}, T_{\text{stim2}}, T_{\text{go}}$ respectively. For convenience, we grouped the 20 tasks into five task families: the Go, Anti, Decision-Making (DM), Delayed Decision-Making (Dly DM), and Matching families.

**Go task family.** This family of tasks includes the Go, RT Go, and Dly Go tasks. In all three tasks, a single stimulus is randomly shown in either modality 1 or 2, and the response should be made in the direction of the stimulus. These three tasks differ in their stimulus onset and offset times. In the Go task, the stimulus appears before the fixation cue goes off. In the RT Go task, the fixation input never goes off, and the network should respond as soon as the stimulus appears. In the Dly Go task, a stimulus appears briefly and is followed by a delay period until the fixation cue goes off. The Dly Go task is similar to the memory-guided saccade task (Funahashi, Bruce and Goldman-Rakic, 1989).

For the Go task,

$$T_{\text{stim1}} \sim \mathbf{U}(500, 1500). \tag{6.11}$$

$\mathbf{U}(t_1, t_2)$ is a uniform distribution between $t_1$ and $t_2$. The unit for time is ms and is omitted for brevity. For the RT Go task,

$$T_{\text{stim1}} \sim \mathbf{U}(500, 2500). \tag{6.12}$$

For the Dly Go tasks,

$$T_{\text{delay1}} \sim \mathbf{U}(\{200, 400, 800, 1600\}). \tag{6.13}$$

Here $\mathbf{U}(\{a_1, \cdots, a_n\})$ denotes a discrete uniform distribution over the set $\{a_1, \cdots, a_n\}$.

**Anti task family.** This family includes the Anti, RT Anti, and Dly Anti tasks. These three tasks are the same as their counterpart Go-family tasks, except that the response should be made to the opposite direction of the stimulus.

**DM family.** This family includes five perceptual decision making tasks: the DM 1, DM 2, Ctx DM 1, Ctx DM 2, and MultSen DM tasks. In each trial, two stimuli are shown simultaneously and are presented till the end of the trial. In DM 1, the two stimuli only appear in modality 1, while in DM 2, the two stimuli only appear in modality 2. In DM 1 and DM 2, the correct response should be made to the direction of the stronger stimulus (the stimulus with higher $\gamma$). In Ctx DM 1, Ctx DM 2, and MultSen DM tasks, each stimulus appears in both modality 1 and 2. In the Ctx DM 1 task, information from modality 2 should be ignored, and the correct response should be made to the stronger stimulus in modality 1. In the Ctx DM 2 task, information from modality 1 should be ignored. In the MultSen DM task, the correct response should be made to the stimulus that has a stronger combined strength in modalities 1 and 2.

The DM 1 and DM 2 tasks are inspired from classical perceptual decision making tasks based on random-dot motion stimuli (Gold and Shadlen, 2007). The two stimuli shown represent momentary motion evidence towards the two target directions. When the two stimuli have the same strengths ($\gamma_1 = \gamma_2$), there is no net evidence towards any target direction, mimicking the condition of 0 motion coherence in the random-dot motion task. A stronger difference in the stimulus strengths emulates a stronger motion coherence. For a coherence $c$ representing net evidence for the direction of stimulus 1, the strengths of stimulus 1 and 2 $(\gamma_1, \gamma_2)$ are set as

$$\gamma_{1,\mathrm{mod}i} = \bar{\gamma} + c,\ \gamma_{2,\mathrm{mod}i} = \bar{\gamma} - c, \tag{6.14}$$

respectively, where $i \in 1,2$ is the modality. Here $\bar{\gamma}$ is the average strength of the two stimuli. For each trial, we draw $\bar{\gamma}$ from a uniform distribution around 1, $\bar{\gamma} \sim \mathbf{U}(0.8, 1.2)$. Indeed, in all DM-family tasks and Dly DM-family tasks, there is a single coherence $c$ in each trial that determines the overall strength of net evidence towards the direction represented by stimulus 1. For all DM family tasks,

$$c \sim \mathbf{U}(\{-0.08, -0.04, -0.02, -0.01, 0.01, 0.02, 0.04, 0.08\}). \tag{6.15}$$

The duration of stimulus 1, which is fixed in each trial, is drawn from the following distribution,

$$T_{\text{stim1}} \sim \mathbf{U}(\{400, 800, 1600\}). \tag{6.16}$$

Indeed, all tasks from the DM family use the same distribution for $T_{\text{stim1}}$. And since the two stimuli are shown simultaneously, $T_{\text{stim1}} = T_{\text{stim2}}$.

The Ctx DM 1 and Ctx DM 2 tasks are inspired from context-dependent decision-making tasks performed by macaque monkeys (Mante et al., 2013). Now each stimulus is presented in both modalities at the same direction, with strengths $\gamma_{1,\text{mod1}}, \gamma_{1,\text{mod2}}$ for stimulus 1, and $\gamma_{2,\text{mod1}}, \gamma_{2,\text{mod2}}$ for stimulus 2. The stimulus strengths are determined by the coherence for modality 1 and 2 ($c_{\text{mod1}}, c_{\text{mod2}}$), so we have

$$\gamma_{1,\text{mod1}} = \bar{\gamma}_{\text{mod1}} + c_{\text{mod1}},\ \gamma_{2,\text{mod1}} = \bar{\gamma}_{\text{mod1}} - c_{\text{mod1}}. \tag{6.17}$$

Similar equation holds for modality 2 as well. $c_{\text{mod1}}$ and $c_{\text{mod2}}$ are drawn independently from the same distribution. In Ctx DM 1, $c = c_{\text{mod1}}$, while in Ctx DM 2, $c = c_{\text{mod2}}$. $\bar{\gamma}_{\text{mod1}}$

and $\bar{\gamma}_{\text{mod2}}$ are also drawn from $\mathbf{U}(0.8, 1.2)$. In the original Mante task (Mante et al., 2013), there is an additional delay period between the stimuli and the response period, which is not included here.

The MultSen DM task mimics a multi-sensory integration task (Raposo, Kaufman and Churchland, 2014). The setup of stimulus is similar to those in the Ctx DM 1 and Ctx DM 2 tasks, except that the network should integrate information from both modalities and the stronger stimulus is the one with higher averaged strength from modality 1 and 2. The overall coherence $c = (c_{\text{mod1}} + c_{\text{mod2}})/2$. We determine all four strengths with the following procedure. First we determine the average strength of stimulus 1 across both modalities, $\gamma_1$, and the average strength of stimulus 2, $\gamma_2$.

$$\gamma_1 = \bar{\gamma} + c,\ \gamma_2 = \bar{\gamma} - c. \tag{6.18}$$

Here $\bar{\gamma}$ and $c$ both follow the same distributions as other DM-family tasks. Then we set

$$\gamma_{1,\text{mod1}} = \gamma_1(1 + \Delta_1),\ \gamma_{1,\text{mod2}} = \gamma_1(1 - \Delta_1), \tag{6.19}$$

where $\Delta_1 \sim \mathbf{U}(0.1, 0.4) \cup \mathbf{U}(-0.4, -0.1)$. Similarly for stimulus 2.

**Dly DM family.** This family includes Dly DM 1, Dly DM 2, Ctx Dly DM 1, Ctx Dly DM 2. These tasks are similar to the corresponding tasks in the DM family, except that in the Dly DM family tasks, the two stimuli are separated in time. The Dly DM 1 and Dly DM 2 tasks are inspired by the classical parametric working memory task developed by Romo and colleagues (Romo et al., 1999). The two stimuli are both shown briefly and are separated by a delay period. Another short delay period follows the offset of the second stimulus.

For all Dly DM family tasks,

$$T_{\text{delay1}} \sim \mathbf{U}(\{200, 400, 800, 1600\}),\ c \sim \mathbf{U}(\{-0.32, -0.16, -0.08, 0.08, 0.16, 0.32\}). \quad (6.20)$$

And, $T_{\text{stim1}} = T_{\text{stim2}} = 300$.

**Matching family.** This family of tasks includes the DMS, DNMS, DMC, DNMC tasks. In these tasks, two stimuli are presented consecutively and separated by a delay period. Each stimulus can appear in either modality 1 or 2. The network response depends on whether or not the two stimuli are "matched." In the DMS and DNMS tasks, two stimuli are matched if they point towards the same direction, regardless of their modalities. In DMC and DNMC tasks, two stimuli are matched if their directions belong to the same category. The first category ranges from 0 to 180 degrees, while the rest from 180 to 360 degrees belongs to the second category. In the DMS and DMC tasks, the network should respond towards the direction of the second stimulus if the two stimuli are matched and maintain fixation otherwise. In the DNMS and DNMC tasks, the network should respond only if the two stimuli are not matched, i.e., a non-match, and fixate when it is a match.

In all Matching family tasks,

$$T_{\text{delay1}} \sim \mathbf{U}(\{200, 400, 800, 1600\}). \quad (6.21)$$

Also, match trials and non-match trials always appear with equal probability.

### 6.4.3 Training procedure

The loss $\mathscr{L}$ to be minimized is computed by time-averaging the squared errors between the network output $\mathbf{z}(t)$ and the target output $\hat{\mathbf{z}}(t)$.

$$\mathscr{L} = \mathscr{L}_{\text{sqe}} \equiv \langle \mathbf{m}_{i,t}(\mathbf{z}_{i,t} - \hat{\mathbf{z}}_{i,t})^2 \rangle_{i,t}. \tag{6.22}$$

Here $i$ is the index of the output units. The squared errors at different time points and of different output units are potentially weighted differently according to the non-negative mask matrix $\mathbf{m}_{i,t}$. For the output ring units, before the response epoch, we have $\mathbf{m}_{i,t} = 1$. The first 100ms of the response epoch is a grace period with $\mathbf{m}_{i,t} = 0$, while for the rest of the response epoch, $\mathbf{m}_{i,t} = 5$. For the fixation output unit, $\mathbf{m}_{i,t}$ is two times stronger than the mask for the output ring units.

The training is performed with Adam, a powerful variant of stochastic gradient descent (Kingma and Ba, 2014). We used the default set of parameters. The learning rate is 0.001, the decay rate for the 1st and 2nd moment estimates are 0.9 and 0.999 respectively.

The recurrent connection matrix is initialized with a scaled identity matrix $q \cdot \mathbf{1}$ (Le, Jaitly and Hinton, 2015), where $\mathbf{1}$ is the identity matrix. We chose $q = 0.54$ such that the gradient is roughly preserved during backpropagation when the network is initialized. The input and output connection weights are initialized as independent Gaussian random variables with mean 0, and standard deviations $1/\sqrt{N_{\text{in}}}$ and $0.4/\sqrt{N_{\text{rec}}}$ respectively. The standard deviation value for the output weights is chosen to prevent saturation of output units after initialization.

During training, we randomly interleaved all the tasks with equal probabilities, except for the Ctx DM 1 and Ctx DM 2 tasks that appear five times more frequently, because

without sufficient training, the network gets stuck at an alternative strategy. Instead of correctly ignoring modality 1 or 2, the network can choose to ignore the context and integrate information from both modalities equally. This strategy gives the network an accuracy close to 75%. During training, we used mini-batches of 64 trials, in which all trials are generated from the same task for computational efficiency.

For continual learning in **Fig. 6.21**, tasks appear sequentially. Each task is trained for 150,000 trials. Ctx DM 1 and Ctx DM 2 are still trained together and interleaved, and so are Ctx Dly DM 1 and Ctx Dly DM 2. We added a regularizer that protects old tasks by setting an additional penalty for deviations of important synaptic weights (or other parameters) (Zenke, Poole and Ganguli, 2017). When training the $\mu$-th task, the regularizer is

$$\mathscr{L}_{\text{cont}} = c_{\text{cont}} \sum_k \Omega_k^{\mu} \left(\theta_k - \tilde{\theta}_k\right)^2 . \tag{6.23}$$

Here $c_{\text{cont}}$ is the overall strength of the regularizer, $\theta_k$ denotes the $k-$th parameter of the network. The value of the anchor parameter $\tilde{\theta}_k$ is the value of $\theta_k$ at the end of the last task (the $(\mu-1)$-th task). No regularizer is used when training the first task. And $\Omega_k^{\mu}$ measures how important the parameter is. Notice that two recent proposals (Zenke, Poole and Ganguli, 2017; Kirkpatrick et al., 2017) for continual learning both use regularizers of this form. The two proposals differ only in how the synaptic importances are computed. We chose the method of Zenke et al. 2017, who argued that the method of Kirkpatrick et al. 2017 measures the synaptic importance locally in the parameter space, resulting in underestimated and inaccurate synaptic importance values for our settings. In Zenke et al. 2017, the importance of one parameter is determined using this parameter's historic contribution to the change in the loss function. For the $k$-th parameter, the contribution

to the change in loss during task $\mu$ is

$$\omega_k^\mu = \sum_{t=t^{\mu-1}}^{t^\mu} g_k(\theta(t))\Delta\theta_k(t), \tag{6.24}$$

where $g_k(\theta(t))$ is the gradient of loss with respect to $\theta_k$ evaluated at $\theta_k(t)$, i.e., $\frac{\partial \mathscr{L}}{\partial \theta_k}|_{\theta_k(t)}$, and $\Delta\theta_k(t)$ is the parameter change taken at step $t$. Therefore $\omega_k^\mu$ tracks how parameter $\theta_k$ contributes to changes in the loss during the $\mu$-th task (from $t^{\mu-1}$ to $t^\mu$). The final synaptic importance is computed by first normalizing $\omega_k^\mu$ with the total change in the synaptic weight $\Delta_k^\mu = \theta_k(t^\mu) - \theta_k(t^{\mu-1})$, and summing $\omega_k^\nu$ for all tasks $\nu < \mu$.

$$\Omega_k^\mu = \sum_{\nu<\mu} \frac{\omega_k^\nu}{(\Delta_k^\nu)^2 + \xi}. \tag{6.25}$$

The additional hyperparameter $\xi$ prevents $\Omega_k^\mu$ from becoming too large. The hyperparameters $c = 0.1$ and $\xi = 0.01$ are determined by a coarse grid search. The final loss is the sum of the squared-error loss and the continual learning regularizer.

$$\mathscr{L} = \mathscr{L}_{\text{sqe}} + \mathscr{L}_{\text{cont}}. \tag{6.26}$$

### 6.4.4 Task variance analysis

A central goal of our analysis was to determine whether individual units within the network are selective to different tasks, or whether units tended to be similarly selective to all tasks. To quantify how selective a unit is in one task, we defined a task variance metric. To compute the task variance $\text{TV}_i(A)$ for task A and unit $i$, we ran the network for many stimulus conditions that span the space of possible stimuli. For example, in the DM family tasks, we ran the network for stimuli with directions ranging from 0 to 360 degrees and

with coherences ranging from almost 0 to 0.2. After running the network for many stimulus conditions, we computed the variance across stimulus conditions (trials) at each time point for a specific unit then averaged the variance across all time points to get the final task variance for this unit. The fixation epoch is excluded from this analysis. This process was repeated for each unit in the network. Therefore

$$\mathrm{TV}_i(A) = \langle \left[ r_i(j,t) - \langle r_i(j,t) \rangle_j \right]^2 \rangle_{j,t}, \tag{6.27}$$

where $r_i(j,t)$ is the activity of unit $i$ on time $t$ of trial $j$. In **Fig. 6.6, 6.7, 6.12**, we only analyzed active units, defined as those that have summed task variance across tasks higher than a threshold, $10^{-3}$. The results do not depend strongly on the choice of the threshold. This procedure prevents units with extremely low task variance from being included in the analysis.

Anti units in **Fig. 6.6** are defined as those units that have higher summed task variance for the Anti family of tasks ($S_{\mathrm{Anti}}$ =Anti, RT Anti, Dly Anti) than for all other tasks. So a unit $i$ is an Anti unit if

$$\sum_{A \in S_{\mathrm{Anti}}} \mathrm{TV}_i(A) > \sum_{A \notin S_{\mathrm{Anti}}} \mathrm{TV}_i(A). \tag{6.28}$$

The clustering of units based on their task variance patterns in **Fig. 6.8** uses K-means clustering from the Python package scikit-learn. To assess how well a clustering configuration is, we computed its silhouette coefficient based on intra-cluster and inter-cluster distances. A higher silhouette coefficient means a better clustering. The optimal number of clusters $\tilde{k}$ is determined by choosing the first $k$ such that the silhouette coefficient for $k+1$ clusters is worse than $k$ clusters.

In **Fig. 6.8**, we visualize the clustering using t-distributed Stochastic Neighbor Embedding (tSNE). For each unit, the normalized task variances across all tasks form a 20 dimensional vector that is then embedded in a 2-dimensional space. For the tSNE method, we used the exact method for gradient calculation, a learning rate of 100, and a perplexity of 30.

The fractional task variance with respect to tasks A and B is

$$\mathrm{FTV}_i(A,B) = \frac{\mathrm{TV}_i(A) - \mathrm{TV}_i(B)}{\mathrm{TV}_i(A) + \mathrm{TV}_i(B)}. \tag{6.29}$$

To obtain a statistical baseline for the FTV distributions as in **Fig. 6.11**, we transform the neural activities of the network with a random orthogonal matrix before computing the task variance. For each network, we generate a random orthogonal matrix $M$ using the Python package Scipy. All network activities are multiplied by this matrix $M$ to obtain a rotated version of the original neural representation.

$$\mathbf{r}_t^{\mathrm{rot}} = M\mathbf{r}_t. \tag{6.30}$$

Since multiplying neural activities by an orthogonal matrix is equivalent to rotating the neural representation in state space, this procedure will preserve results from state space analysis. We then compute task variances and fractional task variances using the rotated neural activities. The FTV distributions using the rotated activities are clearly different from the original FTV distributions.

### 6.4.5 State-space analysis

To compute the representation of a task in the state space, we first computed the neural activities across all possible stimulus conditions, then we averaged across all these conditions. For simplicity of the analysis, we chose to analyze only the steady state responses during the stimulus epoch. We do so by focusing on the last time point of the stimulus epoch, $t_{\text{stim1,end}}$. So the representation of task A is

$$\tilde{\mathbf{r}} = \langle \mathbf{r}(j, t_{\text{stim1,end}}) \rangle_j, \tag{6.31}$$

where $\mathbf{r}(j,t)$ is the vector of network activities at trial $j$ and time $t$ during task A.

For each set of tasks, we performed principal component analysis to get the lower dimensional representation. We repeated this process for different networks. The representations of Go, Anti, Dly Go, and Dly Anti tasks are close to four vertices of a square. As a result, the top two principal components have similar eigenvalues and are therefore interchangeable. To better compare across networks in **Fig. 6.15**, we allowed a rotation and a reflection within the space spanned by the top two PCs. For each network, the rotated and reflected PCs (rPCs) are chosen such that the Go task representation lies on the positive part of the x-axis, and the Dly Go task lies below the x-axis. The representation of Ctx DM 1, Ctx DM 2, Ctx Dly DM 1, and Ctx Dly DM 2 tasks do not form a square, so we only allowed reflections such that Ctx Dly DM 1 is in the first quadrant. The reflected PCs are still PCs.

# Chapter 7

# Training multi-type E-I circuits

## 7.1 Introduction

Training recurrent neural networks is an efficient way to study neural circuit mechanisms of cognition with minimal assumptions. Most models consist of identical neurons that initially are randomly connected. Through training, certain structures will emerge out of the functional need to perform the task at hand. However, a great deal is known about the wiring of biological circuits. Including these knowledge into recurrent network models will make comparison between models and data much easier.

The brain consists of many different types of interneurons. In particular, parvalbumin (PV)-expressing interneurons specialize at targeting somatic areas of pyramidal neurons while somatostatin (SST)-expressing interneurons specialize at targeting dendritic areas of pyramidal neurons (Markram et al., 2004). Besides their specialized connectivity with pyramidal neurons, these neurons also make cell-type specific connectivity within themselves. Vasoactive intestinal peptide-expressing (VIP) neurons primarily target SST neurons, although a weaker projection to PV neurons is also reported (Pi et al., 2013). SST

neurons project to every other type of interneurons while avoiding themselves. PV neurons inhibit themselves. The role of different types of interneurons have been studied experimentally (Kepecs and Fishell, 2014) and theoretically (Yang, Murray and Wang, 2016). The genetic tools to study functions of different types of interneurons only exist in mice and lower animals at this point. Therefore, experimental studies of the functional roles of different types of interneurons have so far been restricted to a limited set of behavior that are available in mice.

We have proposed that dendrite-targeting interneurons can be engaged in cognitive flexibility by supporting pathway-specific gating (Yang, Murray and Wang, 2016). Here we ask whether the network can develop the solution of utilizing dendrite-targeting interneurons in a task that requires pathway-specific gating. We trained a recurrent neural network equipped with multiple types of neurons, including multi-compartmental pyramidal neurons and three types of interneurons modeling PV, SST, and VIP neurons. The network is trained to perform a generalized version of the Mante task (Mante et al., 2013), in which the network is faced with information from multiple pathways and it should base its decision on one of the pathways. Our preliminary results showed that when pyramidal neurons are equipped with multiple dendrites, the network learns to perform pathway-specific gating on dendrites. This study opens doors to investigate more complicated cognitive functions using one of the most biologically detailed recurrent neural network up to date.

## 7.2 Results

### 7.2.1 Training circuit models with multiple cell types

To study how pathway-specific gating could be implemented in a multi-type excitatory-inhibitory neuronal circuit, we build a recurrent neural network model that contains pyramidal neurons and three types of inhibitory neurons (**Fig. 7.1**). Each model pyramidal neuron is multi-compartmental, with five dendrites that feed forward into a somatic compartment. Both dendritic and somatic compartments can receive inputs while only the somatic compartments can project outputs. We include three types of inhibitory neurons that are intended to model PV, SST, and VIP-expressing interneurons in the cortex. Our model neurons capture the mesoscopic-level connectivity between different cell types. More specifically, our model PV neurons target PV neurons and somatic compartments of pyramidal neurons. SST neurons target PV and VIP neurons and dendritic compartments of pyramidal neurons. VIP neurons target SST neurons only. Pyramidal neurons can target all cell types.

After setting the mesoscopic connectivity between different types of neurons, the microscopic connectivity, i.e. the synaptic weights of individual neurons, are set through a supervised learning algorithm. The network is trained to perform a generalized version of the Mante task (Mante et al., 2013). In the original Mante task, the network is presented with stimuli from two pathways, motion and color. Depending on the context, the network should focus on the motion or the color pathway. In the motion context, the network needs to respond to the direction corresponding to the stronger motion stimulus. We generalized the task such that it contains $N_p$ pathways. The network should always base its decision on information from one of the $N_p$ pathways. The identity of

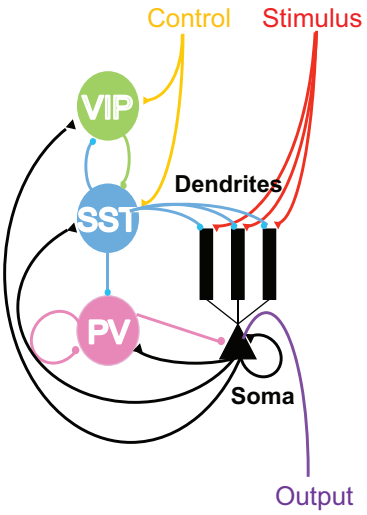


Figure 7.1: Multi-type E-I circuit model schematic. The network model contains multi-compartmental pyramidal neurons and multi types of inhibitory neurons. The structural connections are set according to a canonical microcircuit (Pfeffer et al., 2013). For each population, the number of neurons is set to mimic the proportion it occupies in the mouse prefrontal cortex (Kim et al., 2017).

the relevant pathway is provided by rule inputs targeting the network. There are $N_p$ rule input units. Each corresponds to a pathway to be gated in. Only one of the rule inputs is activated at each trial.

Initial connection weights before training can have a strong impact on whether the network can successfully learn the task. Previously proposed initialization for weights only work for square matrices (Le, Jaitly and Hinton, 2015). In our model, many connections are not allowed because of the mesoscopic structural constraints, for example SST neurons do not target SST neurons. In addition, every connection is sign-constrained, meaning that it is either positive or negative, depending on whether the connection originates from an excitatory or inhibitory neuron. To find a weight initialization that works well for our network model, we systematically tested thousands of weight initialization schemes, and chose the one that works the best overall (see Methods). After choosing the best initialization, the network can learn the generalized Mante task well (**Fig. 7.2**).

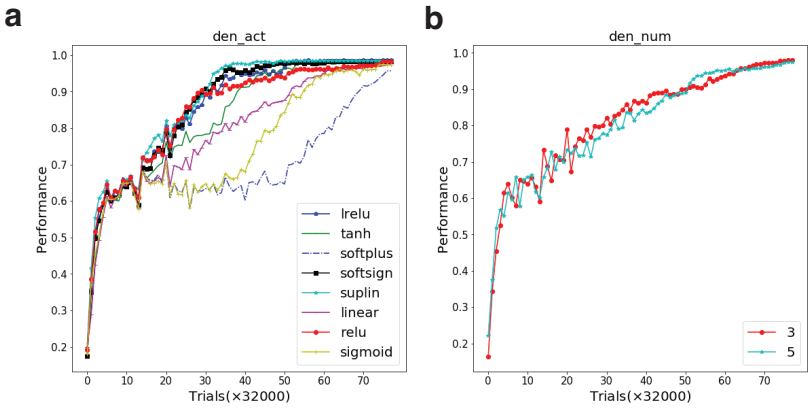


Figure 7.2: Learning curves of the circuit model. After appropriately choosing the initialization, the network model can learn the generalized Mante task well enough when several different dendritic activation functions are used (**a**), and when two different numbers of dendrites-per-neuron are used (**b**).

### 7.2.2 Interneurons develop pathway-specificity

To take the first step in understanding how our multi-type E-I network model solves the generalized Mante task, we studied the neural activity of dendrite and soma of pyramidal neurons in different rules (**Fig. 7.3**). We found that dendrites of pyramidal neurons in the model are strongly selective to rules while somas of pyramidal neurons are not very selective. These results suggest that when equipped with dendrites, the network can learn to perform gating using dendrites instead of somas. In such network, dendrites carry the pathway specificity. By receiving inputs in turn from multiple dendrites, each soma can be mixed for different pathways.

When we trained a network where each pyramidal neuron only has one dendrite, the results are dramatically changed. Dendritic compartments are again selective to the current context or rule. However, instead of being mixed for different pathways, somatic compartments are also pathway-selective in this network.

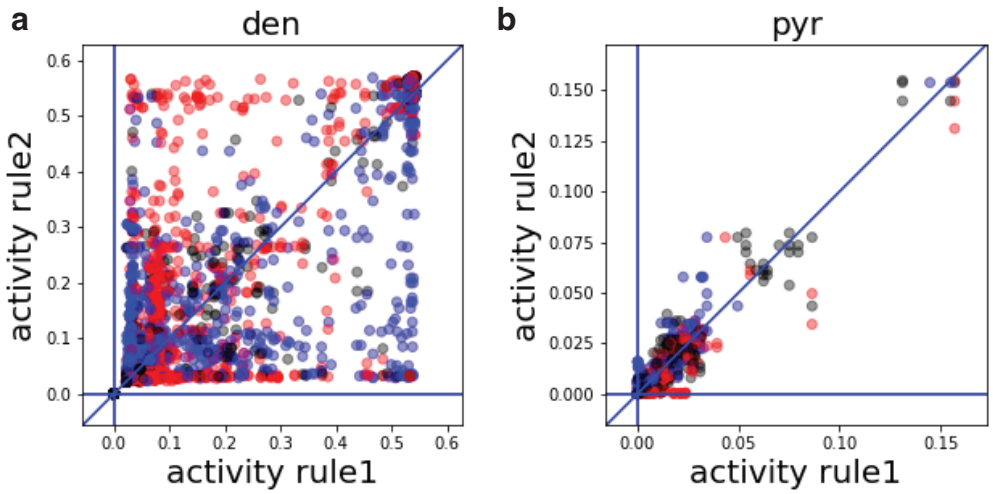


Figure 7.3: Dendrite and soma activity under different rules. (**a**) Activity of dendrites under different contexts. Here we show comparisons for all dendrites under all pairs of contexts. For 5 contexts or rules, there are 10 unique pairs. Each dot is the activity of one dendrite given one pair of rules. (**b**) Similar plot as (**a**) but for soma activity.

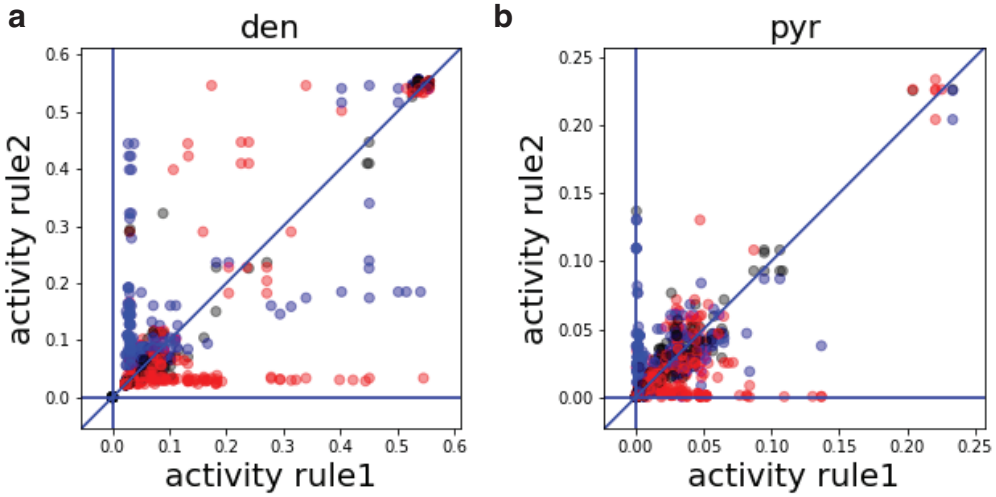


Figure 7.4: Dendrite and soma activity under different rules when each pyramidal neuron has one dendrite. The plot is the same as (**Fig. 7.3**), except that each pyramidal neuron in the model network has only one dendrite. In this network, the pathway selectivity is much stronger on somatic compartments of pyramidal neurons.

## 7.3 Methods

The network setup is similar to the ones used in Chapters 5 and 6, except that there are four types of neurons. Neurons are either excitatory or inhibitory, so the connections are all sign-constrained. The overall structure of the network is set to mimic a canonical microcircuit (Pfeffer et al., 2013).

Each pyramidal neuron contains a fixed number of dendrites. Dendrites are treated as normal units in the network, with fixed output projections to their corresponding somas. The connection weight from a dendrite to a soma is fixed at 1. We used 80 pyramidal neurons, 10 VIP neurons, 6 SST neurons, and 3 PV neurons. Each pyramidal neuron has either five or one dendrite. The activation function for dendrites is the hyperbolic tangent function.

The network receives three kinds of noisy inputs: fixation, stimulus and rule. The fixation and rule inputs target the VIP and SST model neurons, while the stimulus inputs target dendrites of pyramidal neurons. All connection weights from external inputs are constrained to be non-negative.

The read-out, loss function, and regularization are all the same as the ones used in Chapter 6.

The network is trained on a generalized Mante task with five modalities instead of two. Only one modality is relevant to the network in any given trial. The relevant modality is indicated by the rule input activated in the current trial.

# Chapter 8

# Conclusions, discussions, and future directions

Any satisfactory model or theory for any cognitive function should be capable, predictive, and interpretable. Obviously, the model should be capable of the cognitive function in question. It should also be able to predict or at least explain experimental observations regarding the cognitive function. New experimental findings often call for revisions or toppling of previous dominant models. And finally, the model should provide interpretations and insights to the neural mechanism of the cognitive function. Previous computational models usually lack the capability to solve complicated cognitive tasks. Many models are concerned with important but simple cognitive functions and tasks. Models that are capable of sophisticated cognitive tasks typically lack predictive powers at the neuronal level or lack interpretations of their mechanisms. Most models have difficulties explaining neural activity recorded in different areas, layers, and cell types.

In this thesis, we attempt to attack these problems from two fronts. To address the lack of cell-type specificity in most models, we build a neural circuit model in Chapter

2 that hypothesizes functional roles of different types of interneurons in cognitive flexibility. We propose that dendrite-targeting interneurons could control or "gate" the flow of inputs from incoming pathways. A control signal can open the gate for one pathway by exciting a specific subset of VIP and SST neurons, which in turn disinhibit dendritic branches receiving inputs from the pathway to be gated in. In Chapter 2, the interneuronal circuit model we built is essentially feedforward, such that VIP neurons target SST neurons, which in turn target dendrites of pyramidal neurons. The actual neural circuit is highly interconnected. This interconnectivity can lead to interesting neural dynamics. In Chapter 3, we studied how neural dynamics change when varying the cell-density of different types of interneurons. In Chapter 4, we showed that counterintuitive dynamics will naturally arise in a strongly inter-connected circuit. Understanding this counterintuitive dynamics helps us explain experimental phenomena observed in the cortex that have challenged the disinhibitory circuit motif.

On the other front of the attack, we studied neural mechanisms behind more complex cognitive functions. In Chapter 6, we trained and analyzed a recurrent neural network that can perform many cognitive tasks, in order to understand the flexibility and versatility of the prefrontal cortex. Our network is trained to successfully perform 20 cognitive tasks, most of which have been widely used to study cognitive functions, especially in non-human primates. To successfully master all 20 tasks, our network developed clusters of units that are specialized to specific cognitive functions. When instructed to perform a particular cognitive task, the network could flexibly recruit relevant clusters of units that support the cognitive processes currently in need. The use of functionally-specialized clusters can serve as a general principle that explains how the brain is able to perform, and switch between, many different tasks.

Chapter 7 presented preliminary results combining efforts from the two fronts. We trained a recurrent neural network equipped with multi-compartmental pyramidal neurons and multiple types of interneurons. We studied the neural mechanism of the network for a generalized version of the Mante task.

Not surprisingly, our work answered some questions but encourages more. We explored many of these questions in the Discussion section at the end of each chapter. Below I will discuss some more general issues that will frequently go beyond the study of cognitive flexibility.

In Chapters 2-4, we introduced more complexity into our local-circuit models using four distinct types of neurons. However, there are more than 50 types of cells in each area and more than 50 areas in the cortex alone, not to mention subcortical structures. In the end, should we build models or theories that incorporate all types of cells and all areas? Considering that there are around $10^{10}$ neurons in the human cortex, a theory with "simply" $(50 \times 50 =) 2,500$ types of cells is still a tremendous improvement. Ironically, our $10^{10}$-neurons-brain will likely have a hard time remembering a theory of 2,500 components. Will the theory of the future be a 2,500-pages Principles of the Brain, where each page documents our best guess of the role of one cell type in one area?

We are fortunate that this dystopian future is unlikely to happen. Although there are a large number of brain areas and cell types, there also exist principles that can distill the overwhelming complexity of biology into simpler theories. One particularly important principle is the idea of a cortical hierarchy (Felleman and Van Essen, 1991). Instead of treating all cortical areas as distinct structures and describing the property of each area independently, the areas can be described as lying along a cortical hierarchy such that the property of each area can be predicted by its position along the hierarchy. For exam-

ple, the average number of spines on each pyramidal neuron within an area is strongly correlated with the hierarchical position of that area (Chaudhuri et al., 2015). Of course, it is feasible to extend the depiction of areas from a single dimension to multiple dimensions. Discovering principles like the cortical hierarchy greatly reduces the complexity of the theory.

It remains an important task to discover similar principles for the study of different cell types. The goal is to describe all cell types using a small number of factors. Jiang et al. (2015) proposed that cortical interneurons can be classified as one of the three following classes: master regulators that target all neuron types; interneuron-selective interneurons that target other interneurons and disinhibit pyramidal neurons; and pyramidal-neuron-targeting interneurons that target pyramidal neurons and themselves. According to this classification scheme, SST-expressing neurons belong to the master regulators and PV-expressing neurons are pyramidal-neuron-targeting interneurons. Meanwhile, SST neurons mainly target dendrites of pyramidal neurons and PV neurons mainly target perisomatic areas of pyramidal neurons. Therefore, there could be a close relationship between an interneuron's projection pattern onto different cell types and its sub-cellular projection pattern onto pyramidal neurons. SST neurons also have longer time constants in comparison to PV neurons. It is possible that a small number of factors can explain the variation of properties across all types of interneurons, including local- and long-range projection patterns, subcellular projection patterns, and time scales of dynamics.

Another fundamental principle is the concept of canonical microcircuits (Douglas and Martin, 1991), which are circuits repeated through every area in the cortex. Once we understand the function of a canonical microcircuit in one area, the knowledge can be generalized to provide a decent qualitative understanding of the same circuit in other

cortical areas. Combining the principles of cortical hierarchy and canonical microcircuit, we could build a large-scale model of the cortex where a qualitatively similar circuit motif is repeated across cortex, and quantitative parameters of the circuit vary according to the cortical hierarchy (for example, see Chaudhuri et al. (2015)).

In Chapter 2, we proposed that the dendritic disinhibitory circuit can perform pathway-specific gating for cognitive flexibility. However, cognitive flexibility is considered mostly a function of the prefrontal cortex, whereas the interneuronal circuit motif is repeated throughout the cortex, from sensory and motor to prefrontal cortices (Pi et al., 2013; Jiang et al., 2015). A unifying theory of the interneuronal canonical circuit should also explain its functional role in lower-order areas. There is arguably a need for pathway-specific gating in every cortical area, not just the prefrontal cortex. Even area V1 receives inputs from and projects outputs to more than 10 cortical areas in monkeys and mice (Markov et al., 2014; Zingg et al., 2014; **?**). A common function of dendrite- and soma-targeting interneurons across cortical areas could be to regulate how information flow from and to different areas respectively.

In Chapters 5-7, we used the approach of trained network analysis to study potential neural mechanisms of cognitive flexibility. This approach makes it easier than ever to generate network models for complicated functions. However, cautions need to be exercised when studying neural mechanisms using this approach, as it is accompanied by clear, fundamental problems.

The core problem of studying trained neural networks is the issue of interpretability. This issue can take several forms. In the most obvious form, having trained a network does not in itself lend much, if any, understanding to the underlying neural mechanism. So it is necessary to inspect the trained network by analyzing neural activity, network

connectivity, and dynamics (for example, see Mante et al. (2013)). The more fundamental form of the issue is that modelers have little interpretable control over the learned mechanism. Consider a typical bottom-up model, there are usually a handful of free parameters controlled by the modeler, and each parameter has a clear functional interpretation. For example, in the classical drift-diffusion model (Ratcliff, 1978), commonly used to study perceptual decision-making, there are several free parameters, including the boundary level and the drift rate. There exist both informal intuitions and formal mathematical derivations showing how individual parameters influence the model results. However, when training neural networks, modelers can only make structural choices, like choosing the network architecture and neuronal activation functions, and have access to hyperparameters, like the learning rate and relative strengths of various regularizations. The link between what modelers can control (structure and hyperparameters) and what modelers intend to study (neural mechanism) is indirect at best, limiting the model's interpretability. Therefore, trained networks should be used after hand-designed models have failed at performing the task or explaining neural data.

Although cognitive functions typically involve more than one area, most trained networks that are used to study cognitive functions contain a single recurrently-connected model area. It remains unclear how to meaningfully model multiple areas. For deep feed-forward networks, it is relatively straightforward to map different layers to areas along the sensory hierarchy (Yamins et al., 2014). Higher layers correspond to higher-order sensory areas and develop more complex and abstract representations. However, there is no standard way to map different "areas" in a recurrent neural network to areas in the associative cortex. Song, Yang and Wang (2017) modeled different areas by having a separate objective function for the readout from each area. In particular, the decision of the network

is read out from one area used to model the prefrontal cortex, while value of the current state is read out from the other area modeling the orbitofrontal cortex. Generalizing this approach to many areas appear challenging, because it would be difficult to define a distinct objective function for each area. Another approach is to distinguish model areas using structural connectivity. Suppose we build a chain of recurrent networks, with the first area on the chain receiving sensory inputs and the last area projecting to motor outputs. Areas are also recurrently connected with neighboring areas on the chain. It is conceivable that earlier areas on the chain will develop representations more closely related to sensory processing while later areas develop motor representations.

Finally, it is important to acknowledge that the learning algorithm we used, in particular backpropagation, is not biologically realistic. More biologically plausible plasticity rules are currently available but rarely work as well as gradient-based methods. There are also efforts trying to link backpropagation with spike-timing dependent plasticity in feedforward networks (Scellier and Bengio, 2016). What is truly remarkable is that, despite the difference in the learning processes, artificial and biological neural networks can often develop similar solutions to the same computational problem. This phenomenon is reminiscent of convergent evolution, where organisms far apart in the evolutionary tree ended up developing similar features in response to the same kind of environmental pressure. Perhaps, when artificial neural networks developed similar solutions as the brain did, we can call it "convergent optimization". For convergent optimization to happen, there need to be enough structural similarity between the two systems. An important goal for future research would be to understand the condition and extent of convergent optimization.

# Bibliography


Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu and Xiaoqiang Zheng. 2015. "TensorFlow: Large-scale machine learning on heterogeneous systems.". Software available from tensorflow.org.
**URL:** *http://tensorflow.org/*

Abbott, L F and Frances S Chance. 2005. "Drivers and modulators from push-pull and balanced synaptic input." *Prog Brain Res* 149:147–55.

Abbott, LF. 2006. "Where are the switches on this thing." *23 problems in systems neuroscience* pp. 423–31.

Adesnik, Hillel, William Bruns, Hiroki Taniguchi, Z. Josh Huang and Massimo Scanziani. 2012. "A neural circuit for spatial summation in visual cortex." *Nature* 490(7419):226–231.
**URL:** *http://www.nature.com/nature/journal/v490/n7419/full/nature11526.html*

Adler, A. and W.-B. Gan. 2015. "Somatostatin interneurons exhibit diverse activities during motor learning." *Program No. 357.10. 2015 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience, 2015. Online.* .

Akam, Thomas and Dimitri M Kullmann. 2010. "Oscillations and filtering networks support flexible routing of information." *Neuron* 67(2):308–20.

Ali, Afia B and Alex M Thomson. 2008. "Synaptic alpha 5 subunit-containing GABAA receptors mediate IPSPs elicited by dendrite-preferring cells in rat neocortex." *Cereb Cortex* 18(6):1260–71.

Ardid, Salva and Xiao-Jing Wang. 2013. "A tweaking principle for executive control: neuronal circuit mechanism for rule-based task switching and conflict resolution." *J Neurosci* 33(50):19504–17.

Ascher, P and L Nowak. 1988. "The role of divalent cations in the N-methyl-D-aspartate responses of mouse central neurones in culture." *J Physiol* 399:247–66.

Averbeck, Bruno B and Daeyeol Lee. 2007. "Prefrontal neural correlates of memory for sequences." *Journal of Neuroscience* 27(9):2204–2211.

Badre, David. 2008. "Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes." *Trends Cogn Sci* 12(5):193–200.

Bakker, Bram. 2002. Reinforcement learning with long short-term memory. In *Advances in neural information processing systems*. pp. 1475–1482.

Baldo, J V, A P Shimamura, D C Delis, J Kramer and E Kaplan. 2001. "Verbal and design fluency in patients with frontal lobe lesions." *J Int Neuropsychol Soc* 7(5):586–96.

Bar-Ilan, Lital, Albert Gidon and Idan Segev. 2012. "The role of dendritic inhibition in shaping the plasticity of excitatory synapses." *Front Neural Circuits* 6:118.

Barak, Omri, David Sussillo, Ranulfo Romo, Misha Tsodyks and L F Abbott. 2013. "From fixed points to chaos: three models of delayed discrimination." *Prog Neurobiol* 103:214–22.

Barak, Omri and Misha Tsodyks. 2014. "Working models of working memory." *Current opinion in neurobiology* 25:20–24.

Barak, Omri, Misha Tsodyks and Ranulfo Romo. 2010. "Neuronal population coding of parametric working memory." *J Neurosci* 30(28):9424–30.

Bayraktar, T, E Welker, T F Freund, K Zilles and J F Staiger. 2000. "Neurons immunoreactive for vasoactive intestinal polypeptide in the rat primary somatosensory cortex: morphology and spatial relationship to barrel-related columns." *J Comp Neurol* 420(3):291–304.

Benda, Jan and Andreas VM Herz. 2003. "A universal model for spike-frequency adaptation." *Neural computation* 15(11):2523–2564.

Bengio, Yoshua, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard and Zhouhan Lin. 2015. "Towards biologically plausible deep learning." *arXiv preprint arXiv:1502.04156* .

Bengio, Yoshua, Nicolas Boulanger-Lewandowski and Razvan Pascanu. 2013. Advances in optimizing recurrent networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE pp. 8624–8628.

Bengio, Yoshua, Patrice Simard and Paolo Frasconi. 1994. "Learning long-term dependencies with gradient descent is difficult." *Neural Networks, IEEE Transactions on* 5(2):157–166.

Bengio, Yoshua, Thomas Mesnard, Asja Fischer, Saizheng Zhang and Yuhuai Wu. 2015. "An objective function for STDP." *arXiv preprint arXiv:1509.05936* .

Benna, Marcus K and Stefano Fusi. 2016. "Computational principles of synaptic memory consolidation." *Nat Neurosci* 19(12):1697–1706.

Berg, Esta A. 1948. "A simple objective technique for measuring flexibility in thinking." *The Journal of general psychology* 39(1):15–22.

Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley and Yoshua Bengio. 2010. Theano: A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf.* pp. 1–7.

Binzegger, Tom, Rodney J Douglas and Kevan AC Martin. 2004. "A quantitative map of the circuit of cat primary visual cortex." *Journal of Neuroscience* 24(39):8441–8453.

Branco, Tiago, Beverley A Clark and Michael Häusser. 2010. "Dendritic discrimination of temporal input sequences in cortical neurons." *Science* 329(5999):1671–5.

Britten, K H, M N Shadlen, W T Newsome and J A Movshon. 1993. "Responses of neurons in macaque MT to stochastic motion signals." *Vis Neurosci* 10(6):1157–69.

Brunton, Bingni W, Matthew M Botvinick and Carlos D Brody. 2013. "Rats and humans can optimally accumulate evidence for decision-making." *Science* 340(6128):95–98.

Cajal, Santiago Ramón. 1911. "Histologie du syste me nerveux de l'Homme et des verte be s." *Maloine (Paris)* 2:891–942.

Carnevale, Federico, Victor de Lafuente, Ranulfo Romo, Omri Barak and Néstor Parga. 2015. "Dynamic control of response criterion in premotor cortex during perceptual detection under temporal uncertainty." *Neuron* 86(4):1067–77.

Caruana, Rich. 1997. "Multitask learning." *Machine Learning* pp. 41–75.

Chaisangmongkon, Warasinee, Sruthi K Swaminathan, David J Freedman and Xiao-Jing Wang. 2017. "Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions." *Neuron* 93(6):1504–1517.

Chaudhuri, Rishidev, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy and Xiao-Jing Wang. 2015. "A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex." *Neuron* 88(2):419–31.

Chen, Simon X, An Na Kim, Andrew J Peters and Takaki Komiyama. 2015. "Subtype-specific plasticity of inhibitory circuits in motor cortex during motor learning." *Nat Neurosci* .

Chiu, Chiayu Q, Gyorgy Lur, Thomas M Morse, Nicholas T Carnevale, Graham C R Ellis-Davies and Michael J Higley. 2013. "Compartmentalization of GABAergic inhibition by dendritic spines." *Science* 340(6133):759–62.

Christophel, Thomas B, P Christiaan Klink, Bernhard Spitzer, Pieter R Roelfsema and John-Dylan Haynes. 2017. "The distributed nature of working memory." *Trends Cogn Sci* 21(2):111–124.

Churchland, Mark M, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu and Krishna V Shenoy. 2012. "Neural population dynamics during reaching." *Nature* 487(7405):51–6.

Cichon, Joseph and Wen-Biao Gan. 2015. "Branch-specific dendritic Ca(2+) spikes cause persistent synaptic plasticity." *Nature* 520(7546):180–5.

Cohen, J D, K Dunbar and J L McClelland. 1990. "On the control of automatic processes: a parallel distributed processing account of the Stroop effect." *Psychol Rev* 97(3):332–61.

Cole, Michael W, Jeremy R Reynolds, Jonathan D Power, Grega Repovs, Alan Anticevic and Todd S Braver. 2013. "Multi-task connectivity reveals flexible hubs for adaptive task control." *Nat Neurosci* 16(9):1348–55.

Cole, Michael W, Joset A Etzel, Jeffrey M Zacks, Walter Schneider and Todd S Braver. 2011. "Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex." *Front Hum Neurosci* 5:142.

Cole, Michael W, Patryk Laurent and Andrea Stocco. 2013. "Rapid instructed task learning: a new window into the human brain's unique capacity for flexible cognitive control." *Cogn Affect Behav Neurosci* 13(1):1–22.

Cunningham, John P and Byron M Yu. 2014. "Dimensionality reduction for large-scale neural recordings." *Nat Neurosci* 17(11):1500–9.

Dan, Yang and Mu-Ming Poo. 2006. "Spike timing-dependent plasticity: from synapse to perception." *Physiological reviews* 86(3):1033–1048.

Dauphin, Yann N, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli and Yoshua Bengio. 2014. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*. pp. 2933–2941.

Diamond, Adele. 2013. "Executive functions." *Annual review of psychology* 64:135–168.

Dipoppa, Mario, Adam Ranson, Michael Krumin, Marius Pachitariu, Matteo Carandini and Kenneth D Harris. 2016. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. Technical Report biorxiv;058396v1.
**URL:** *http://biorxiv.org/lookup/doi/10.1101/058396*

Douglas, R J and K A Martin. 1991. "A functional microcircuit for cat visual cortex." *J Physiol* 440:735–69.

Druckmann, Shaul, Linqing Feng, Bokyoung Lee, Chaehyun Yook, Ting Zhao, Jeffrey C Magee and Jinhyun Kim. 2014. "Structured synaptic connectivity between hippocampal regions." *Neuron* 81(3):629–40.

Eccles, J C, P Fatt and K Koketsu. 1954. "Cholinergic and inhibitory synapses in a pathway from motor-axon collaterals to motoneurones." *J Physiol* 126(3):524–62.

Eliasmith, Chris, Terrence C Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, Charlie Tang and Daniel Rasmussen. 2012. "A large-scale model of the functioning brain." *Science* 338(6111):1202–5.

Elston, Guy N, Ruth Benavides-Piccione and Javier Defelipe. 2005. "A study of pyramidal cell structure in the cingulate cortex of the macaque monkey with comparative notes on inferotemporal and primary visual cortex." *Cereb Cortex* 15(1):64–73.

Engel, Tatiana A and Xiao-Jing Wang. 2011. "Same or different? A neural circuit mechanism of similarity-based pattern match decision making." *Journal of Neuroscience* 31(19):6982–6996.

Ercsey-Ravasz, Mária, Nikola T Markov, Camille Lamy, David C Van Essen, Kenneth Knoblauch, Zoltán Toroczkai and Henry Kennedy. 2013. "A predictive network model of cerebral cortical connectivity based on a distance rule." *Neuron* 80(1):184–97.

Felleman, D J and D C Van Essen. 1991. "Distributed hierarchical processing in the primate cerebral cortex." *Cereb Cortex* 1(1):1–47.

Festa, Dylan, Guillaume Hennequin and Máté Lengyel. 2014. Analog memories in a balanced rate-based network of EI neurons. In *Advances in Neural Information Processing Systems*. pp. 2231–2239.

Fino, Elodie and Rafael Yuste. 2011. "Dense inhibitory connectivity in neocortex." *Neuron* 69(6):1188–203.

Freedman, David J and John A Assad. 2016. "Neuronal mechanisms of visual categorization: an abstract view on decision making." *Annu Rev Neurosci* 39:129–47.

Freund, T F and G Buzsáki. 1996. "Interneurons of the hippocampus." *Hippocampus* 6(4):347–470.

Fu, Min, Xinzhu Yu, Ju Lu and Yi Zuo. 2012. "Repetitive motor learning induces coordinated formation of clustered dendritic spines in vivo." *Nature* 483(7387):92–5.

Fu, Yu, Jason M Tucciarone, J Sebastian Espinosa, Nengyin Sheng, Daniel P Darcy, Roger A Nicoll, Z Josh Huang and Michael P Stryker. 2014. "A cortical circuit for gain control by behavioral state." *Cell* 156(6):1139–52.

Fu, Yu, Megumi Kaneko, Yunshuo Tang, Arturo Alvarez-Buylla and Michael P Stryker. 2014. "A cortical disinhibitory circuit for enhancing adult plasticity." *Elife* 4:e05558.

Funahashi, Ken-ichi and Yuichi Nakamura. 1993. "Approximation of dynamical systems by continuous time recurrent neural networks." *Neural networks* 6(6):801–806.

Funahashi, S, C J Bruce and P S Goldman-Rakic. 1989. "Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex." *J Neurophysiol* 61(2):331–49.

Fuster, Joaquin. 2015. *The Prefrontal Cortex*. Fifth edition ed. Academic Press.

Gao, Peiran and Surya Ganguli. 2015. "On simplicity and complexity in the brave new world of large-scale neuroscience." *Curr Opin Neurobiol* 32:148–55.

Gelb, Adhémar and Kurt Goldstein. 1925. "Psychologische analysen hirnpathologischer fälle." *Psychologische Forschung* 6(1):127–186.

Gentet, Luc J, Yves Kremer, Hiroki Taniguchi, Z Josh Huang, Jochen F Staiger and Carl C H Petersen. 2012. "Unique functional properties of somatostatin-expressing GABAergic neurons in mouse barrel cortex." *Nat Neurosci* 15(4):607–12.

Gerstner, W. 2000. "Population dynamics of spiking neurons: fast transients, asynchronous states, and locking." *Neural Computation* 12(1):43–89.

Gerstner, Wulfram and Werner Kistler. 2002. "Spiking Neuron Models Cambridge University Press.".

Gillespie, Daniel T. 1996. "The mathematics of Brownian motion and Johnson noise." *American Journal of Physics* 64(3):225–240.

Gjorgjieva, Julijana, Guillaume Drion and Eve Marder. 2016. "Computational implications of biophysical diversity and multiple timescales in neurons and synapses for circuit performance." *Current opinion in neurobiology* 37:44–52.

Gold, Joshua I and Michael N Shadlen. 2007. "The neural basis of decision making." *Annu Rev Neurosci* 30:535–74.

Goodman, Dan and Romain Brette. 2008. "Brian: a simulator for spiking neural networks in python." *Front Neuroinform* 2:5.

Gough, Harrison G. 1956. "California psychological inventory.".

Graupner, Michael and Nicolas Brunel. 2012. "Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location." *Proc Natl Acad Sci U S A* 109(10):3991–6.

Hahnloser, Richard LT. 1998. "On the piecewise analysis of networks of linear threshold neurons." *Neural Networks* 11(4):691–697.

Hanes, D P, W F Patterson, 2nd and J D Schall. 1998. "Role of frontal eye fields in countermanding saccades: visual, movement, and fixation activity." *J Neurophysiol* 79(2):817–34.

Hardt, Moritz, Benjamin Recht and Yoram Singer. 2015. "Train faster, generalize better: Stability of stochastic gradient descent." *arXiv preprint arXiv:1509.01240* .

Hawrylycz, Michael et al. 2016. "Inferring cortical function in the mouse visual system through large-scale systems neuroscience." *Proceedings of the National Academy of Sciences* 113(27):7337–7344.
**URL:** *http://www.pnas.org/content/113/27/7337*

Helmstaedter, Moritz, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung and Winfried Denk. 2013. "Connectomic reconstruction of the inner plexiform layer in the mouse retina." *Nature* 500(7461):168–74.

Higgins, David, Michael Graupner and Nicolas Brunel. 2014. "Memory maintenance in synapses with calcium-based plasticity in the presence of background activity." *PLoS Comput Biol* 10(10):e1003834.

Hines, M L and N T Carnevale. 1997. "The NEURON simulation environment." *Neural*

*Comput* 9(6):1179–209.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. "Long short-term memory." *Neural computation* 9(8):1735–1780.

Ibrahim, Leena A., Lukas Mesik, Xu-ying Ji, Qi Fang, Hai-fu Li, Ya-tang Li, Brian Zingg, Li I. Zhang and Huizhong Whit Tao. 2016. "Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition." *Neuron* 89(5):1031–1045.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0896627316000520*

Ishikawa, Taro, Yoshinori Sahara and Tomoyuki Takahashi. 2002. "A single packet of transmitter does not saturate postsynaptic glutamate receptors." *Neuron* 34(4):613–21.

Isoda, Masaki and Okihide Hikosaka. 2007. "Switching from automatic to controlled action by monkey medial frontal cortex." *Nat Neurosci* 10(2):240–8.

Izhikevich, Eugene M. 2007. "Solving the distal reward problem through linkage of STDP and dopamine signaling." *Cerebral cortex* 17(10):2443–2452.

Jackson, Jesse, Inbal Ayzenshtat, Mahesh M. Karnani and Rafael Yuste. 2016. "VIP+ interneurons control neocortical activity across brain states." *Journal of Neurophysiology* 115(6):3008–3017.
**URL:** *http://jn.physiology.org/content/115/6/3008*

Jadi, Monika, Alon Polsky, Jackie Schiller and Bartlett W Mel. 2012. "Location-dependent effects of inhibition on local spiking in pyramidal neuron dendrites." *PLoS Comput Biol* 8(6):e1002550.

Jiang, Xiaolong, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel and Andreas S Tolias. 2015. "Principles of connectivity among morphologically defined cell types in adult neocortex." *Science* 350(6264):aac9462.

Johnston, Kevin, Helen M Levin, Michael J Koval and Stefan Everling. 2007. "Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching." *Neuron* 53(3):453–62.

Karnani, Mahesh M, Masakazu Agetsuma and Rafael Yuste. 2014. "A blanket of inhibition: functional inferences from dense inhibitory connectivity." *Curr Opin Neurobiol* 26:96–102.

Kastellakis, George, Denise J Cai, Sara C Mednick, Alcino J Silva and Panayiota Poirazi. 2015. "Synaptic clustering within dendrites: An emerging theory of memory formation." *Prog Neurobiol* .

Keller, Georg B., Tobias Bonhoeffer and Mark Hübener. 2012. "Sensorimotor mismatch signals in primary visual cortex of the behaving mouse." *Neuron* 74(5):809–815.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0896627312003844*

Kepecs, A and S Raghavachari. 2007. "Gating information by two-state membrane potential fluctuations." *J Neurophysiol* 97(4):3015–23.

Kepecs, Adam and Gordon Fishell. 2014. "Interneuron cell types are fit to function." *Nature* 505(7483):318–26.

Kiani, Roozbeh, Timothy D Hanks and Michael N Shadlen. 2008. "Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment." *Journal of Neuroscience* 28(12):3017–3029.

Kim, Yongsoo, Guangyu Robert Yang, Kith Pradhan, Kannan Umadevi Venkataraju, Mihail Bota, Luis Carlos García Del Molino, Greg Fitzgerald, Keerthi Ram, Miao He, Jesse Maurica Levine, Partha Mitra, Z Josh Huang, Xiao-Jing Wang and Pavel Osten. 2017. "Brain-wide Maps Reveal Stereotyped Cell-Type-Based Cortical Architecture and Subcortical Sexual Dimorphism." *Cell* 171(2):456–469.e22.

Kim, Yongsoo, Kannan Umadevi Venkataraju, Kith Pradhan, Carolin Mende, Julian Taranda, Srinivas C Turaga, Ignacio Arganda-Carreras, Lydia Ng, Michael J Hawrylycz, Kathleen S Rockland, H Sebastian Seung and Pavel Osten. 2015. "Mapping social behavior-induced brain activation at cellular resolution in the mouse." *Cell Rep* 10(2):292–305.

Kingma, Diederik and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* .

Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran and Raia Hadsell. 2017. "Overcoming catastrophic forgetting in neural networks." *Proc Natl Acad Sci U S A* 114(13):3521–3526.

Kleindienst, Thomas, Johan Winnubst, Claudia Roth-Alpermann, Tobias Bonhoeffer and Christian Lohmann. 2011. "Activity-dependent clustering of functional synaptic inputs on developing hippocampal dendrites." *Neuron* 72(6):1012–24.

Koch, C, T Poggio and V Torre. 1982. "Retinal ganglion cells: a functional interpretation of dendritic morphology." *Philos Trans R Soc Lond B Biol Sci* 298(1090):227–63.

Koechlin, Etienne, Chrystèle Ody and Frédérique Kouneiher. 2003. "The architecture of cognitive control in the human prefrontal cortex." *Science* 302(5648):1181–5.

Kuchibhotla, Kishore V., Jonathan V. Gill, Grace W. Lindsay, Eleni S. Papadoyannis, Rachel E. Field, Tom A. Hindmarsh Sten, Kenneth D. Miller and Robert C. Froemke. 2016. "Parallel processing by cortical inhibition enables context-dependent behavior." *Nature Neuroscience* advance online publication.
**URL:** *http://www.nature.com/neuro/journal/vaop/ncurrent/full/nn.4436.html*

Kvitsiani, D, S Ranade, B Hangya, H Taniguchi, J Z Huang and A Kepecs. 2013. "Distinct behavioural and network correlates of two interneuron types in prefrontal cortex." *Nature* 498(7454):363–6.

Laje, Rodrigo and Dean V Buonomano. 2013. "Robust timing and motor patterns by taming chaos in recurrent neural networks." *Nat Neurosci* 16(7):925–33.

Larkman, A U. 1991. "Dendritic morphology of pyramidal neurones of the visual cortex of the rat: III. Spine distributions." *J Comp Neurol* 306(2):332–43.

Larkum, Matthew E, Walter Senn and Hans-R Lüscher. 2004. "Top-down dendritic input increases the gain of layer 5 pyramidal neurons." *Cereb Cortex* 14(10):1059–70.

Le, Quoc V, Navdeep Jaitly and Geoffrey E Hinton. 2015. "A simple way to initialize recurrent networks of rectified linear units." *arXiv preprint arXiv:1504.00941* .

Lee, Jung H, Christof Koch and Stefan Mihalas. 2017. "A Computational Analysis of the Function of Three Inhibitory Cell Types in Contextual Visual Processing." *Front Comput Neurosci* 11:28.

Lee, Jung Hoon and Stefan Mihalas. 2016. "Visual processing mode switching regulated by VIP cells." *bioRxiv* p. 084632.
**URL:** *http://biorxiv.org/content/early/2016/10/31/084632*

Lee, Seung-Hee, Alex C Kwan, Siyu Zhang, Victoria Phoumthipphavong, John G Flannery, Sotiris C Masmanidis, Hiroki Taniguchi, Z Josh Huang, Feng Zhang, Edward S Boyden, Karl Deisseroth and Yang Dan. 2012. "Activation of specific interneurons improves V1 feature selectivity and visual perception." *Nature* 488(7411):379–83.

Lee, Soohyun, Illya Kruglikov, Z Josh Huang, Gord Fishell and Bernardo Rudy. 2013. "A disinhibitory circuit mediates motor integration in the somatosensory cortex." *Nat Neurosci* 16(11):1662–70.

Lee, SooHyun, Jens Hjerling-Leffler, Edward Zagha, Gord Fishell and Bernardo Rudy. 2010. "The largest group of superficial neocortical GABAergic interneurons expresses ionotropic serotonin receptors." *J Neurosci* 30(50):16796–808.

Levy, Robert B and Alex D Reyes. 2012. "Spatial profile of excitatory and inhibitory

synaptic connectivity in mouse primary auditory cortex." *Journal of Neuroscience* 32(16):5609–5619.

Litwin-Kumar, Ashok, Robert Rosenbaum and Brent Doiron. 2016. "Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes." *J Neurophysiol* 115(3):1399–409.

Lovett-Barron, Matthew, Gergely F Turi, Patrick Kaifosh, Peter H Lee, Frédéric Bolze, Xiao-Hua Sun, Jean-François Nicoud, Boris V Zemelman, Scott M Sternson and Attila Losonczy. 2012. "Regulation of neuronal input transformations by tunable dendritic inhibition." *Nat Neurosci* 15(3):423–30, S1–3.

Lu, Jiang-teng, Cheng-yu Li, Jian-Ping Zhao, Mu-ming Poo and Xiao-hui Zhang. 2007. "Spike-timing-dependent plasticity of neocortical excitatory synapses on inhibitory interneurons depends on target cell type." *J Neurosci* 27(36):9711–20.

Ma, Wen-pei, Bao-hua Liu, Ya-tang Li, Z Josh Huang, Li I Zhang and Huizhong W Tao. 2010. "Visual representations by cortical somatostatin inhibitory neurons–selective but with weak and delayed responses." *J Neurosci* 30(43):14371–9.

Machens, Christian K, Ranulfo Romo and Carlos D Brody. 2010. "Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex." *J Neurosci* 30(1):350–60.

Major, Guy, Matthew E Larkum and Jackie Schiller. 2013. "Active properties of neocortical pyramidal neuron dendrites." *Annu Rev Neurosci* 36:1–24.

Mansouri, Farshad A, Kenji Matsumoto and Keiji Tanaka. 2006. "Prefrontal cell activities related to monkeys' success and failure in adapting to rule changes in a Wisconsin Card Sorting Test analog." *J Neurosci* 26(10):2745–56.

Mante, Valerio, David Sussillo, Krishna V Shenoy and William T Newsome. 2013. "Context-dependent computation by recurrent dynamics in prefrontal cortex." *Nature* 503(7474):78–84.

Marder, Eve, Marie L Goeritz and Adriane G Otopalik. 2015. "Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms." *Current opinion in neurobiology* 31:156–163.

Markov, N T, M M Ercsey-Ravasz, A R Ribeiro Gomes, C Lamy, L Magrou, J Vezoli, P Misery, A Falchier, R Quilodran, M A Gariel et al. 2014. "A weighted and directed interareal connectivity matrix for macaque cerebral cortex." *Cereb Cortex* 24(1):17–36.

Markram, Henry, Maria Toledo-Rodriguez, Yun Wang, Anirudh Gupta, Gilad Silberberg

and Caizhi Wu. 2004. “Interneurons of the neocortical inhibitory system.” *Nat Rev Neurosci* 5(10):793–807.

Marlin, Joseph J and Adam G Carter. 2014. “GABA-A receptor inhibition of local calcium signaling in spines and dendrites.” *J Neurosci* 34(48):15898–911.

Martens, James and Ilya Sutskever. 2011. Learning recurrent neural networks with hessian-free optimization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 1033–1040.

Mejias, Jorge F. and André Longtin. 2014. “Differential effects of excitatory and inhibitory heterogeneity on the gain and asynchronous state of sparse cortical networks.” *Frontiers in Computational Neuroscience* 8.
**URL:** *http://journal.frontiersin.org/article/10.3389/fncom.2014.00107/abstract*

Mesik, Lukas, Wen-pei Ma, Ling-yun Li, Leena A. Ibrahim, Z. J. Huang, Li I. Zhang and Huizhong W. Tao. 2015. “Functional response properties of VIP-expressing inhibitory neurons in mouse visual and auditory cortex.” *Frontiers in Neural Circuits* 9.
**URL:** *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4460767/*

Meyer, Hanno S, Daniel Schwarz, Verena C Wimmer, Arno C Schmitt, Jason N D Kerr, Bert Sakmann and Moritz Helmstaedter. 2011. “Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5A.” *Proc Natl Acad Sci U S A* 108(40):16807–12.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. 2013. “Distributed representations of words and phrases and their compositionality.” *Advances in neural information processing systems* pp. 3111–3119.

Miller, E K, C A Erickson and R Desimone. 1996. “Neural mechanisms of visual working memory in prefrontal cortex of the macaque.” *J Neurosci* 16(16):5154–67.

Miller, E K and J D Cohen. 2001. “An integrative theory of prefrontal cortex function.” *Annu Rev Neurosci* 24:167–202.

Miller, G A. 1956. “The magical number seven plus or minus two: some limits on our capacity for processing information.” *Psychol Rev* 63(2):81–97.

Miller, Paul, Carlos D Brody, Ranulfo Romo and Xiao-Jing Wang. 2003. “A recurrent network model of somatosensory parametric working memory in the prefrontal cortex.” *Cerebral Cortex* 13(11):1208–1218.

Milner, Brenda. 1963. “Effects of different brain lesions on card sorting: The role of the frontal lobes.” *Archives of neurology* 9(1):90–100.

Mirabella, Giovanni, Giuseppe Bertini, Inés Samengo, Bjørg E Kilavik, Deborah Frilli, Chiara Della Libera and Leonardo Chelazzi. 2007. "Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories." *Neuron* 54(2):303–18.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg and Demis Hassabis. 2015. "Human-level control through deep reinforcement learning." *Nature* 518(7540):529–33.

Montufar, Guido F, Razvan Pascanu, Kyunghyun Cho and Yoshua Bengio. 2014. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*. pp. 2924–2932.

Munoz, Douglas P and Stefan Everling. 2004. "Look away: the anti-saccade task and the voluntary control of eye movement." *Nat Rev Neurosci* 5(3):218–28.

Murphy, Brendan K and Kenneth D Miller. 2003. "Multiplicative gain changes are induced by excitation or inhibition alone." *J Neurosci* 23(31):10040–51.

Murphy, Brendan K and Kenneth D Miller. 2009. "Balanced amplification: a new mechanism of selective amplification of neural activity patterns." *Neuron* 61(4):635–48.

Nevian, Thomas and Bert Sakmann. 2006. "Spine Ca2+ signaling in spike-timing-dependent plasticity." *J Neurosci* 26(43):11001–13.

Nevian, Thomas, Matthew E Larkum, Alon Polsky and Jackie Schiller. 2007. "Properties of basal dendrites of layer 5 pyramidal neurons: a direct patch-clamp recording study." *Nat Neurosci* 10(2):206–14.

Newsome, William T, Kenneth H Britten and J Anthony Movshon. 1989. "Neuronal correlates of a perceptual decision." *Nature* 341(6237):52–54.

Neymotin, Samuel A, George L Chadderdon, Cliff C Kerr, Joseph T Francis and William W Lytton. 2013. "Reinforcement learning of two-joint virtual arm reaching in a computer model of sensorimotor cortex." *Neural computation* 25(12):3263–3293.

Niell, Cristopher M. and Michael P. Stryker. 2010. "Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex." *Neuron* 65(4):472–479.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0896627310000590*

Olshausen, B A, C H Anderson and D C Van Essen. 1993. "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of infor-

mation." *J Neurosci* 13(11):4700–19.

Ozeki, Hirofumi, Ian M. Finn, Evan S. Schaffer, Kenneth D. Miller and David Ferster. 2009. "Inhibitory stabilization of the cortical network underlies visual surround suppression." *Neuron* 62(4):578–592.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0896627309002876*

Padoa-Schioppa, Camillo and John A Assad. 2006. "Neurons in the orbitofrontal cortex encode economic value." *Nature* 441(7090):223–6.

Pakan, Janelle MP, Scott C. Lowe, Evelyn Dylda, Sander W. Keemink, Stephen P. Currie, Christopher A. Coutts and Nathalie L. Rochefort. 2016. "Behavioral-state modulation of inhibition is context-dependent and cell type specific in mouse visual cortex." *eLife* 5:e14985.
**URL:** *https://elifesciences.org/content/5/e14985v4*

Pascanu, Razvan, Guido Montufar and Yoshua Bengio. 2013. "On the number of response regions of deep feed forward networks with piece-wise linear activations." *arXiv preprint arXiv:1312.6098* .

Pascanu, Razvan, Tomas Mikolov and Yoshua Bengio. 2012. "On the difficulty of training recurrent neural networks." *arXiv preprint arXiv:1211.5063* .

Petreanu, Leopoldo, Tianyi Mao, Scott M Sternson and Karel Svoboda. 2009. "The subcellular organization of neocortical excitatory connections." *Nature* 457(7233):1142–5.

Pfeffer, Carsten K, Mingshan Xue, Miao He, Z Josh Huang and Massimo Scanziani. 2013. "Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons." *Nat Neurosci* 16(8):1068–76.

Phillips, Elizabeth AK and Andrea R. Hasenstaub. 2016. "Asymmetric effects of activating and inactivating cortical interneurons." *eLife* 5:e18383.
**URL:** *https://elifesciences.org/content/5/e18383v2*

Pi, Hyun-Jae, Balázs Hangya, Duda Kvitsiani, Joshua I Sanders, Z Josh Huang and Adam Kepecs. 2013. "Cortical interneurons that specialize in disinhibitory control." *Nature* 503(7477):521–4.

Pinto, Lucas and Yang Dan. 2015. "Cell-Type-Specific Activity in Prefrontal Cortex during Goal-Directed Behavior." *Neuron* 87(2):437–450.

Poirazi, Panayiota, Terrence Brannon and Bartlett W Mel. 2003. "Pyramidal neuron as two-layer neural network." *Neuron* 37(6):989–99.

Popescu, Gabriela, Antoine Robert, James R Howe and Anthony Auerbach. 2004. "Reaction mechanism determines NMDA receptor response to repetitive stimulation." *Nature* 430(7001):790–3.

Potjans, Tobias C and Markus Diesmann. 2014. "The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model." *Cereb Cortex* 24(3):785–806.

Potjans, Wiebke, Abigail Morrison and Markus Diesmann. 2009. "A spiking neural network model of an actor-critic learning agent." *Neural computation* 21(2):301–339.

Ragan, Timothy, Lolahon R Kadiri, Kannan Umadevi Venkataraju, Karsten Bahlmann, Jason Sutin, Julian Taranda, Ignacio Arganda-Carreras, Yongsoo Kim, H Sebastian Seung and Pavel Osten. 2012. "Serial two-photon tomography for automated ex vivo mouse brain imaging." *Nat Methods* 9(3):255–8.

Rajan, Kanaka, Christopher D Harvey and David W Tank. 2016. "Recurrent network models of sequence generation and memory." *Neuron* 90(1):128–42.

Rajan, Kanaka and LF Abbott. 2006. "Eigenvalue spectra of random matrices for neural networks." *Physical review letters* 97(18):188104.

Raposo, David, Matthew T Kaufman and Anne K Churchland. 2014. "A category-free neural population supports evolving demands during decision-making." *Nat Neurosci* 17(12):1784–92.

Ratcliff, Roger. 1978. "A theory of memory retrieval." *Psychological review* 85(2):59.

Reimer, Jacob, Emmanouil Froudarakis, Cathryn R. Cadwell, Dimitri Yatsenko, George H. Denfield and Andreas S. Tolias. 2014. "Pupil fluctuations track fast switching of cortical states during quiet wakefulness." *Neuron* 84(2):355–362.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0896627314008915*

Renart, Alfonso, Jaime De La Rocha, Peter Bartho, Liad Hollender, Néstor Parga, Alex Reyes and Kenneth D Harris. 2010. "The asynchronous state in cortical circuits." *science* 327(5965):587–590.

Reverberi, Carlo, Kai Görgen and John-Dylan Haynes. 2012. "Compositionality of rule representations in human prefrontal cortex." *Cereb Cortex* 22(6):1237–46.

Rigotti, Mattia, Daniel Ben Dayan Rubin, Xiao-Jing Wang and Stefano Fusi. 2010. "Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses." *Front Comput Neurosci* 4:24.

Rigotti, Mattia, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller and Stefano Fusi. 2013. "The importance of mixed selectivity in complex cognitive tasks." *Nature* 497(7451):585–90.

Rodgers, Chris C and Michael R DeWeese. 2014. "Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents." *Neuron* 82(5):1157–70.

Roelfsema, Pieter R and Arjen van Ooyen. 2005. "Attention-gated reinforcement learning of internal representations for classification." *Neural computation* 17(10):2176–2214.

Roitman, Jamie D and Michael N Shadlen. 2002. "Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task." *J Neurosci* 22(21):9475–89.

Romo, R, C D Brody, A Hernández and L Lemus. 1999. "Neuronal correlates of parametric working memory in the prefrontal cortex." *Nature* 399(6735):470–3.

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

Rudolph, Michael and Alain Destexhe. 2003. "A fast-conducting, stochastic integrative mode for neocortical neurons in vivo." *J Neurosci* 23(6):2466–76.

Rudy, Bernardo, Gordon Fishell, SooHyun Lee and Jens Hjerling-Leffler. 2011. "Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons." *Dev Neurobiol* 71(1):45–61.

Rumelhart, David E, Geoffrey E Hinton and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report California Univ San Diego La Jolla Inst for Cognitive Science.

Rushworth, M F S, K A Hadland, D Gaffan and R E Passingham. 2003. "The effect of cingulate cortex lesions on task switching and working memory." *J Cogn Neurosci* 15(3):338–53.

Sakai, Katsuyuki. 2008. "Task set and prefrontal cortex." *Annu Rev Neurosci* 31:219–45.

Saleem, Aman B, Asl Ayaz, Kathryn J Jeffery, Kenneth D Harris and Matteo Carandini. 2013. "Integration of visual motion and locomotion in mouse visual cortex." *Nat Neurosci* 16(12):1864–9.

Sanders, Honi, Michiel Berends, Guy Major, Mark S Goldman and John E Lisman. 2013. "NMDA and GABAB (KIR) conductances: the "perfect couple" for bistability." *J Neurosci* 33(2):424–9.

Sasaki, Ryo and Takanori Uka. 2009. "Dynamic readout of behaviorally relevant signals from area MT during task switching." *Neuron* 62(1):147–57.

Scellier, Benjamin and Yoshua Bengio. 2016. "Towards a biologically plausible backprop." *arXiv preprint arXiv:1602.05179* .

Schiller, J, G Major, H J Koester and Y Schiller. 2000. "NMDA spikes in basal dendrites of cortical pyramidal neurons." *Nature* 404(6775):285–9.

Scott, William A. 1962. "Cognitive complexity and cognitive flexibility." *Sociometry* pp. 405–414.

Seybold, Bryan A., Elizabeth A. K. Phillips, Christoph E. Schreiner and Andrea R. Hasenstaub. 2015. "Inhibitory actions unified by network integration." *Neuron* 87(6):1181–1192.
**URL:** *http://www.sciencedirect.com/science/article/pii/S0896627315007709*

Shoemaker, Patrick A. 2011. "Neural bistability and amplification mediated by NMDA receptors: Analysis of stationary equations." *Neurocomputing* 74(17):3058–3071.

Siegel, Markus, Timothy J Buschman and Earl K Miller. 2015. "Cortical information flow during flexible sensorimotor decisions." *Science* 348(6241):1352–5.

Siegelmann, Hava T and Eduardo D Sontag. 1995. "On the computational power of neural nets." *Journal of computer and system sciences* 50(1):132–150.

Silberberg, Gilad and Henry Markram. 2007. "Disynaptic inhibition between neocortical pyramidal cells mediated by Martinotti cells." *Neuron* 53(5):735–46.

Smith, Spencer L, Ikuko T Smith, Tiago Branco and Michael Häusser. 2013. "Dendritic spikes enhance stimulus selectivity in cortical neurons in vivo." *Nature* 503(7474):115–20.

Sohal, Vikaas S, Feng Zhang, Ofer Yizhar and Karl Deisseroth. 2009. "Parvalbumin neurons and gamma rhythms enhance cortical circuit performance." *Nature* 459(7247):698–702.

Song, H Francis, Guangyu R Yang and Xiao-Jing Wang. 2016. "Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible frame-

work." *PLoS Comput Biol* 12(2):e1004792.

Song, H Francis, Guangyu R Yang and Xiao-Jing Wang. 2017. "Reward-based training of recurrent neural networks for cognitive and value-based tasks." *Elife* 6:e21492.

Song, H Francis, Henry Kennedy and Xiao-Jing Wang. 2014. "Spatial embedding of structural similarity in the cerebral cortex." *Proc Natl Acad Sci U S A* 111(46):16580–5.

Song, Sen, Per Jesper Sjöström, Markus Reigl, Sacha Nelson and Dmitri B Chklovskii. 2005. "Highly nonrandom features of synaptic connectivity in local cortical circuits." *PLoS Biol* 3(3):e68.

Spiro, Rand J et al. 1988. "Cognitive Flexibility Theory: Advanced Knowledge Acquisition in Ill-Structured Domains." *Technical Report No. 441.* .

Sridharan, Devarajan and Eric I Knudsen. 2015. "Selective disinhibition: A unified neural mechanism for predictive and post hoc attentional selection." *Vision Res* 116(Pt B):194–209.

Stevenson, Ian H and Konrad P Kording. 2011. "How advances in neural recording affect data analysis." *Nat Neurosci* 14(2):139–42.

Stoet, Gijsbert and Lawrence H Snyder. 2004. "Single neurons in posterior parietal cortex of monkeys encode cognitive set." *Neuron* 42(6):1003–12.

Stroop, J Ridley. 1935. "Studies of interference in serial verbal reactions." *Journal of experimental psychology* 18(6):643.

Sussillo, David and L F Abbott. 2009. "Generating coherent patterns of activity from chaotic neural networks." *Neuron* 63(4):544–57.

Sussillo, David, Mark M Churchland, Matthew T Kaufman and Krishna V Shenoy. 2015. "A neural network that finds a naturalistic solution for the production of muscle activity." *Nat Neurosci* 18(7):1025–33.

Sussillo, David and Omri Barak. 2013. "Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks." *Neural Comput* 25(3):626–49.

Sutskever, Ilya, James Martens, George Dahl and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*. pp. 1139–1147.

Sutton, Richard S and Andrew G Barto. 1998. *Reinforcement learning: An introduction*.

Vol. 1 MIT press Cambridge.

Taniguchi, Hiroki, Miao He, Priscilla Wu, Sangyong Kim, Raehum Paik, Ken Sugino, Duda Kvitsiani, Duda Kvitsani, Yu Fu, Jiangteng Lu, Ying Lin, Goichi Miyoshi, Yasuyuki Shima, Gord Fishell, Sacha B Nelson and Z Josh Huang. 2011. "A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex." *Neuron* 71(6):995–1013.

Thomson, Alex M, David C West, Yun Wang and A Peter Bannister. 2002. "Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2–5 of adult rat and cat neocortex: triple intracellular recordings and biocytin labelling in vitro." *Cerebral cortex* 12(9):936–953.

Thurstone, Louis Leon. 1938. "Primary mental abilities.".

Tschentscher, Nadja, Daniel Mitchell and John Duncan. 2017. "Fluid intelligence predicts novel rule implementation in a distributed frontoparietal control network." *J Neurosci* 37(18):4841–4847.

Tsodyks, Misha V., William E. Skaggs, Terrence J. Sejnowski and Bruce L. McNaughton. 1997. "Paradoxical effects of external modulation of inhibitory interneurons." *Journal of Neuroscience* 17(11):4382–4388.
**URL:** *http://www.jneurosci.org/content/17/11/4382*

Urban-Ciecko, Joanna and Alison L Barth. 2016. "Somatostatin-expressing neurons in cortical networks." *Nat Rev Neurosci* 17(7):401–9.

Urban-Ciecko, Joanna, Erika E Fanselow and Alison L Barth. 2015. "Neocortical somatostatin neurons reversibly silence excitatory transmission via GABAb receptors." *Curr Biol* 25(6):722–31.

Vogels, T P, H Sprekeler, F Zenke, C Clopath and W Gerstner. 2011. "Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks." *Science* 334(6062):1569–73.

Vogels, Tim P and L F Abbott. 2009. "Gating multiple signals through detailed balance of excitation and inhibition in spiking networks." *Nat Neurosci* 12(4):483–91.

Wall, Nicholas R, Mauricio De La Parra, Jordan M Sorokin, Hiroki Taniguchi, Z Josh Huang and Edward M Callaway. 2016. "Brain-Wide Maps of Synaptic Input to Cortical Interneurons." *J Neurosci* 36(14):4000–9.

Wallis, J D, K C Anderson and E K Miller. 2001. "Single neurons in prefrontal cortex encode abstract rules." *Nature* 411(6840):953–6.

Wang, X J. 1999. “Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory.” *J Neurosci* 19(21):9587–603.

Wang, X-J, J Tegnér, C Constantinidis and P S Goldman-Rakic. 2004. “Division of labor among distinct subtypes of inhibitory neurons in a cortical microcircuit of working memory.” *Proc Natl Acad Sci U S A* 101(5):1368–73.

Wang, Xiao-Jing. 2002. “Probabilistic decision making by slow reverberation in cortical circuits.” *Neuron* 36(5):955–68.

Wang, Xiao-Jing. 2013. The prefrontal cortex as a quintessential “cognitive-type” neural circuit. In *Principles of Frontal Lobe Function*, ed. D. T. Stuss and R. T. Knight. Second edition ed. New York: Cambridge University Press pp. 226–248.

Weigl, Egon. 1927. “Zur Psychologie sogenannter Abstraktionsprozesse. I. Untersuchungen über das" Ordnen".” *Zeitschrift für Psychologie und Physiologie der Sinnesorgane. Abt. 1. Zeitschrift für Psychologie* .

Welsh, Marilyn C, Bruce F Pennington and Dena B Groisser. 1991. “A normative-developmental study of executive function: A window on prefrontal function in children.” *Developmental neuropsychology* 7(2):131–149.

Wimmer, Ralf D, L Ian Schmitt, Thomas J Davidson, Miho Nakajima, Karl Deisseroth and Michael M Halassa. 2015. “Thalamic control of sensory selection in divided attention.” *Nature* 526(7575):705–9.

Wong, Kong-Fatt and Xiao-Jing Wang. 2006. “A recurrent network mechanism of time integration in perceptual decisions.” *J Neurosci* 26(4):1314–28.

Xue, Mingshan, Bassam V Atallah and Massimo Scanziani. 2014. “Equalizing excitation-inhibition ratios across visual cortical neurons.” *Nature* 511(7511):596–600.

Yamins, Daniel L K, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert and James J DiCarlo. 2014. “Performance-optimized hierarchical models predict neural responses in higher visual cortex.” *Proc Natl Acad Sci U S A* 111(23):8619–24.

Yang, Guang, Cora Sau Wan Lai, Joseph Cichon, Lei Ma, Wei Li and Wen-Biao Gan. 2014. “Sleep promotes branch-specific formation of dendritic spines after learning.” *Science* 344(6188):1173–8.

Yang, Guangyu Robert, John D. Murray and Xiao-Jing Wang. 2016. “A dendritic disinhibitory circuit mechanism for pathway-specific gating.” *Nature Communications* 7:12815.
**URL:** *http://www.nature.com/ncomms/2016/160920/ncomms12815/full/ncomms12815.html*

Yuste, Rafael. 2015. “From the neuron doctrine to neural networks.” *Nat Rev Neurosci* 16(8):487–97.

Zenke, Friedemann, Ben Poole and Surya Ganguli. 2017. “Improved multitask learning through synaptic intelligence.” *arXiv preprint arXiv:1703.04200* .

Zhang, Siyu, Min Xu, Tsukasa Kamigaki, Johnny Phong Hoang Do, Wei-Cheng Chang, Sean Jenvay, Kazunari Miyamichi, Liqun Luo and Yang Dan. 2014. “Long-range and local circuits for top-down modulation of visual cortex processing.” *Science* 345(6197):660–5.

Zingg, Brian, Houri Hintiryan, Lin Gou, Monica Y Song, Maxwell Bay, Michael S Bienkowski, Nicholas N Foster, Seita Yamashita, Ian Bowman, Arthur W Toga and Hong-Wei Dong. 2014. “Neural networks of the mouse neocortex.” *Cell* 156(5):1096–111.

Zipser, D and R A Andersen. 1988. “A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons.” *Nature* 331(6158):679–84.

Zylberberg, Ariel, Diego Fernández Slezak, Pieter R Roelfsema, Stanislas Dehaene and Mariano Sigman. 2010. “The brain’s router: a cortical network model of serial processing in the primate brain.” *PLoS Comput Biol* 6(4):e1000765.