Relating changes in cortical state to circuit structure and dynamics

by

Matthew P. Getz

B.S., B.A., University of Florida, 2012

M.S., City College of New York, 2016

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Matthew P. Getz

It was defended on

November 30th 2022

and approved by

Marlene Cohen, Ph.D, Professor, Neuroscience

Bard Ermentrout, Ph.D, Professor, Mathematics

Matthew Smith, Ph.D, Associate Professor, Biomedical Engineering, Carnegie Mellon University

John Maunsell, Ph.D, Professor, Neurobiology, University of Chicago

Committee Co-Chair: Caroline Runyan, Ph.D, Assistant Professor, Neuroscience

Dissertation Advisor & Committee Co-Chair: Brent Doiron, Ph.D, Professor, Mathematics

Relating changes in cortical state to circuit structure and dynamics

Matthew P. Getz, PhD

University of Pittsburgh, 2022

Variability in neural activity is often tied to cognitive or behavioral substrates, yet in linking neural dynamics to behavior, most theoretical work has ignored changing cortical state. In this dissertation I will present two pieces of work which seek to explicitly relate cortical state changes to circuit structure and dynamics. We find the role of inhibitory interneurons appears to be a unifying theme in the interaction between cognitive variables and neural dynamics.

In the first part we ask what circuit properties underlie how cortical state affects information flow through a neural network. We find that for a linear decoder's performance to change as a function of state, it must be restricted to a subset of the population. Curiously, the decoder's performance change is shaped not by the population of cells being decoded but rather the collection of cells which project to the decoded population. This result has an interesting implication: understanding information flow through cortical circuits may rest on understanding inhibitory interneuron response properties.

In the second part I will turn to the correlation between normalization and attention and argue that, despite being a conspicuous relationship between a cognitive variable (attention) and a circuit-dynamic variable (normalization), it nevertheless is insufficient to adequately constrain circuit models. We instead find other correlated heterogeneities better constrain mechanistic models of attention and in particular point to the necessity of strongly recurrent networks in constructing these relationships. We then demonstrate network properties which support this collection of correlated heterogeneities showing how these depend on the structure of inhibition.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation reflects a snapshot in time of work that has been ongoing for many years. While this document has my name at the top, it is in fact the product of many generous individuals supporting both the work and myself. Any successes in this document are shared with many others; mistakes and failures are my own.

I am deeply grateful to have had such a wonderful committee to shepherd me through school. Foremost to my advisor Brent Doiron for taking me on as a student. I credit most of what I know about how theoretical neuroscience works, from the actual grind to general knowledge, to him, and appreciate his providing a [highly unfiltered] lens to see academia through; to Marlene Cohen and Matt Smith, for being immensely supportive and insightful advisors in their own rights, regularly offering advice, encouragement, and unique perspectives; to Bard Ermentrout, for teaching me applied math and for being a truly singular source of inspiration and an endless source of entertainment, sometimes constructive; to Caroline Runyan, for her immediate and enthusiastic encouragement; and to John Maunsell, for accepting me as a community member and being incredibly attentive to my requests despite having [as far as I can tell] no skin in the game save his own curiosity.

I am grateful to the many members of the Doiron lab who offered support and camaraderie throughout my time in Pittsburgh and Chicago. Despite his curmudgeonly ways Jeff Dunworth helped me immensely in figuring out how to be a lab member, and in debugging code and math. Chengcheng Huang has been a constant fixture throughout my time in grad school. I learned a lot from her, and enjoyed all the time we've spent together. Hannah Bos, Mike Leone, and Jay Pina were all instrumental in my development, as well as good friends outside the lab. Danielle Rager was a wonderful counterpoint in all aspects, and I enjoyed our interactions crossing from programming to films. Tevin Rouse was a great intellectual companion, especially during the lockdowns, and kept me on my toes with his persistent inquisitiveness. Following the lab's move to Chicago, Olivia Gozel and Gregory Handy became my closest colleagues. I've enjoyed many discussions and outings with them, and they have assisted me with many thankless tasks, and for that, I thank them. I am

fortunate to have learned how to do better science from all of these individuals, but also to have spent much time out of lab enjoying other activities together and in so doing, learning what makes a truly wonderful scientific community.

I am grateful to the communities both in Pittsburgh and Chicago who accepted and aided me through multiple unique experiences. I am particularly thankful for Katy Friason and Pati Stan for their outstanding friendship. Doug Ruff's encyclopedic knowledge helped me get off the ground in visual neuroscience, and his unflagging support helped me grow significantly.

I am incredibly thankful to the many individuals outside my academic circles who helped in this journey, most importantly Elaine Wilson and Mirna Turina.

Despite not fully knowing what I'd gotten myself into, my parents Ken and Katy were unquestioning and unlimited in their assistance of me in this endeavor. Together with my brother Michael, and grandmothers Linda and Jean, all have offered me a vast amount of support in their own ways throughout this process. I cannot say if I'd have made it to this point without them.

I am fortunate to say that I owe a debt of gratitude to many others whose names would be too numerous to list. I hope I don't fail to insure they know.

1.0 Introduction

The notion of cognition embodies the often complex operations performed by the brain which endow an organism with rich functional capabilities. It encapsulates the familiar sense of thinking in terms of decision-making, planning, remembering, perceiving; but also taking action - moving. In this way cognition and behavior are bound in a continual loop. The importance of this relationship for understanding activity in the brain was evinced by Simon and Kaplan in the early 90's: "Cognitive science, defined as the study of intelligence and its computational processes, can be approached ... [by studying] human (or animal) intelligence, seeking to abstract a theory of intelligent processes from the behavior of intelligent organisms" [134]. On the other hand, uncovering the "computational processes" underlying cognition (or "intelligence") depends on understanding how neural activity is structured. As put by Sejnowski and Churchland, "Once it became evident that the operations of the brain were essential for thoughts and actions, discovering the biological basis for mental functions was an abiding objective" [130].

In general relating mental functions to biological processes is a hard problem. There are myriad neuron species [85, 143, 39], cortical and subcortical regions [40, 52] and layers to consider, with neural circuits themselves constantly undergoing changes in their activity and connectivity across temporal and spatial scales through learning. One way to progress is to recognize that there are different levels of organization: observing a single neuron's activity will not explain the ability of someone riding a bicycle. Instead, the activity of groups of neurons needs to be considered. But is that enough? The problem now becomes how to build relations across levels. In bridging from neurons and circuits to behavior, this thesis focuses on what might be construed as an important intermediate level: the state of a circuit.

A working definition of a circuit's state (and that which we use with more formality in Chapter 2) is its operating point. Importantly, changes in the operating point of a circuit are intimately related both to an organism's actions and a circuit's activity. Many of these changes are due to the state of the animal [70], or controlled through cognitive processes. The importance of incorporating state context into uncovering structure-function relation-

ships is illustrated by lower animals: the complete circuit diagram of both the nematode *Caenorhabditis elegans* and the stomatogastric ganglion of the crab are known, yet it is still unclear how exactly activity in these circuits maps to the organism's behavior due in part to the large variety of ways in which the state of the circuits can be modified [87, 75]. Therefore a critical step in aligning neural activity with behavior lies in developing a deeper understanding of the way in which state affects neural activity. In the context of much theoretical work, this explicit step has been largely absent. The content of this thesis articulates two studies which have sought to make progress in this regard.

1.1 Cortical states of arousal and attention

Activation of sensory receptors elicits activity in neural impulses that are carried from an organism's periphery to its central nervous system (excluding animals with nerve nets or similar). The classical view of cortical neural responsiveness thus derives from the constellation of such stimulus-driven impulses to which a cortical neuron responds. However, it is well known that signals endogenous to the brain can affect neural activity as well, independently of any stimulus drive. These may arise through feedback signals from higher-order cortical areas to lower ones [155, 72], or from midbrain and brainstem neuromodulatory structures like the locus coeruleus and basal forebrain, with diffuse projections throughout cortex [58]. Among the many effects of feedback drive and neuromodulation are changes in synaptic efficacy, neural baseline firing rates, covariability in pairwise neural activity, and cellular response gain [57, 41] (mechanisms of state modulation are discussed in section 1.2).

Early observations of cortical activity uncovered two clearly differentiable states: sleep, or low arousal, in which low frequency oscillatory activity dominates throughout cortex; and the awake state, characterized by high-frequency asynchronous activity [111]. In general, cortical state is defined along a continuum between these two extremes [57]. We now focus on two particular mechanisms of state modulation, arousal and attention. Each of these is well-studied in the electrophysiological and neuroimaging literature, with clearly defined behavioral and physiological correlates which make linking neural activity to behavior possible

in both situations.

1.1.1 Arousal

Arousal can be thought of as a general sense of alertness. Particularly in the context of medically-induced anesthesia, the oscillatory frequency of low-pass filtered signals across cortex has been used to quantify arousal states [17]. Given the highly desynchronized nature of activity during active wakefulness, however, a different measure is needed to capture global arousal states. Experiments in awake, behaving, animals have identified pupil diameter as a reliable measure of an animal's alertness [93]. Pupil diameter affords an excellent metric to capture arousal state because it is easily measured throughout the experiment, is continuous in time and can thus be related to temporally correlated neural activity. In this way, experiments have shown that arousal state is related to an animal's ability to detect a change in an auditory tone sequence in a non-monotonic fashion [92]. This is likely familiar to the reader: if one is barely awake on the one hand or manic on the other, it is hard to detect subtle changes in one's environment, whereas a moderate level of arousal state is most consistent with nuanced engagement with the external world. Having experimental access to a mouse's arousal state enabled a suite of interesting discoveries at the neuronal level. It was shown that inhibitory interneurons in visual cortex were a particular target of neuromodulatory action as a function of arousal [102], leading to disinhibition of excitatory pyramidal cells which resulted in gain modulation of those excitatory cells [45]. These experiments were thereby able to establish a link between the cortical state of an animal, its behavior, and its coincident neural activity.

1.1.2 Attention

In contrast to arousal which induces a global state change across cortex, attention can affect circuits in a localized fashion. Perceptually, attention acts to select relevant stimuli in either a top-down - endogenous - fashion in which an organism chooses to what or where to deploy attentional resources, or in a bottom-up - exogenous - fashion, also referred to as attentional capture [21]. Attentional capture is marked by surprising selection-dominant

stimuli which draw one's focus to an object or area. As an experimental paradigm, attention has been critical in bridging across cortical state, behavior, and neural computation. Non-human primate studies in which monkeys must perform a visual change-detection task have shown that when a monkey's attention is cued to a stimulus more likely to change on a given trial, the monkey's ability to discriminate changes in the angle of an oriented grating increases [24]. Furthermore, electrophysiological recordings have revealed that on average, the gain and firing rates of neurons encoding the relevant stimulus increase with attention [144, 91], and shared variability between neurons within a cortical region decrease [24, 120]. On the other hand, correlated variability has been shown to increase across hierarchically distinct, connected cortical regions [121], suggesting that attention facilitates communication between brain areas as well [74].

Endogenous covert (that is, without eye movement) attention can be deployed to spatial locations or stimulus features (or both simultaneously). We will largely concern ourselves with the former, but there are important distinctions between the two which merit mentioning. Attending to a location in visual space affects the activity of cortical neurons whose receptive fields represent that location. On the other hand, feature attention affects neurons tuned to the relevant stimulus feature, across hemispheres [25]. Additionally, while spatial and feature attention have been shown to scale neural tuning functions in a contrast-invariant way [91, 144], there is evidence that feature attention may sharpen population response curves [89].

The scaling of neural responses with attention has generally been attributed to a change in response gain. The nature of this gain change has been debated insofar as its effect on a neuron's contrast response function (CRF). Some evidence supported a leftward translation along the contrast axis of the CRF, indicative of a change in neural sensitivity [116]. Other reports pointed to a multiplicative scaling of the CRF (response gain), possibly with a vertical translation in the CRF as well (activity gain). Naturally, as with much of biology, the answer seems to be a bit of everything [152], however response gain appears to be the more consistent, robust effect [90]. A consistent descriptor across these conditions, however, is that attention acts like a contrast controller since neural responses under attention appear similar to changing stimulus contrast in a neuron- and stimulus-specific fashion (Figure

1a). Emphasis should be placed on the phrase *acts like* because attention is not a contrast amplifier in a physiological sense; experimental evidence does not support this claim [78].

An apparent limitation in the gain control framework of attention arises from studies in which multiple stimuli were presented within a recorded neuron's receptive field (RF). Given the large RFs of higher order visual neurons, it is possible to fit spatially distinct stimuli within one, and to differentially cue attention to different RF locations (Figure 1b). In general, cueing attention to one stimulus in the RF reduces the effect of the second stimulus on the rate response. This led to the proposed *biased competition* model in which stimuli within a RF are effectively competing for a cell's responsiveness [114]. While this model falls short of a complete characterization, its emphasis on the mutually antagonistic effects of multiple stimuli in a neuron's RF is an important one [90]. A better model depends on the principle of *normalization* which we review in Section 1.3 and link to attention in Chapter 3.

Many anatomical regions have been identified which contribute to visual attention. Neuroimaging evidence, together with stimulation experiments in primate frontal cortices, supports a top-down view of attentional control in which the source of attentional effects lies in fronto-parietal regions such as prefrontal cortex. Subcortical regions like the superior colliculus and pulvinar of the thalamus have also been implicated in attentional control. All of these anatomical areas (frontal, parietal, visual, thalamic) have been associated to constitute a distributed "attention network" [43].

While it does not feature in this thesis, it bears noting that a body of literature has identified a prominent role for oscillatory activity in attentional processes, hypothesizing that rhythms in the theta frequency band alternately lock attention on (high activity phase) or enable it to switch (low activity phase) [42]. There is at least cursory relevance in this observation: we have already discussed the characterization of cortical state in terms of the frequency and scale of correlated activity [57]. Neuromodulation is additionally implicated in the modification of LFP oscillatory activity in cortex [142]. It would seem an appropriate future direction to ask to what extent these ideas can be brought under the framework described in this work.

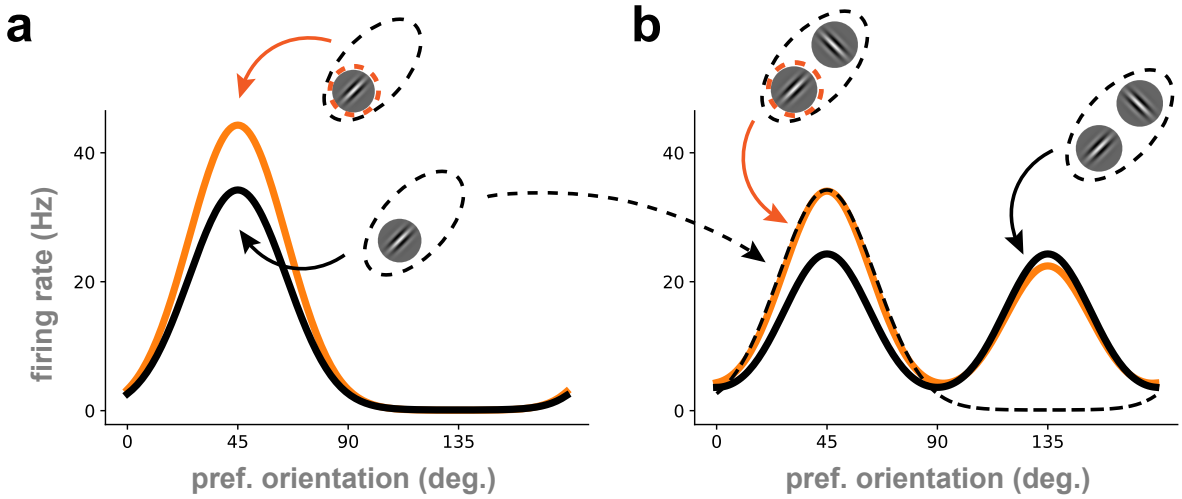

Figure 1: *Network activity in the context of attention and normalization* **a** Illustration of spatial attention effects from a network model (details in Chapter 3). Black dashed ovals represent a neural receptive field (RF). Inset gabors indicate the driving stimulus; surrounding dashed orange circles indicate location of attentional focus. Solid black curve is network response in absence of attention, solid orange curve is attended response. **b** Illustration of attention in the context of two stimuli within the RF with conventions as in (a). Dashed black curve is 45^o stimulus alone (equivalent to solid black curve in (a)).

1.2 Control of cortical state

The use of anesthetics in humans and animals is an example of cortical state control [17]. Analogous to the discovery and study of neural tuning properties, the use of anesthesia has enabled researchers to consider the differences between anesthetized and awake brain states on circuit function [36]. Of course, anesthesia is a largely unnatural perturbation of brain state. As we argued above, understanding the function of neural circuits depends on relating natural behaviors to cellular processes [76]. Accordingly, recent work in rodents has focused on uncovering endogenous state controllers in the context of awake behavior [11]. Two primary forms of state control are neuromodulation and cortical feedback. We give a brief description of each, which motivates the work in the following chapters.

1.2.1 Neuromodulatory processes

The simplistic, stimulus-driven feedforward view of neural responsiveness considered in the previous section involves only one form of direct, local, synaptic transmission through neurotransmitters like glutamate and γ-aminobutyric acid (GABA) which act on a fast, millisecond timescale through ionotropic receptors on the postsynaptic cell. However, neuromodulators like serotonin (5-hydroxytryptamine; 5-HT), acetylcholine (ACh), and norepinephrine (NE) can have a distributed effect on neurons within a given cortical region [87, 142]. Neuromodulation can alter a wide suite of biophysical properties of neurons, including synaptic efficacies [141]. Neuromodulatory effects are also slower, mediated (with some exceptions, such as ACh's action on ionotropic nicotinic receptors [29]) through metabotropic receptors. Depending on the task in which an animal is engaged, this separation of timescales implies that the effects of neuromodulation can be considered constant over a timeframe in which stimuli are inducing rapid changes in neural activity.

Neuromodulation is also generally diffuse, with targets throughout all of cortex, in contrast to the highly localized nature of synaptic transmission. Hence, in addition to being constant in time, for local circuits, neuromodulatory effects can be assumed constant over space. Brainstem and midbrain nuclei serve as the major sources of neuromodulatory input

to cortex. For example, the locus coeruleus in the brainstem is a major source of noradrenergic outputs, while the basal forebrain is a primary source of cholinergic projections [17]. Both may play a role in affecting visual processing in rodents during running. Noradrenaline in visual cortex has been shown to enhance response gain and reduce firing rate variability, consistent with effects observed during locomotion [106]. Yet another recent study implicated cholinergic inputs from the basal forebrain in modulating gain in visual cortex during locomotion [45].

1.2.2 Cortical feedback

In further contrast to the feedforward processing view of cortex, it is well known that there exist long-range, interregional feedback projections throughout cortex. While far less-characterized, these feedback signals have nevertheless been shown to affect processing at early stages of sensory systems [139, 140]. Given that feedback activity is synaptic in nature, it can be much more spatially targeted. Two recent studies illustrate both the power and specificity of feedback connections. Zagha *et al.* [155] explored the role of primary motor (M1) cortex in modulating activity in primary somatosensory (S1) cortex. They showed that M1 stimulation caused activation of S1 neurons in a layer-dependent fashion and independent of thalamus, implicating direct feedback projections in the process. Further, they showed that representations of complex whisker manipulations were more accurately decoded from S1 in the presence of M1 activation, demonstrating the ability of feedback activity to affect processing. In rodent primary visual (V1) cortex, Keller *et al.* [72] uncovered the existence of what they term a feedback receptive field (fbRF) through the use of inverse images. These images were full-screen with portions excised. Excitatory neurons in V1 layer 2/3 showed a response field that was larger than the classical feedforward RF. Importantly, it was shown that the fbRF of a given V1 cell arises from higher order visual areas with RFs that are offset relative to the cell in V1. In this way, neural activity is modulated by the context in which a stimulus is embedded. Again, we see that feedback projections strongly affect neural responsiveness by controlling the state of cortical circuits in a more localized fashion.

What this very brief description of two different modulatory processes in the brain -

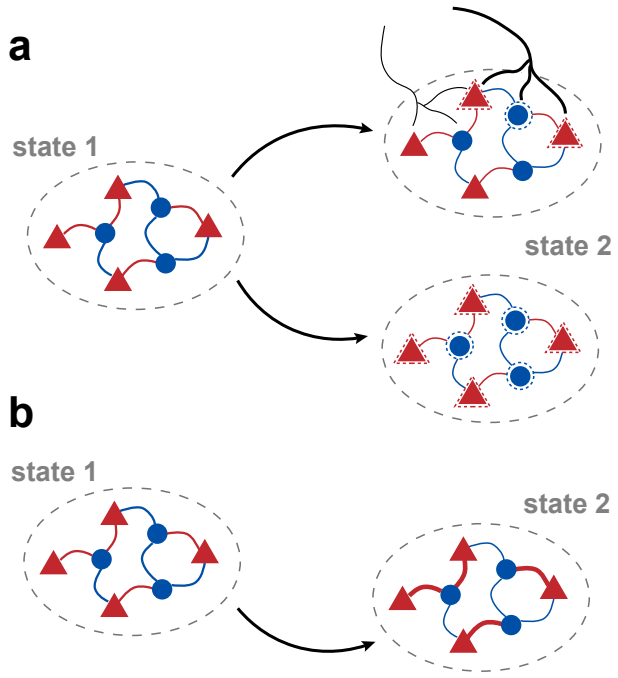

Figure 2: *Modeling cortical state controllers* **a** Illustration of possible modeled gain control in a recurrent circuit: localized through feedback projections (top right, dashed outlines) or through diffuse neuromodulation (bottom right, dashed outlines). **b** Illustration of altered synaptic efficacy through e.g., neuromodulatory mechanisms.

neuromodulation and corticortical feedback - illustrates is just how ubiquitous they are. Yet, their effect on neural circuits has been largely left out of models which look to understand neural representations, encoding, and decoding process (section 1.4). Interpreting these models has nevertheless relied on animal studies in which behavioral and brain states are changing, suggesting that drawing mechanistic conclusions from these models may be limited without considering what processes in the brain are different across states. This section also highlights how we can incorporate the effects of modulatory processes on circuits: we can first assume they operate on a sufficiently slow timescale that from one state to the next the effect of a modulus may be justifiably assumed constant. The exact implementation of the modulus will depend on the nature of the modeled process (Figure 2). Changes in cellular gain can arise from both feedback projections and neuromodulation in either a spatially diffuse or localized fashion (Figure 2a) whereas synaptic weights can also change through neuromodulatory processes (Figure 2b).

1.3 Normalization as a nonclassical neural response property

Determining how cortical state affects neural activity depends on an understanding of neural response properties generally. Early experimental descriptions of neural responsiveness were often performed in anesthetized animals, thereby (perhaps inadvertently) controlling for cortical state [64]. Foundational descriptions of neural tuning to stimulus features [12] and responsiveness as a function of stimulus contrast were then established [129]. Many of these basic properties could be captured reasonably well with a simple linear summation model in which thalamic inputs combine to determine cortical response properties [64, 98]. However it soon became apparent that this basic conceptualization couldn't capture the full repertoire of cortical neuron responses. A nonlinear, multi-stage model was introduced which incorporated divisive normalization to the linear summation step [59]. That is, after passing an input term through a rectified power function (in which the power typically equalled two), the response was then divided by the sum of a constant term and the sum of other model unit outputs (e.g., equation 37). This normalization stage readily explained many of the

peculiarities of stimulus-conditioned responses.

Normalization encapsulates the broad concept that neural activity is determined by the net activity across a distributed neural population. This involvement of local population activity explains contrast response saturation, surround suppression, and cross orientation suppression, among other phenomena, in single neurons. Classically, neurons respond to stimuli presented within their receptive fields (RFs) and are silent to isolated stimuli beyond their RF. Surround suppression is the effect when stimuli larger than the RF induce a reduction in firing rates. This effect is tuning-dependent; an annulus in the surround will exert a maximal effect when aligned to the neuron's preferred orientation, and minimally when aligned to the neuron's null orientation [22]. Another tuning-dependent property is cross-orientation suppression. Cross-orientation suppression is the effect whereby the presence of a secondary non-optimal stimulus in a cell's RF often causes a reduction in the response relative to optimal stimulus presented alone (Figure 1b, compare dashed black line (single stimulus) and solid black line (two stimuli) at 45^o. Cross-orientation suppression is discussed in more detail in section 3.2) [119]. Contrast response saturation has been shown both theoretically [105] and experimentally [135] to depend on recurrent inputs. Further, attention's ability to amplify the saturation point of a neuron's response is further evidence that it is not a biophysical constraint but rather a network property. Hence, what normalization really brings to the fore and makes explicit in terms of a descriptive model (section 3.2) is the dependence of neural response properties on the local network structure in addition to feedforward inputs.

1.4 Neural encoding and decoding

In isolation, neurons are very reliable: current drive which is not so strong as to induce adaptation or depolarization block will trigger spikes whenever threshold is reached. However the source of this current drive is important - synapses are highly malleable, and may vary in the extent to which they affect postsynaptic voltage changes. Embedded in a network, *in vivo*, neural activity becomes highly variable [154]. Neurons communicate through spikes,

but the *rate* at which spikes are emitted by a cell is recorded as a key variable of information encoded by that cell. Typically, this variable is inferred through repeated presentation of a salient stimulus during recording of a neuron's activity, and subsequent averaging of this activity over stimulus presentations. In the context of a stimulus parameterized by a continuous variable - for example, the orientation of a bar in a two-dimensional plane - the firing rate of a neuron sensitive (or: tuned) to that stimulus can exhibit continuous deviation proportional to the stimulus parameter [64]. This has been taken as evidence of a neuron's encoded representation. What is clear in the modern view is that what we can at best say is that the neuron's *expected* activity encodes a particular variable. When a neuron's activity is observed for each presented stimulus, what is in fact observed is a probability distribution over firing rates as a function of the stimulus (see section **??**). A cell's encoding of a particular stimulus variable must have practical relevance to an organism. Given the hierarchical structure of cortical connectivity, this relevance can be envisioned as a cell's representation being decodable by a downstream region. But given the variability in activity from stimulus presentation to presentation (or, trial to trial), how does a downstream area cope with this stochasticity?

One solution is redundancy - have more than one neuron carry the same information and average across *neurons* [48, 47]. Another is to encode variables in a distributed fashion, where neural activity spans a lower dimensional subspace. Both of these solutions depend on populations of neurons to code stimulus variables, not neurons in isolation. Further, cortical circuits are highly recurrent, that is, interconnected within and across cortical layers. The idea of a neuron in isolation coding for a particular component of the external world might only make sense if the collection of inputs to that neuron are weak and considered statistically independent, which is known not to be the case [111]. Population coding depends on understanding interactions between cells, the nature of their interconnections, and, by extension, the structure of shared variability [145].

In seeking to link neural dynamics to behavioral actions, neuroscientists have long used statistical estimation tools to approximate the extent to which information can be read out from a population of cells thereby limiting (or facilitating) behavioral performance [103, 132] (Appendix A.3). This derives in part from psychophysical experiments which define behav-

ioral performance by an animal's ability to discriminate two subtly different stimuli. One can consider an observation of a stimulus as a sample from some underlying probability distribution conditioned on the stimulus s. To measure an observer's ability to differentiate between these distributions, a commonly used metric of discriminability is d', defined as the difference in the estimated means of the distributions divided by the standard deviation (assuming the variances are the same; see Appendix A.3) [54, 136]. Intuitively, the activity of neural circuits should relate to the capacity of an organism to perform and action accurately, albeit to some level of abstraction. In this vein d' has been applied to neural population responses as well [136, 24]. d' is in turn related to another measure frequently used in the analysis of neural population coding, Fisher information (FI). FI has proven a highly useful tool for studying the properties of neural tuning that affect discriminability [132], and how noise in population activity affects coding [65] (see next section). FI also determines the Cramér-Rao bound, which places a limit on the accuracy of unbiased decoders (Appendix A.3). This fact enables a natural interpretation of FI in terms of "readout" by higher cortical areas, in a feedforward processing view of the brain. Downstream neurons attempting to decode a signal carried by their synaptic inputs will be more accurate if the feedforward connections have a higher FI. Consistently, various studies have developed biologically-motivated constructs to implement decoding schemes which can approach these bounds [108, 32]. Intriguingly, experimental studies have shown that neurons often encode stimulus representations with either the same [16] or a higher fidelity than an organism is able to functionally report [137]. This suggests that other factors in an organism's central nervous system conspire to bound information flow. In order to address a potential explanation, we must first discuss the role of correlated variability in affecting neural codes.

1.4.1 The effect of noise correlations on neural codes

As introduced above, despite a neuron's reliability under controlled inputs, neural responses are highly variable *in vivo*, even to repeated presentations of the exact same stimulus. From a neural coding perspective, if this variability is sufficiently small or uncorrelated across cells it can be averaged out over a large enough population. However, if fluctuations

are correlated across cells, they cannot be averaged away and will result in a degradation of information [156]. To illustrate this point, consider the signal to noise ratio of a population of N statistically identical cells with mean signal μ and variance σ^2 [156]. Then the population signal to noise ratio (SNR), defined by the ratio of the mean response to the standard deviation, is given by

$$\text{SNR} = \frac{N\mu}{\sigma\sqrt{N + cN(N-1)}} \tag{1}$$

where $c \in [0, 1]$ is the correlation in the signals. If $c = 0$ then the population is completely uncorrelated and the SNR grows with increasing population size like $\sqrt{N}$ times the SNR of an individual unit. However if $c > 0$ then the SNR saturates since

$$\text{SNR} = \frac{N\mu}{\sigma\sqrt{N + cN(N-1)}} \xrightarrow{N\to\infty} \frac{\mu}{\sigma\sqrt{c}}. \tag{2}$$

While overly simplistic, this example captures an important point which has occupied the field for decades: what is the effect of correlated variability on population codes?

Abbott and Dayan [1] later showed that in a population of neurons with translation-invariant tuning curves, FI would increase with both the size of the population and with certain structures of covariability; only with limited range correlations did they find that correlations would hamper FI. Even so, for a sufficiently large population the length constant of the correlations would need to increase as well to limit information. While this particular solution would need to be finely tuned, it pushed forward the search for how the *structure* of correlated variability impacted population codes.

Recent work has identified the existence of general coding bounds [97, 69]. Termed *differential correlations* [97], they are correlations whose structure matches the derivative of the stimulus tuning function. The trouble with this correlation function is that it induces coordinated shifts in the representation of the stimulus by the population. Since the variability is thus aligned to the coding direction, it cannot be averaged out and therefore bounds the information readout. While the existence of these correlations is currently a topic of debate [123, 67, 137], as we discuss in Chapter 2, modulation of information flow with cortical state only makes sense if this bound is not saturated in an unmodulated context. This prediction only arises from the explicit consideration of brain state in studying information flow.

1.5 *E/I* networks

We will use standard excitatory-inhibitory (E/I) rate networks to address the questions in this thesis, the general form of which is given by:

$$\tau_E \frac{dr_E(t)}{dt} = -r_E(t) + f_E\left(\sum_E J_{EE} r_E(t) + \sum_I J_{EI} r_I(t) + s_E(t)\right) \tag{3}$$

$$\tau_I \frac{dr_I(t)}{dt} = -r_I(t) + f_I\left(\sum_E J_{IE} r_E(t) + \sum_I J_{II} r_I(t) + s_I(t)\right). \tag{4}$$

This class of model characterizes neural activity in terms of firing rates r_E, r_I as a function f of recurrent inputs J and feedforward inputs s. Since neural activity is encoded in discrete spike events, this model is necessarily both an assumption and an abstraction. Nevertheless, one can think of neurons as encoding a stimulus variable $s(t)$ in their firing rate, and a firing rate in turn as representing the probability of a spike occurring in some small time window. In this way, the average activity over a population of cells represents information about the stimulus [48, 47]. A model of neural firing rates is hence appropriate to consider neural encoding and decoding questions which do not take fine temporal codes on the order of spike times into account. Additionally, the electrophysiological studies which motivate much of our work analyze activity in terms of spike counts over disjoint temporal windows or trial-average activity of a particular neuron (again, over spike counts) [24, 101, 150]. In this case as well, rate models are suitable to describe the relevant dynamics captured by these datasets (a more substantial review of their derivation is included in Appendix A.1).

1.6 Outline

The evidence we have presented shows that the brain does not operate in a static regime [28]. In order to understand the neural bases underlying behavior it is necessary to take cortical state into account. However, most previous modeling studies have not, and that is where this work comes in. Chapter 2 considers the fact that on constrained laboratory tasks, an animal's performance changes as a function of cortical state. Using Fisher information

(FI), an information theoretic measure, as a proxy for performance, we formally show that one may derive conditions under which FI changes as a function of cortical state. From this, we are able to then anticipate which neural circuit mechanisms underlie these changes. In Chapter 3 we sought to constrain mechanisms of a well-studied state modulator: attention. We use the fact that attentional affects on a cell are correlated with normalization effects on a cell. Within this context, we formulate a simple model description to determine which experimental observations afford the best model constraints. We then use these constraints to argue that the relationship between normalization and attention depends on the circuit structure in which a neuron is embedded.

2.0 Subpopulation Codes Permit Information Modulation Across Cortical States

This chapter has been published as a preprint [50].

2.1 Overview

Cortical state is modulated by myriad cognitive and physiological mechanisms. Yet it is still unclear how changes in cortical state relate to changes in neuronal processing. Previous studies have reported state dependent changes in response gain or population-wide shared variability, motivated by the fact that both are important determinants of the performance of any population code. However, if the state-conditioned cortical regime is well-captured by a linear input-output response (as is often the case), then the linear Fisher information (FI) about a stimulus available to a decoder is invariant to state changes. In this study we show that by contrast, when one restricts a decoder to a subset of a cortical population, information within the subpopulation can increase through a modulation of cortical state. A clear example of such a subpopulation code is one in which decoders only receive projections from excitatory cells in a recurrent excitatory/inhibitory (E/I) network. We demonstrate the counterintuitive fact that when decoding only from E cells, it is exclusively the I cell response gain and connectivity which govern how information changes. Additionally, we propose a parametrically simplified approach to studying the effect of state change on subpopulation codes. Our results reveal the importance of inhibitory circuitry in modulating information flow in recurrent cortical networks, and establish a framework in which to develop deeper mechanistic insight into the impact of cortical state changes on information processing in these circuits.

2.2 Introduction

Cortical circuits encode information about stimulus or action variables in their population activity [126]. These circuits then act on other cortical, musculoskeletal or endocrine systems to drive behavior. Studying cortical processing as an information transmission problem is a useful step toward relating neuronal activity and behavioral responses [104, 112, 76]. Theories of population coding require an understanding of both the sensitivity of trial-averaged stimulus tuning [132] and the structure of population-wide trial-to-trial variability [1, 9, 73]. Any changes in these two measures of neuronal response must be considered in order to evaluate whether the stimulus information available to a decoder of the population has increased or decreased. Such analysis has helped interpret experimental observations; for instance attention, well known to improve behavior in complex visual tasks [107], has been found to increase firing rates and neural selectivity while decreasing pairwise noise correlations along the visual pathway [91, 89, 24, 120, 110]. These effects are therefore taken to coincide with an attention-mediated improvement in information processing [142].

Biologically-motivated models of neuronal circuits define a network's state as the dynamical regime in which it is operating [62]. Top-down modulatory processes engaged during shifts in attention or arousal induce changes in the state of a cortical network through processes such as neuromodulation, feedback projections, and synaptic rearrangement [44, 93, 57, 87]. Any shift in network state will affect how a network responds to stimuli, observed as a shift in trial averaged tuning as well as response variability. Uncovering network mechanisms which enable cognitive processes like attention to improve behavioral performance must therefore take network state into account. Yet prior studies of information processing have focused on parametric models that are agnostic to underlying circuit mechanisms, so that any shift in tuning is made without considering a concomitant shift in variability (or vice versa) [132, 6, 9, 69, 1, 65]. There thus remains a gap in our mechanistic understanding of the way in which information changes as a function of cortical state.

In this study we explore how changes in cortical state affect information processing in neural circuits. We find that in any circuit, the information available to a linear decoder is invariant to network state when the decoder reads out from the entire population [68].

However, in many cases only a subset of the neurons project to a given decoder: inhibitory neurons have predominantly local projections [143] while excitatory neurons are often subdivided based on their outward projections [133, 151]. From the vantage of information processing we label this a *subpopulation code*. We show that when a network's state is modulated then the information available to a linear decoder of a subpopulation code is malleable. Intriguingly, we also find that the information encoded in a subpopulation does not explicitly depend on the activity and connectivity within the subpopulation. Rather, it is only those neurons within the circuit which connect *to* the projection cells that directly shape information flow. We thereby demonstrate that it may not be possible to draw significant conclusions on state dependent cortical processing without a more complete view of the network in question. Towards this end, we provide a framework for a circuit dissection that exposes how modulations of cortical state impact information processing in neural circuits.

2.3 Results

There are a vast array of mechanisms through which the brain modulates cortical network state. Two of those commonly studied, neuromodulators and feedback projections, can both uniquely affect circuit dynamics, from cellular excitability to transient synaptic weight changes [87, 57]. The neuronal correlates of these state changes are shifts in the firing rate, neuronal sensitivity, and correlations of populations of neurons. In relating these changes in neural response statistics to behavior we are really asking how a modulus influences cortical processing. Fisher information (FI) provides a means of addressing this question as it measures a decoder's ability to discriminate between two stimuli. It has been argued that in the regime in which sensory cortex lives, linear FI is equivalent to FI (and this likely extends to other areas of cortex as well, where linear decoders have best fit the relationship of neural activity to behavior) [117, 73]. We therefore focus on linear FI in this work, which is given by

$$\mathrm{FI} = \frac{d\mathbf{r}}{d\theta}^T \Sigma^{-1} \frac{d\mathbf{r}}{d\theta}. \tag{5}$$

Here $\frac{d\mathbf{r}}{d\theta}$ is the population response gain (change in rate for a change in a stimulus parameter θ) and Σ is the covariance matrix of the population response.

Now consider a recurrently coupled network of excitatory (E) and inhibitory (I) populations (Figure 3a) with dynamics described by

$$\tau_i \dot{r_i} = -r_i(t) + f_i\left(\sum_j J_{ij} r_j(t) + I_{0,i}(\theta, t)\right), \tag{6}$$

$$I_{0,i}(\theta, t) = b + s_{\text{ext}}(\theta) + \sigma_i(\sqrt{1-c}\zeta_i(t) + \sqrt{c}\zeta_c(t)). \tag{7}$$

Here i is the index of a unit in the E or I population, b is baseline input for all cells (potentially heterogeneous but assumed constant for simplicity), $s_{\text{ext}}(\theta)$ is an external drive that depends on the stimulus parameter θ, J_{ij} is the connection weight from unit j to unit i, r_i is the firing rate of unit i in the network, f_i is the transfer function of unit i, ζ_i is noise private to unit i, ζ_c is noise common to all units in the network and c scales the amount of shared variability (Eq. 18). We restrict our analysis to steady-state responses in which for sufficiently small input noise, the responses of the firing rate vector $\mathbf{r}$ can be linearized around the operating point (Figure 3c) [31]. Equation 6 then becomes:

$$\tau \delta \dot{\mathbf{r}} = -\delta \mathbf{r}(t) + \mathbf{L} \cdot (\mathbf{J} \cdot \delta \mathbf{r}(t) + I_0(\theta, t)) \tag{8}$$

where $\delta\mathbf{r}$ are the dynamics of the response around the steady-state solution and $\mathbf{L}$ is the linearization matrix about the steady state rate (Eq. 17; [35]).

In the context of this system, a network modulation is a change in the steady-state rate without a change in stimulus input (s_{ext}). We therefore define a modulation as a perturbation which changes the operating point of a network for a fixed stimulus (illustrated in Figure 3c). The space of perturbations is thus restricted to the space of network parameters (see also Figure 4a); some aspect of the network is changing such that we must relinearize about a new steady-state response. In this study we will consider moduli of two types: additive changes to the top-down input current (representative of cortical feedback [72]) or transient synaptic weight changes (e.g. neuromodulator-induced synaptic plasticity [87, 142]).

Previous modeling studies have captured the effects of attention on the response statistics of cortical circuits with a change in the top-down input drive to the network [68, 49, 63].

We introduced a similar top-down drive to a model recurrent E/I network to induce a state change (see Methods: Ring network). This modulus resulted in significant changes to the network response to a fixed stimulus: by contrast to the unmodulated network (state U), in the modulated network (state M) firing rates decreased across the excitatory population (Figure 3b) resulting in a decrease in the squared gain of the population (Figure 3d) and decorrelation across E cells (Figure 3e). Given that these statistics define FI, we expect these dramatic changes to the network's response statistics to affect information flow in this circuit. What we instead observed is that FI is invariant to this modulation (Figure 3f). We show in the next section (Eq. 10) that this is true for *all* possible moduli under our definition. This not only apparently contradicts the necessity that behavioral improvements would be reflected in enhanced information processing in cortex but also questions the effect of these neural response statistics on cortical processes more generally.

2.3.1 Modulation can improve information flow in subpopulation codes

In order to develop intuition for the invariance of FI to modulation we turn to an E/I network with a single E and single I unit (Figure 4a). The responses r_E and r_I to an input stimulus s and a perturbation $s+\delta s$ can be described as a joint probability distribution over the rates (the top row of Figure 4b). When a modulation is applied, the shape and center of the response distributions for s and $s+\delta s$ both vary in E-I firing rate space (Figure 4b, purple ellipses). For instance, modulation could induce a rotation of the distributions in r_E-r_I space (Figure 4b, state M), or change the correlations between r_E and r_I (Figure 4b, state M'). Overall discriminability between s and $s+\delta s$ does not change in either of these scenarios, however, since the total overlap of the distributions does not change. Hence the decoding error is constant across network state which is consistent with the result of the full network model described above (Figure 3f).

This would seem to suggest that modulation of cortical circuits is functionally insignificant, and that optimization for sensory discrimination should act on other mechanisms, such as feedforward thalamic inputs. However until now, we have made a tacit assumption that all neurons encode the stimulus variable. Yet from the standpoint of a downstream cortical

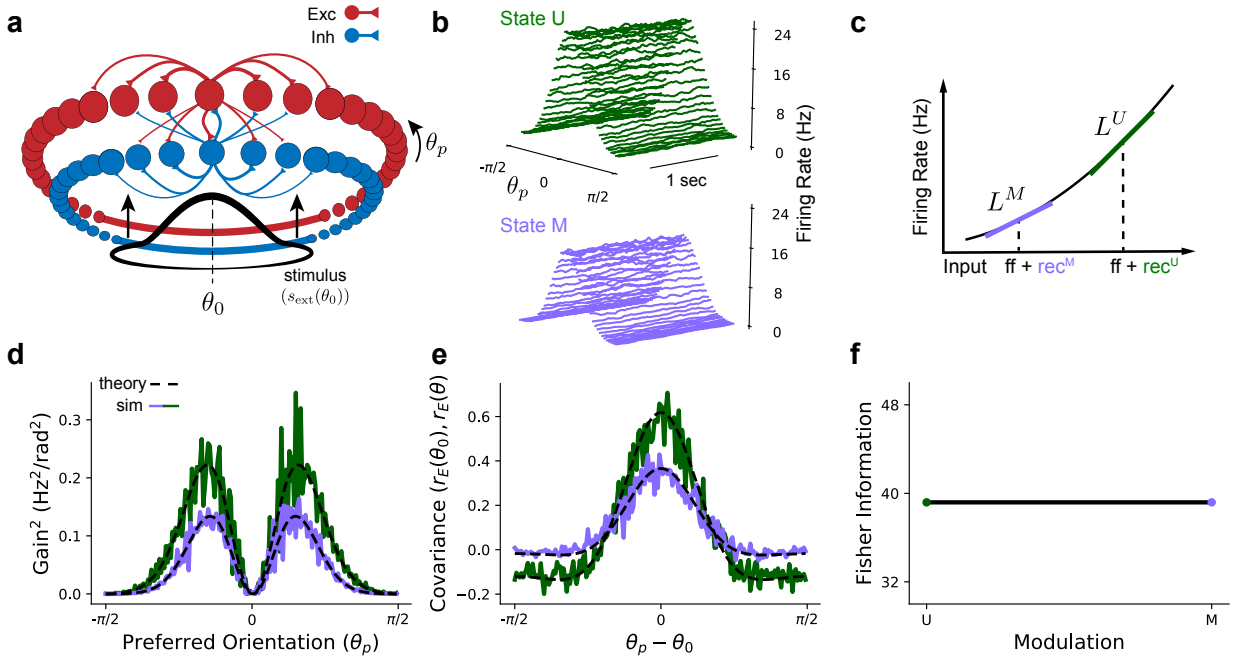

Figure 3: *Changes in network activity do not imply changes in information.* **a** Network schematic. Units are arranged on a ring indexed by θ where position on the ring corresponds to that unit's preferred value of θ. Red units are excitatory and blue are inhibitory. Size of connection line indicates strength of connection. Not all connections are shown. **b**. Firing rates for the excitatory network units across time in the unmodulated (U) and modulated (M) states. **c**. Illustration of the effect of state changes on the input/output response for a single unit in the excitatory population before (green) and after (purple) modulation. **d**. Fit of the squared gain across the E population in a nonlinear model (solid lines) with a linear theory (black dashed lines). **e**. Fit of integrated cross-covariance relative to a single E unit (with preferred orientation θ_0) with a linear theory. Colors as in (b). **f**. Plot of Fisher information as a function of modulation.

area only those neurons which project to it will matter in the readout. In particular, we know from anatomical studies that excitatory (pyramidal) neurons are the dominant projection cells in cortex. Hence information read out of a population by downstream areas must be predominantly conveyed through excitatory pathways [124]. Since our decoder is functionally equivalent to a downstream region decoding an upstream signal, we therefore reconsider the same problem by decoding from the excitatory population alone. This corresponds to a projection of the joint r_E/r_I firing rate distribution onto the r_E axis (Figure 4b, bottom row). Similarly to the joint distributions, the distributions of excitatory firing rates can change significantly with state as well. Critically, we observe a decrease in the overlap of the r_E distributions with modulation (decrease in the error), indicating an increase in the discriminability in the E population following each modulation. Thus, while FI for the full network is invariant to state modulation, we see that FI restricted to the E population alone can change with state.

To formalize the arguments above in terms of FI, we define external stimuli for this E/I network as $s_{\text{ext},E} = k_E s, s_{\text{ext},I} = k_I s$ in equation 7, where k_α is a sensitivity term which scales the size of a feedforward stimulus drive to population $\alpha \in \{E, I\}$. Then the linear FI for the full E/I network is given by [68]:

$$\text{FI}(P_{\text{ext}}, P_{\text{net}}) = \frac{\sigma_I^2 k_E^2 + \sigma_E^2 k_I^2 - 2\sigma_E \sigma_I k_E k_I c}{\sigma_E^2 \sigma_I^2 (1 - c^2)} = \text{FI}(P_{\text{ext}}). \tag{9}$$

This equation is independent of all network parameters (P_{net}), depending only on the external input parameters, P_{ext} (Figure 4a). Since we have defined a modulation to only affect the network state (and consequently only impact network parameters P_{net}), FI for the full population must be invariant to modulation.

This result is true in general for any network size; linear FI reduces to a simple form which depends only on the input gain and input covariance since the output gain and output covariance depend on the network linearization in the same way which subsequently cancels (see Methods). Thus, in a linearized system, FI reduces to:

$$\text{FI} = \Phi^T \Sigma_{\text{ext}}^{-1} \Phi, \tag{10}$$

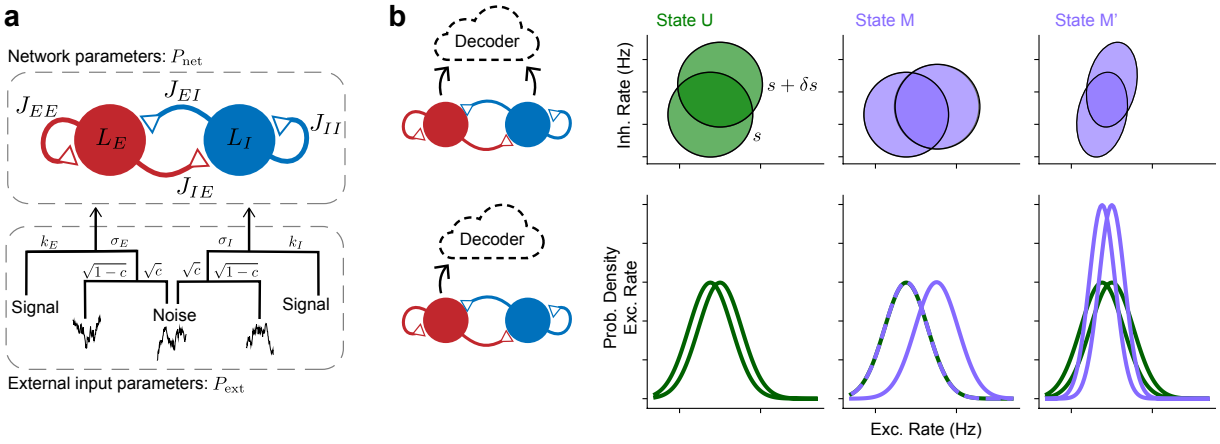

Figure 4: *Projection to lower dimensions allows for improved discrimination with modulation* **a**. Illustration of E/I circuit parameters, partitioned into external input parameters (P_{ext}; not affected by modulation) and network parameters (P_{net}; subject to change through modulation). **b**. (Top row) 95th percentile distributions of idealized steady state rates at contrasts s and $s + \delta s$ for unmodulated (green) and modulated (purple) networks. A decoder reading out from the full network (top left) has access to the joint r_E/r_I distribution. Modulation of a linear model does not change discriminability in high dimensions because the overlap (error) between the joint r_E/r_I distributions over s and $s + \delta s$ does not change from U to M or M'. (Bottom row) Projection of the joint r_E/r_I distribution onto the r_E axis is the same as a decoder restricted to observing only the E population (bottom left). The state changes in the r_E/r_I space permit increased discriminability in the E population due to decreased overlap of the distributions.

where Σ_{ext} is the input covariance matrix and $\Phi = \frac{ds_{\text{ext}}}{d\theta}$ is the input gain, that is, the derivative of the stimulus input with respect to the tuning parameter θ. Since equation 10 is simply the N-dimensional analogue of equation 9 and similarly depends only on P_{ext}, this result explains the invariance of FI to modulation for any network (as in Figure 3d).

We now seek an expression for FI in terms of only the E population consistent with a readout restricted to projection neurons. For the two-unit E/I network, the information read out from the E population alone is defined as [68]:

$$\text{FI}_E = \frac{G_E^2}{V_E} \tag{11}$$

where $G_E = \frac{dr_E}{ds}$ is the gain and V_E is the variance of the E population. This expression is the information analogue to a projection of the readout onto the r_E axis. Because it involves restricting readout to a subset of the population we refer to this as *subpopulation coding*. After some algebra we can write FI_E as

$$\text{FI}_E(P_{\text{ext}}, P_{\text{net}}) = \frac{(k_E - k_I x)^2}{(\sigma_I x - \sigma_E c)^2 + \sigma_E^2(1-c^2)} = \text{FI}_E(P_{\text{ext}}, x(P_{\text{net}})), \tag{12}$$

where $x = \frac{L_I J_{EI}}{1+L_I J_{II}}$. It is now apparent that the information in this subpopulation *does* depend on network state through the variable x, which is a function of the network parameters P_{net}. Therefore FI_E can change with modulation (compare equations 12 and 9). Surprisingly, equation 12 reveals that this dependence on network state comes only from L_I, J_{EI} and J_{II}; that is, the information gleaned from the E population depends only on inputs to the network, together with the linearization of the I population and the I connectivity (Figure 5a, blue), but not explicitly on either the E gain or excitatory recurrent connections (Figure 5a, gray). Said differently, it is only those units which project *into* the readout population dictate the extent to which information readout changes with network state.

From the view of a linear decoder restricted to the E population, the recurrent network can be reduced to a feedfoward inhibition model, where E receives inputs $I_{0,e}$ from external sources, and $-x \cdot I_{0,i}$ from the I population (Figure 5b) [86]. Here, x is the *effective coupling* from the inhibitory to the excitatory population (Figure 5b). Hence the effective stimulus gain of the E population is $\frac{d}{ds}(I_{0,e} - xI_{0,i}) = k_E - k_I x$ and the variance of the effective

total input is $\mathrm{Var}(I_{0,e} - xI_{0,i}) = (\sigma_I x - \sigma_E c)^2 + \sigma_E^2(1 - c^2)$, which are the numerator and denominator of FI_E (Eq. 12), respectively.

The effective coupling parameter x affects both the gain and the variance of the E population responses (Eq. 12; [86]). First suppose that the I population does not receive signal input, meaning that $k_I = 0$. Then FI_E can change only through the denominator, corresponding to the variance of the effective input. FI_E is maximized at $x = \frac{\sigma_E}{\sigma_I} c$, when the correlated noise from I population cancels the correlated component of the input noise to E population and minimizes the variance of the total input (Figure 5c, top). We refer to a modulation which pushes x closer to this value as *correlation canceling*. When there is no correlation in the inputs to E and I populations ($c = 0$), I merely contributes noise to E through $\sigma_I x$. In this case, FI_E is maximal when $x = 0$, meaning that there is no projection from I to E.

By contrast, when I receives signal input ($k_I > 0$), I has a subtractive effect on the E stimulus response since I projects the same signal to E with a negative sign (Figure 5b). Therefore modulating the gain is now weighted against affecting variability to enhance information. For x small, reducing x further to minimize the subtractive impact on the E response is optimal, even if correlations are large (Figure 5c, bottom). As a result we label our network as being in the *gain reduction* regime. However, this brings up an important constraint since network inhibition plays an important stabilizing role as well. Thus, any changes in x must ensure network stability as well (Figure 5c, hatched region). Note that while we have reduced the information dependence to a single network hyper-parameter, x, the network's stability depends on *all* of the network parameters P_{net}.

The above analysis of a two-unit network corresponds to a homogeneous neuron population with identical tuning dependence on a stimulus variable. Next, we consider a population of neurons with distributed tuning preference over the encoded range of a stimulus variable, such as orientation.

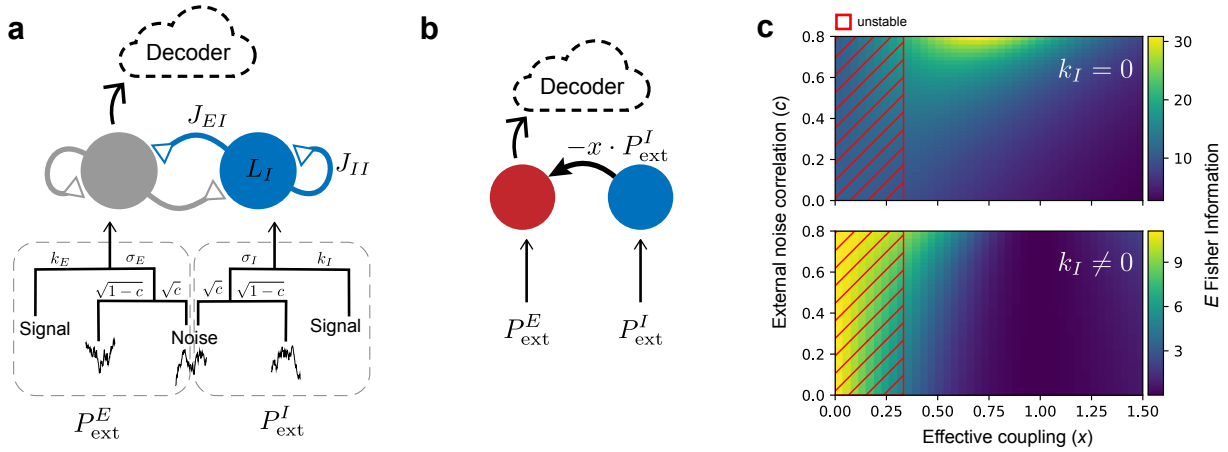

Figure 5: *Changes in FI_E depend only on inputs to E.* **a**. Illustration of the network parameters that affect FI_E in the network described in Figure 4 (blue). Network parameters which do not affect FI_E are in gray. P_{ext} has been split into E- and I-specific external inputs (P^E_{ext} and P^I_{ext}, respectively). **b**. Reduced network model. FI_E depends only on feedforward inputs to E. **c**. FI_E as a function of $x(J_{EI}, J_{II}, L_I)$ and c. Red hatched region indicates unstable solution for a single choice of parameters P_{net}. (top) $k_I = 0$; $x_{max} = \frac{\sigma_E c}{\sigma_I}$. (bottom) $k_I = 1$; $x_{max} = \frac{\sigma_E^2 - \sigma_E \sigma_I c}{\sigma_E \sigma_I c - \sigma_I^2}$. Stability line for parameters $L_E = 10$, $J_{EE} = 0.18$, $J_{IE} = 0.24$.

2.3.2 Subpopulation codes with distributed tuning

We return to a recurrent network in which N excitatory units connect to N inhibitory units (as in Figure 3a). A stimulus such as an oriented visual grating is now given by $s_{\text{ext}}(\theta)$ with each unit's preferred tuning parameter corresponding to a particular orientation θ.

We again consider FI_E as in the previous section, a decoder observing only the E population, which for a collection of N excitatory units takes the form (Methods)

$$\text{FI}_E = (\Phi_E - X\Phi_I)^T[\Sigma^E_{\text{ext}} - \Sigma^C_{\text{ext}}X^T - X\Sigma^C_{\text{ext}} + X\Sigma^I_{\text{ext}}X^T]^{-1}(\Phi_E - X\Phi_I) \tag{13}$$

where Φ_α is the input gain Φ restricted to population $\alpha \in \{E, I\}$, $\Sigma^\alpha_{\text{ext}}$ is the input covariance matrix to α and Σ^C_{ext} denotes the input covariance between E and I, and $X = \mathbf{J}_{EI}(\mathbf{L}_I^{-1} + \mathbf{J}_{II})^{-1}$ is the effective coupling matrix from the inhibitory to the excitatory population (Figure 5b). Equation 13 is simply the N-dimensional analogue of equation 12, thereby confirming that our preceding analysis extends to arbitrary dimension in FI_E. What again distinguishes FI from FI_E is that the former depends only on the structure of the input statistics to the network whereas the latter depends additionally on those network parameters restricted to the I population, $\mathbf{L}_I, \mathbf{J}_{EI}$ and $\mathbf{J}_{II}$. In particular, if we look back at the ring network which motivated this study (Figure 3), FI_E now increases.

By expanding the dimension of our recurrent network we have expanded the space of possible moduli. A complete characterization of the whole parameter space is out of the scope of this work, we are interested here in whether the mechanisms for increasing FI_E observed in the two-unit E/I network relate to phenomena observable in the N-dimensional network, namely gain modulation and correlation cancellation. We model a neuromodulatory effect as a transient rescaling in synaptic weights, $\mathbf{J}_{EI}$ [87] (Figure 6ai,aii; blue connections). Reducing $\mathbf{J}_{EI}$ here diminishes the effective projection from I to E thereby disinhibiting E.

We again see that the net input gain and covariance to the E population produces FI_E. To understand the impact each has on FI_E we changed X in equation 13 from unmodulated (X^U) to modulated (X^M) in either the net input gain ($\Phi_E - X\Phi_I$) or the covariance term ($\Sigma^E_{\text{ext}} - \Sigma^C_{\text{ext}}X^T - X\Sigma^C_{\text{ext}} + X\Sigma^I_{\text{ext}}X^T$) alone while keeping the other X's fixed in the unmodulated state. Both substitutions result in increased FI_E (Figure 6b), with a change in the gain term

resulting in an approximate three-fold increase and a change in the covariance resulting in an approximate doubling of information. What this illustrates is that both gain and covariance changes play significant roles in improving information readout. Furthermore, their joint effect is multiplicative, as the net increase in FI_E with modulation is almost six-fold.

We now explore the mechanism by which X increases gain and decreases covariance. In our framework external inputs to the network do not change with a state modulation by assumption, thus the E input gain Φ_E and input covariance Σ^E_{ext} are fixed across modulation (Figure 6ci,ei). Since $\mathbf{J}_{EI}$ decreases in magnitude, the suppression of signal gain from I inputs ($X\Phi_I$) is reduced (Figure 6cii) leading to an increase in the net E input gain (Figure 6d). Additionally, this modulation resulted in a reduction of the projected I covariance, that is, the variability in E due to the I to E connection (Figure 6eii). In particular, the modulation of the projected I covariance results in the partial cancellation of correlations and a net reduction in E variability (Figure 6f). These joint improvements in signal and noise therefore combined nonlinearly to induce the increase in FI_E (Figure 6b). In conclusion, the two-unit model accurately anticipated how modulation can affect population codes through a combined effect on the effective gain and covariance.

Previous information-theoretic studies have identified a source of correlated variability, termed differential correlations, which causes information to saturate in a population [97, 69]. The differential correlation imposes an upper bound on the efficacy of population codes. As we show in the Supplemental, differential correlations limit FI_E as well. However, modulation can still increase information in a neural population provided the system has not yet saturated the bound (Figure 9). Recent studies have attempted to estimate information-limiting correlations in primary visual cortex in mouse and monkey, arguing for the presence of differential correlations which bound the information encodable by the neural population [96, 66]. However a powerful recent study recording from tens of thousands of neurons across thousands of trials did not find evidence for information saturation in mouse V1 [137]. All of these studies have considered the collective activity of V1 as the full population, encoding an oriented visual grating, say, and relating it to behavioral performance. Previous anatomical experiments have shown that feedforward projections from V1 are rather segmented into partially overlapping patches [124]. In our view, a decoder would represent the downstream

area connected to a particular upstream patch. This partitioning of cortical areas could explain how, even in the presence of information-limiting correlations, neural processing can be kept away from the information saturation bound and therefore modulated with network state.

2.3.3 The implications of subpopulation codes for divergent cortical pathways

The results we have described thus far are not unique to a partitioning of E/I networks by excitatory and inhibitory neurons. Rather, they are defined in terms of readout (or: observed) and non-readout (unobserved) populations. Given the extensive branching of corticocortical projections, this differentiation is relevant for pathways diverging from within the same cortical area to project to disjoint downstream targets. In this way, local recurrent connections or parallel yet interconnected pathways can influence one another's information processing.

The benefits of our framework in analyzing information flow through the circuit can be seen by considering two recurrently connected E populations stabilized by a single I population (Figure 7ai,bi). A decoder reading out from both E populations is equivalent to FI_E computed in the preceding sections. As before, FI_E is invariant to changes in E activity and to all E connections. This can easily be seen from equation 13 applied to this network in which $X = \begin{pmatrix} x_{E_1 I} \\ x_{E_2 I} \end{pmatrix} = \begin{pmatrix} \frac{L_I J_{E_1 I}}{1+L_I J_{II}} \\ \frac{L_I J_{E_2 I}}{1+L_I J_{II}} \end{pmatrix}$. What this formulation nicely reveals is that the components of X are simply the two x's for the subnetworks E_1/I and E_2/I (see effective connectivity diagrams, Figure 7aii axes). In particular, each component of X is the effective projection from I into each element E_1, E_2 of the readout population. Similar to the case of only one E population (Figure 5), if we examine the information landscape with the chosen parameter set, reducing $x_{E_1 I}$ and $x_{E_2 I}$ would lead to an increase in information, consistent with a disinhibitory mechanism (Figure 7aii). For example, movement toward the $x_{E_1 I}$ axis would be achieved by weakening the $I \to E_2$ connection (the alternative of changing L_I or J_{II} would of course affect both $x_{E_1 I}$ and $x_{E_2 I}$). However it must be ensured that this modulation does not lead to instabilities in the network, a condition which again depends on all network parameters P_{net} (red region, Figure 7aii).

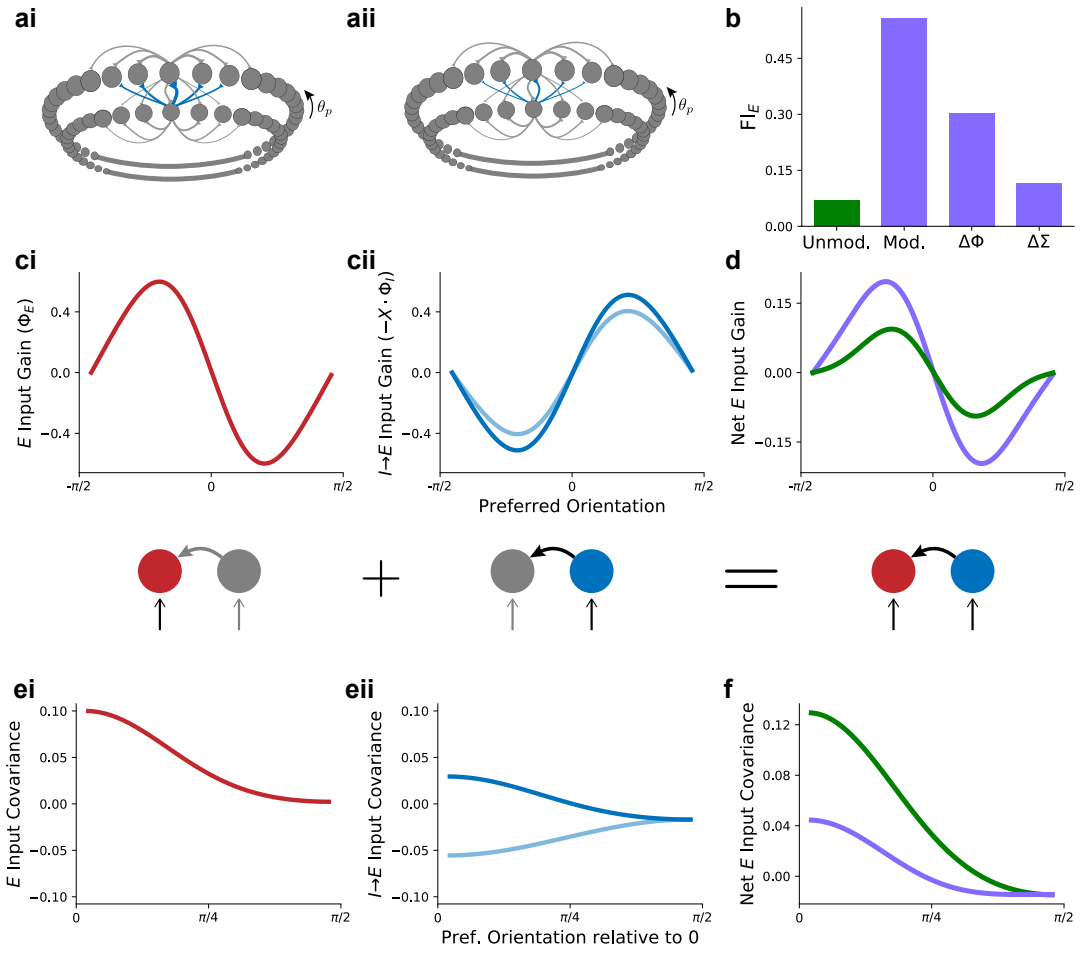

Figure 6: *Modulation affects inputs to E.* **a**. Network schematic of modulation via changes in synaptic strength. **b**. FI_E for unmodulated (green), modulated (mod.), modulated with changes in gain only ($\Delta\Phi$) and modulated with changes in covariance only ($\Delta\Sigma$). **c-f**. 0^o is chosen to be the unit whose preferred orientation matches the peak of the stimulus. **c**. (i) External input gain *to* E units is constant with modulation. (ii) I inputs to E are scaled by X and depend on modulation (light blue). Schematics as in Figure 5b indicate which network elements determine the plotted values above and below. **d**. Net external input gain to E before (green) and after (purple) modulation. **e** (i) Illustration of a row from the input covariance to E. (ii) The effective input noise from I to E, $-\Sigma^C_{\text{ext}}X^T - X\Sigma^C_{\text{ext}} + X\Sigma^I_{\text{ext}}X^T$, before (dark blue) and with (light blue) modulation. **f**. Total input covariance to E. Colors as in (d).

Now suppose E_1 and E_2 project to different targets, and consider the downstream decoder reading out from only E_1 (Figure 7bi). In this case we want to analyze FI_{E_1} and $X = \begin{pmatrix} x_{E_1E_2} & x_{E_1I} \end{pmatrix}$ now becomes effective couplings from E_2 and I. The two components of X, $x_{E_1E_2}$ and x_{E_1I}, are composed of all effective paths from populations E_2 and I, respectively, *to* E_1 (Figure 7bii axis diagrams; Eqs. 24, 25). In this case, FI_{E_1} is high when $x_{E_1E_2}$ and x_{E_1I} are both large or small (Figure 7bii). Therefore, information of the E_1 population can be increased by jointly increasing or decreasing $x_{E_1E_2}$ and x_{E_1I}, which can be achieved by, for example, strengthening or weakening $J_{E_1E_2}$ and J_{E_1I} (Figure 10). Note that there is a region of inaccessible values of $x_{E_1E_2}$ and x_{E_1I} (Figure 7bii, white region), due to the restrictions Dale's law places on the signs of the connection weights (i.e. $J_{\alpha E} \geq 0$; see Supplemental Material). Similarly, the parameter space for FI_E is restricted to positive values of x_{E_1I} and x_{E_2I} (Figure 7aii). The inaccessible region of xs depends on connection strengths and cellular gains (Figure 8).

In a general network of multiple units, we find that X comprises all paths through the unobserved units together with all projections to each readout unit (Figure 7c and Supplemental Material). The analysis is similar to previous works which decompose the network response covariance into structural motifs [145, 128].

In sum, the modulation of information flow in cortical networks is affected by all input connections to the readout population. This poses a thornier issue for populations projecting to divergent targets than for E networks with the same decoder since activity along one pathway can influence information flow along another connected path. As we speculate in the Discussion, this result could motivate compartmentalization, or clustering, of activity in cortex.

2.3.4 Parametric considerations in the theory of subpopulation codes

While we have thus far argued for the biological importance of our theory, it has a mathematical benefit as well. In order to understand the effect of modulation on information flow in a network, traditional analysis would require knowledge of all network parameters. For a network of N units this is $N(N+1)$ different parameters if we consider all possible

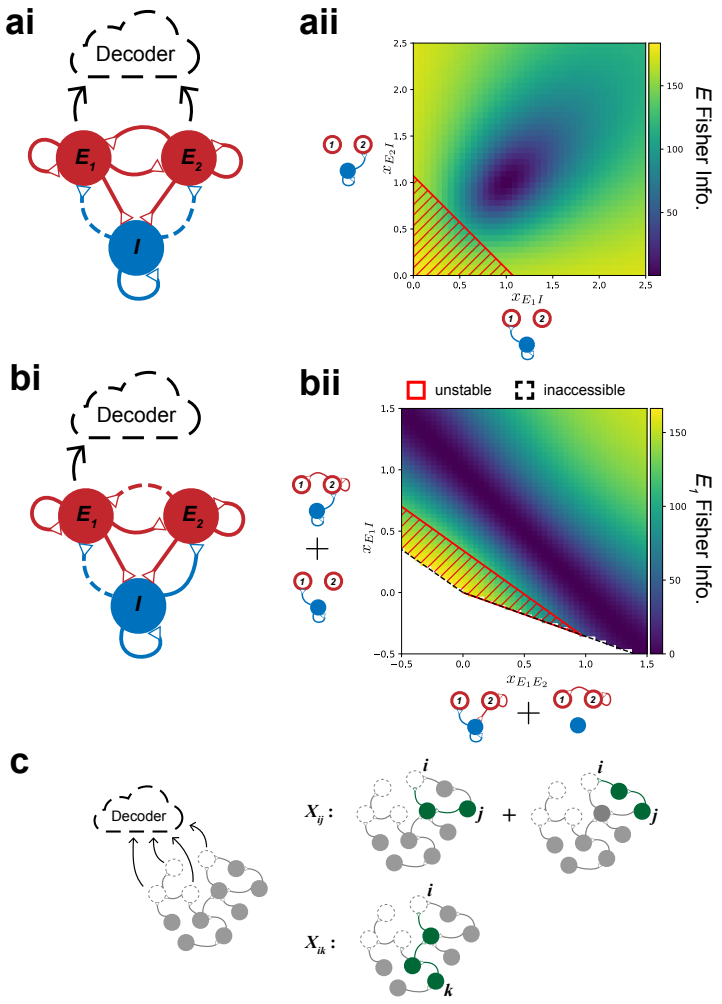

Figure 7: *Information flow through E subpopulation* **a**. (i) Network schematic. Dashed lines illustrate which connections were varied in computing stability bounds in (ii). Readout is from both E_1 and E_2. (ii) FI_E for x_{E_1I} and x_{E_2I} for fixed input covariance and stimulus. Axes show the effective connections which comprise each x. Red region indicates instability. **b**. (i) As in (ai) for readout from only E_1. (ii) As in (aii) where the white region indicates inaccessible values of $x_{E_1E_2}$ and x_{E_1I} for the chosen parameter set. **c**. Illustration of the general form of elements of X. A generic network is shown (grey) with only a subset of units available to the decoder (grey dashed circles). Two elements of X are illustrated in terms of the pathways through the network which contribute to them (green).

connections (N^2) and each unit's stimulus-response linearization (N). In our theory, however, if readout is restricted to m of the N total units one must only understand how modulation affects $m(N-m)$ parameters, the number of elements in X. This can offer a significant advantage. For instance, in the preceding example we reduced the relevant parameter space from 12 to 2 when decoding from E_1.

Other considerations such as system stability will yet demand full knowledge of the system. In spite of this, our theory still affords some benefit (Figure 8). We return to the question of decoding from E_1 in the same network as in the preceding section (Figure 7) by varying four connection strengths ($J_{E_1E_2}$, J_{IE_2}, J_{E_1I} and J_{E_2I}; Figure 8ai). If we look at how FI_{E_1} changes when viewed as a function of these four connection strengths, both the stability boundaries and the information landscapes change across panels (Figure 8aii). Describing a consistent theory for how modulation can enhance information flow in this context would be exceedingly difficult; there are still five unexplored connections. By contrast the same parameter changes replotted in terms of $x_{E_1E_2}$ and x_{E_1I} show a different picture (Figure 8bi): the information space is invariant, and only stability and accessibility boundaries change (Figure 8bii). Given that a modulation can drive a change either within or across plots, this representation affords a much clearer picture of how to modulate the network once network stability is taken into account. We comment, however, that the connection parameters we manipulated here still offer a peek behind the curtain; without knowledge of which connectivity parameters most influence X it would not have necessarily been *a priori* obvious to consider $J_{E_1E_2}$ vs. J_{E_1I}, and one might have had to explore all nine connectivity values to arrive at the same conclusion.

2.4 Discussion

In this study we explored how information within a neural population changes as a function of cortical state. We have shown that for cortical state to affect information flow in a neural circuit, a linear decoder must only observe a subset of the complete neural population, consistent with cortical anatomy in which only a subset of neurons project to downstream

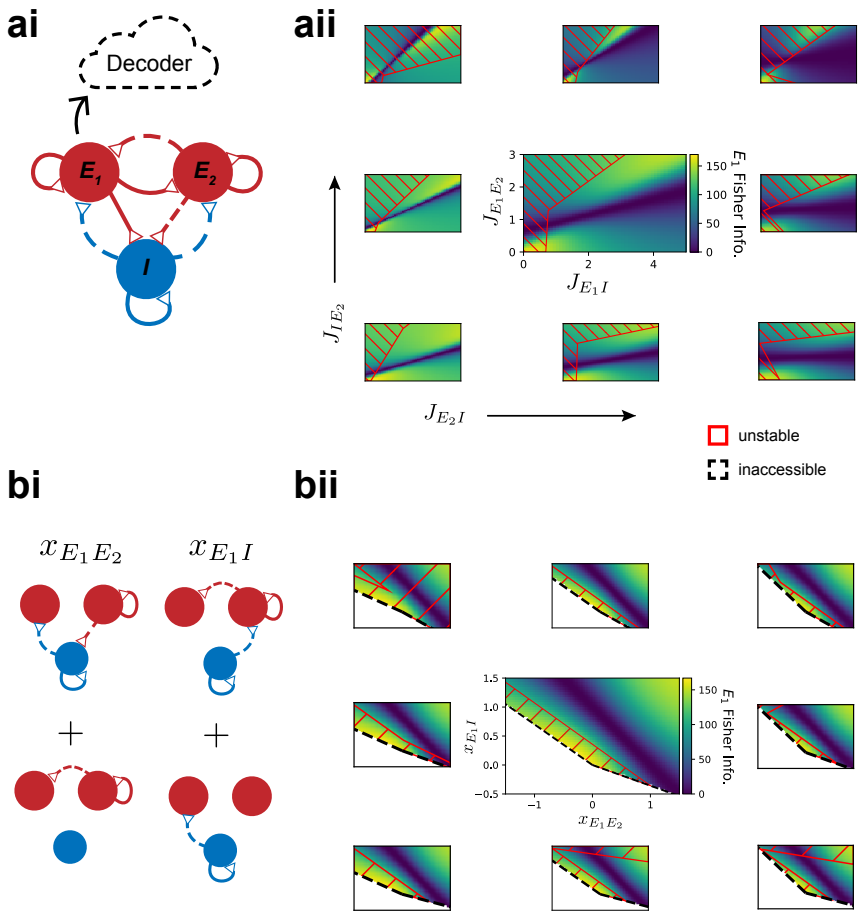

Figure 8: *Parametric benefits of the theory.* **a**. (i) Schematic of the network. Dashed lines represent connections varied within and across plots (stability boundaries computed as in Figure 7). (ii) FI_E as a function of varying connectivities (J's) for a semi-naive surf of J-space. Red regions indicate instability. Direct connections from E_2 and I to E_1 were varied within a plot. Connections between E_2 and I were varied across plots. **b**. (i) Illustration of connections contributing to $x_{E_1E_2}$ and x_{E_1I} with varied connections dashed. (ii) Same data and arrangement as in (aii) plotted as a function of $x_{E_1E_2}$ and x_{E_1I}. Central plot in (aii) and (bii) is same parameter set as Figure 7bii.

targets [55]. Moreover, we observed the counter-intuitive result that under changes in cortical state, it is only those neurons which are not decoded from (i.e. do not project to a downstream region) that shape linear readout from a neural population that does project downstream. This suggests a strong role for inhibition in shaping information flow in cortex, a view which is gaining broad support [61]. Furthermore, we showed that subpopulation linear Fisher information depends only on the structure of the external (e.g. thalamocortical) inputs to the network, together with the non-readout connections and input/output linearization, effectively reducing subpopulation codes to a feedforward circuit.

In the ring model with distributed tuning, we assumed that any current modulation is independent of the tuning preference of neurons (Figure 3). While tuning-dependent modulation has been used to model selective attention [37, 82], it trivially adds information to the network. Specifically, denoting current modulation as $M(\theta)$, we would have $\Phi = \frac{ds_{\text{ext}}}{d\theta} + \frac{dM}{d\theta}$. Hence, in this work we only considered modulations that do not trivially introduce information to the network, such as transient synaptic weight changes and tuning-independent current modulation.

Throughout this study we largely focused on a classic E/I dichotomy in which excitatory neurons conveys information to downstream regions while inhibitory neurons solely act locally. However our analysis proved relevant to other common circuit motifs such as those in which local excitatory cells shape projection cell output or divergent projection cells mutually interact. As an example, intracortical pyramidal neurons in motor cortex innervate corticostriatal cells which project from motor cortex to various subcortical and peripheral targets [133]. Additionally, deep layers of cortex are also the source of reciprocal connections between divergent projection neurons: corticothalamic neurons which project subcortically and intratelencephalic neurons which project within cortex [56]. Given the high divergence of cortical projection pathways and the preponderance of clustered network architectures across cortex, our results support an information-processing benefit to a clustered organization [84]. By limiting the interaction between neural clusters performing distinct computations, the brain could achieve better control over the degree to which activity in one cluster affects the information flow in another.

Previous studies have considered readout from only the excitatory neurons [86, 99, 68].

However these studies either did not explore the impact of modulation on the decoder [86] or considered only homogeneous tuning inputs [68], leaving open the question of whether and how these results extended to networks with distributed tuning. A recent study [99] explored the coding capacity of inhibitory neurons, thus comparing E vs. I information, and examined how changes in E/I connectivity (J_{EI}, J_{IE}) affected information content in E and I populations separately. Our results clarify why varying these connections was the only way to affect information in their network: their use of a linear system fixed the cellular gain (L in our theory), and the only other terms the effective connectivity (X) depend on are the J's.

The effects of state modulation on neural circuits are highly diverse, capable of adjusting a range of properties from cell-intrinsic - such as excitability and neurotransmitter release [51, 87] - to population-wide, such as oscillatory activity and noise correlations [142]. The mechanisms underlying state changes are equally diverse, involving many classes of neuromodulator and different sources of feedback drive [57]. Neuromodulators can be broadly distributed, such as cholinergic projections from midbrain to cortex, or highly specific, like dopaminergic targeting of individual cortical layers [141, 142]. This seemingly endless flexibility poses a challenge for determining how these moduli relate to neural processing. Our theory goes some way towards identifying the circuit components which are actually affecting neural processing. For example, we found that reducing the synaptic strength between I and E can increase gain and decrease covariance in excitatory units. These effects are mirrored by acetylcholine (ACh), which has been shown to reduce synaptic efficacy of intracortical connections [142].

Neuromodulation can affect the responsiveness of neurons in many ways that our model does not capture, for example, by reducing burst spiking and altering firing adaptation [57]. Of course, the omission of a spiking mechanism is a clear limitation of rate models, and future work should address the way in which modulation of spiking properties affects neural coding across states. Our choice of firing rate models are nevertheless able to capture the main effects observed in many experimental paradigms such as primate electrophysiology experiments in which analysis is performed on spike counts [24], or calcium imaging experiments [138]. In fact, we anticipate that the spatially broad recording capacity of modern calcium imaging,

together with neural subtype indicators will provide the data our model has identified as important to understanding the circuit mechanisms of information modulation, namely, the activity of local (largely inhibitory) interneurons.

Our results also highlight the general difficulty in assigning a functional role to specific network components in affecting information flow. If we consider an arbitrary connection J_{EI} this can emerge in multiple paths from the non-readout to the readout population, thereby affecting multiple values in X (Supplemental Material; Figure 10). Similarly, the linearization L_α of unit α contributes to all values of X in which connections from unit α are present. These issues are natural consequences of the recurrent nature of cortical circuits. In spite of this, the ability to map a recurrent circuit to an effectively feedforward model with an interpretable hyperparameter, X, will facilitate uncovering the mechanics of state-dependent information modulation.

Understanding how neural processing depends on brain state is a major goal in systems neuroscience [35]. Even in simple subcortical systems like the crab stomatogastric ganglion in which the full connectivity structure is known, a modulus's effect on circuit dynamics is still difficult to generalize as it depends upon the relationship between the network state and the nature of the modulus [88]. Expanding to cortical circuits of much larger size poses a daunting challenge, however, our theory significantly reduces the space of parameters which need to be measured, as well as provides a degree of interpretability by representing information flow in terms of effective pathways within a network. Our results highlight the importance of locally projecting neurons in shaping the information in neurons that project to downstream areas.

2.5 Methods

2.5.1 Linear theory

We considered a linearization of a nonlinear firing rate model (equation 6). For sufficiently small noise, equation 6 can be linearized around the steady state rate, with fluctuations in

rate around the operating point given by an N-dimensional extension of traditional linear response theory:

$$\tau \delta \dot{\mathbf{r}} = W \delta \mathbf{r} + D\zeta, \tag{14}$$

$$W = \left(\begin{array}{c|c} -1 + \mathbf{L}_E \mathbf{J}_{EE} & -\mathbf{L}_E \mathbf{J}_{EI} \\ \hline \mathbf{L}_I \mathbf{J}_{IE} & -1 - \mathbf{L}_I \mathbf{J}_{II} \end{array}\right), \tag{15}$$

$$D = T_n \left(\begin{array}{c|c|c} \mathbf{L}_E \sigma_E \sqrt{1-c} & 0 & \mathbf{L}_E \sigma_E \sqrt{c} B_E \\ \hline 0 & \mathbf{L}_I \sigma_I \sqrt{1-c} & \mathbf{L}_I \sigma_I \sqrt{c} B_I \end{array}\right) \tag{16}$$

where ζ is the $(N+\nu)$-dimensional vector of external input noise (here $\nu = 1$ is the dimension of the shared external input variability), W is an $N \times N$ matrix of effective weights, D is $N \times (N+\nu)$ matrix which scales the noise terms, B_α is the $N_\alpha \times \nu$ matrix which determines the input covariance structure, σ_α is the amplitude of the noise to a given unit, $\tau = \left(\begin{array}{c|c} \tau_E 1 & 0 \\ \hline 0 & \tau_I 1 \end{array}\right)$ where 1 is the $N_\alpha \times N_\alpha$ identity matrix. $\mathbf{L}_\alpha$ is the diagonal matrix of derivatives of the transfer function at the steady state rate whose i^{th} diagonal element is given by

$$L_i = \frac{df_i}{d\theta}\big(\sum_j J_{ij}\bar{r} + \hat{I}_{0,i}(\theta)\big), \tag{17}$$

with $\bar{r}$ the steady-state solution to equation 6 and $\hat{I}_{0,i}(\theta) = b + s_{\text{ext}}(\theta)$. In order to avoid injecting pure white noise into the system, we consider a temporally smoothed noise process

$$\tau_n \dot{\zeta}_i = -\zeta_i + \sqrt{\tau_n}\xi_i \tag{18}$$

where ξ_n is a white noise process and i is an index over all possible noise sources. Consequently, T_n is a time-scaling constant given by $T_n = \sqrt{2\tau_n}$. This linearized stochastic system then enables us to estimate the full covariance matrix of the spatially extended model [46]. In particular, the covariance matrix is given by $\Sigma = W^{-1}D(W^{-1}D)^T$. It should be noted we have written the connectivity parameters $\mathbf{J}$ as positive values; the negative sign of inhibition is made explicit in the dynamical equations and carried through the relevant derivations.

2.5.2 Gain calclulation

The population response gain is given in general by the derivative of a population's response with respect to an input parameter. In the reduced E/I model we have $s_{\text{ext},\alpha} = k_\alpha s$ such that $G = \frac{d\bar{\mathbf{r}}}{ds}$. In the ring network, the parameter of interest is the angular variable θ such that $G = \frac{d\bar{\mathbf{r}}}{d\theta} = \mathbf{L}\frac{dI}{d\theta} = \mathbf{L}(\frac{ds_{\text{ext}}}{d\theta} + \frac{dM}{d\theta} + \mathbf{J}\frac{d\bar{\mathbf{r}}}{d\theta}) = \mathbf{L}(\frac{ds_{\text{ext}}}{d\theta} + \frac{dM}{d\theta}) + \mathbf{LJ}G \implies G = (\mathbf{1} - \mathbf{LJ})^{-1}\mathbf{L}(\frac{ds_{\text{ext}}}{d\theta} + \frac{dM}{d\theta}) = (-W^{-1})\mathbf{L}(\frac{ds_{\text{ext}}}{d\theta} + \frac{dM}{d\theta})$. Here we have let I equal the argument of f in equation 6. We have generalized this derivation with the inclusion of a current modulus $M(\theta)$ such that $I = \mathbf{J} \cdot \mathbf{r} + I_0 + M(\theta)$ where I_0 is given by equation 7; a transient weight change would simply affect $\mathbf{J}$. Finally, letting $\Phi = \frac{ds_{\text{ext}}}{d\theta} + \frac{dM}{d\theta}$ we arrive at the following compact expression: $G = -W^{-1}\mathbf{L}\Phi$.

2.5.3 Fisher information analysis

2.5.3.1 Full FI

As mentioned above we write the (long time) covariance matrix $\Sigma = (W^{-1}D)(W^{-1}D)^T = W^{-1}DD^T(W^{-1})^T$. Notice that we can rewrite D such that $D = \mathbf{L}D_{\text{ext}}$ where $D_{\text{ext}}D_{\text{ext}}^T = \Sigma_{\text{ext}}$. Hence, we have

$$\Sigma = W^{-1}DD^T(W^{-1})^T = W^{-1}\mathbf{L}\Sigma_{\text{ext}}\mathbf{L}(W^{-1})^T. \tag{19}$$

The linear Fisher information (FI) is thus given by

$$G^T\Sigma^{-1}G = -\Phi^T\mathbf{L}^T(W^{-1})^T(W^T\mathbf{L}^{-1}\Sigma_{\text{ext}}^{-1}\mathbf{L}^{-1}W(-W^{-1}\mathbf{L}\Phi) \tag{20}$$

$$= \Phi^T\Sigma_{\text{ext}}^{-1}\Phi \tag{21}$$

since $\mathbf{L} = \mathbf{L}^T$.

2.5.3.2 FI_E derivation

We computed FI for the E population similarly by calculating $\text{FI}_E = G_E^T\Sigma_E^{-1}G_E$ as follows. Partitioning the input covariance matrix into a block structure such that $\Sigma_{\text{ext}} =$

$\begin{pmatrix} \Sigma^E_{\text{ext}} & \Sigma^C_{\text{ext}} \\ \Sigma^C_{\text{ext}} & \Sigma^I_{\text{ext}} \end{pmatrix}$ and applying equation 19 leads to

$$\Sigma_E = \Omega^{-1}\mathbf{L}_E[\Sigma^E_{\text{ext}} - \Sigma^C_{\text{ext}}\mathbf{L}_I((-\mathbf{J}_{EI})W_{II}^{-1})^T - (-\mathbf{J}_{EI})W_{II}^{-1}\mathbf{L}_I\Sigma^C_{\text{ext}} + \mathbf{J}_{EI}W_{II}^{-1}\mathbf{L}_I\Sigma^I_{\text{ext}}\mathbf{L}_I(\mathbf{J}_{EI}W_{II}^{-1})^T](\Omega^{-1}\mathbf{L}_E)^T \quad (22)$$

where $\Omega = W_{EE} - W_{EI}W_{II}^{-1}W_{IE}$ and we have made use of the fact that $W_{EI} = -\mathbf{L}_E\mathbf{J}_{EI}$. By a similar argument we have that

$$G_E = -\Omega^{-1}\mathbf{L}_E \begin{pmatrix} \mathbf{1} & -\mathbf{J}_{EI}W_{II}^{-1}\mathbf{L}_I \end{pmatrix} \begin{pmatrix} \Phi_E \\ \Phi_I \end{pmatrix}. \quad (23)$$

Plugging these two expressions into the FI_E equation at the beginning of this section and identifying the term $X = -\mathbf{J}_{EI}W_{II}^{-1}\mathbf{L}_I = -\mathbf{J}_{EI}(-1-\mathbf{L}_I\mathbf{J}_{II})^{-1}\mathbf{L}_I$ gives the final equation 13.

2.5.4 Subpopulation codes in general: FI_α

The derivation shown above in the section "FI_E derivation" is flexible to shifts in the block structure. Thus for any subpopulation α of the full network the same arguments apply for a partitioning of the inputs with the mappings $E \to \alpha$ and $I \to U$ where the unobserved network elements are denoted by U.

2.5.4.1 Derivation of X for divergent E populations

Here we derive X for FI_E and FI_{E_1} in the two recurrently connected E populations and single I population (Figure 7). From equation 13 we have that, for FI_E, $X = \begin{pmatrix} J_{E_1I} \\ J_{E_2I} \end{pmatrix}(L_I^{-1} + J_{II})^{-1} = \begin{pmatrix} x_{E_1I} \\ x_{E_2I} \end{pmatrix}$ where, since L_I and J_{II} are scalars, $x_{E_1I} = \frac{L_I J_{E_1I}}{1+L_I J_{II}}$ and $x_{E_2I} = \frac{L_I J_{E_2I}}{1+L_I J_{II}}$.

We now derive FI_{E_1} after the preceding section. Let $U = \{E_2, I\}$. Then for E_1 readout, $X = J_{E_1U}(W_{UU})^{-1}L_U = \begin{pmatrix} x_{E_1E_2} & x_{E_1I} \end{pmatrix}$ where

$$x_{E_1E_2} = \frac{(J_{E_1E_2}(-1 - L_I J_{II})L_{E_2} - J_{E_1I}w_{IE_2}L_{E_2})}{det(W_{UU})} = \frac{L_{E_2}J_{E_1E_2}w_{II} - L_I J_{E_1I}L_{E_2}J_{IE_2}}{w_{E_2E_2}w_{II} - w_{E_2I}w_{IE_2}} \quad (24)$$

	U	M	M'
J_{EE}	0.975	0.769	4.74
J_{IE}	0.25	0.018	0.416
J_{EI}	0.25	0.018	0.208
J_{II}	0	0.206	0
L_E	5/8	5/8	1/5
L_I	8/5	8/5	5

Table 1: Parametric solutions to Figure 4

and

$$x_{E_1I} = \frac{(-J_{E_1E_2}w_{E_2I}L_I + J_{E_1I}(-1 + L_{E_2}J_{E_2E_2})L_I)}{det(W_{UU})} = \frac{-L_{E_2}J_{E_1E_2}L_IJ_{E_2I} + L_IJ_{E_1I}w_{E_2E_2}}{w_{E_2E_2}w_{II} - w_{E_2I}w_{IE_2}}. \tag{25}$$

Here we have written $w_{\alpha\beta}$ to denote the individual elements of the W matrix for this network, defined in equation 15 (for analysis of this network see: Supplemental Material). In particular, $w_{\alpha\beta} = -\delta_{\alpha\beta} + L_\alpha J_{\alpha\beta}$ where δ is the Kronecker delta function.

2.5.5 Model parameters

2.5.5.1 E/I network (Figure 4)

The response distributions were given by $r = -W^{-1}\mathbf{L}I$ and $\Sigma = (W^{-1}\mathbf{L})\Sigma_{\text{ext}}(W^{-1}\mathbf{L})^T$. Here we used the W notation defined along with equation 15. Letting $A = -W^{-1}\mathbf{L}$ with superscripts to denote the relevant figure panel in which the parameter set was used, we have: $A^U = \begin{pmatrix} .5 & -.8 \\ .8 & .5 \end{pmatrix}$, $A^M = \begin{pmatrix} .94 & -.01 \\ .01 & .94 \end{pmatrix}$, $A^{M'} = \begin{pmatrix} .4 & -.4 \\ .8 & .5 \end{pmatrix}$. Input drive $I = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$; $\delta I = \begin{pmatrix} 2 \\ 0.5 \end{pmatrix}$; input covariance $\Sigma_{\text{ext}} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. Additionally, L's and J's which also solve the equations are given in Table 1.

2.5.5.2 Ring network (Figures 3 and 6)

The ring model consisted of $N_E = 180$ E units and $N_I = 180$ I units where location on the ring corresponded to a unit's preferred tuning θ. Parameters were derived in part from Rubin *et al.* [118].

For all units the transfer function is threshold quadratic: $f = k\lfloor I_0 \rfloor_+^m$. Here $k = 0.04$ and $m = 2$. The stimulus is a wrapped gaussian with amplitude c_α: $s_{\text{ext}}(\theta) = c_\alpha g(\theta; \theta_0, \sigma_{\text{ext}})$ where

$$g(\theta; \theta_0, \sigma_{\text{ext}}) = k_{\text{ext}} \sum_{n=-\infty}^{\infty} \exp\left(\frac{-(\theta - \theta_0 + Nn)^2}{2\sigma_{\text{ext}}^2}\right). \tag{26}$$

The baseline input $b = 10, k_{\text{ext}} = 40$ and the stimulus was centered at $\theta_0 = 90^o$. Modulation was introduced as a uniform current bias $M(\theta) = \mu$ so that $I = I_0 + \mu_\alpha$ with $\mu_E = -1, \mu_I = 2$ (Figure 3), $\mu_E = -2, \mu_I = 4$ (Figure 9). Modulation in Figure 6 was modeled as a change in connectivity weight (see Table 2).

Connectivity decayed with distance, given by $g(\theta; \theta_\alpha, \sigma_{\alpha\beta})$ with $\sigma_{\alpha\beta} = 32^o$ for all $\alpha, \beta \in \{E, I\}$. Note in the figures, angles were expressed in radians.

The input covariances partitioned as in equation 22 were given by: $\Sigma_{\text{ext}}^\gamma = c_\gamma g(0, \sigma_\gamma), \gamma \in \{E, I, C\}$ (see Table 2).

2.5.5.3 E_1/E_2 network (Figures 7, 8)

We used equation 13 to generate the plots of FI_E vs. x. Input variance was 0.003025 while input covariance was 0.00242. To compute the stability boundaries in Figures 7 and 8, unless varied explicitly in the Figure, weights were set to $w_{E_1E_1} = w_{E_2E_2} = 0.5, w_{II} = -2, w_{E_2E_1} = 1, w_{IE_1} = w_{IE_2} = -w_{E_2I} = 1.4$ ($w_{\alpha\beta}$ notation defined in Methods: Derivation of X for divergent E populations).

2.6 Supplemental Material

Here we provide detail of the analyses of divergent excitatory pathways (Figure 7, 8) including stability and accessibility bounds, a discussion of differential correlations in our

	Value	Units	Figure
τ_E	20	msec	$3, 6$
τ_I	10	msec	$3, 6$
τ_n	1	msec	$3, 6$
J_{EI}	0.023		3
	0.03		6^M
	0.043		6
J_{EE}	0.044		3
	0.027		6
J_{IE}	0.042		3
	0.03		6
J_{II}	0.018		3
	0.042		6
σ_E	1		
σ_I	1		
σ_{ext}	30^o		
	40^o		6
σ_E	30^o		
c_E	0.1		
σ_I	20^o		
c_I	0.5		
σ_C	30^o		
c_C	0.2		

Table 2: Superscript M in the Figure value denotes the value to which a parameter changed with modulation.

model [97, 69] and a brief description of the general decomposition of X into paths through the network (Figure 7).

2.6.1 Impact of low-rank variability on modulation of subpopulation codes

Numerous experimental studies have demonstrated that the covariance structure in cortical circuits contains a significant low-rank component [63]. We therefore consider a covariance matrix which decomposes into full and low-rank components $\Sigma = \Sigma_0 + \epsilon v v^T$. Here Σ_0 is full-rank and v is an N-dimensional vector. We denote the elements of v which act on the E population by the vector v_E. Computing FI_E for this covariance structure results in (see: Supplemental, FI_E for low-rank covariance)

$$\mathrm{FI}_E = I_0\big(1 - \frac{\epsilon}{1 + \epsilon\langle v_E, v_E\rangle}(|v_E|\cos(\psi_{\Phi,v}))^2\big) \tag{27}$$

where $\psi_{\Phi,v}$ denotes the angle between Φ and v. We have defined $\langle x, y\rangle = x^T\Sigma_0^{-1}y$ and $I_0 = \Phi^T(\Sigma_0)\Phi$. Thus we see two potential (nontrivial) mechanisms for enhancing information flow: through changes in I_0 or by reducing $\cos(\psi_{\Phi,v})$. As the numerical results explored yielded either parallel or orthogonal vectors, we focus here only on the first mechanism. In particular, let v be the simplest case: the unit vector (Figure 9A, top). A modulation in the form of a constant input bias to all cells in the network model described in the previous section reduces correlations for nearby units while slightly increasing correlations for dissimilarly tuned units (Figure 9B, solid lines, green to magenta). This modulation results in a linear increase in FI_E as a function of I_0 (Figure 9C, solid line, green to magenta).

Equation 27 allows us to additionally consider an important case for v. It has been shown that the nature of correlated variability which limits information in a neural network is that which causes a shift in the population's response in the direction of the encoded variable [97]. These are called differential correlations as for a population response $\mathbf{r}$ they take the form of $\mathbf{r}' = \frac{d\mathbf{r}}{d\theta}$ (other studies have often used the notation f instead of $\mathbf{r}$ [97]). To this end consider a network in the presence of differential correlations such that $v = \Phi$ and equation 27 simplifies to

$$\mathrm{FI}_E = \frac{I_0}{1 + \epsilon I_0}. \tag{28}$$

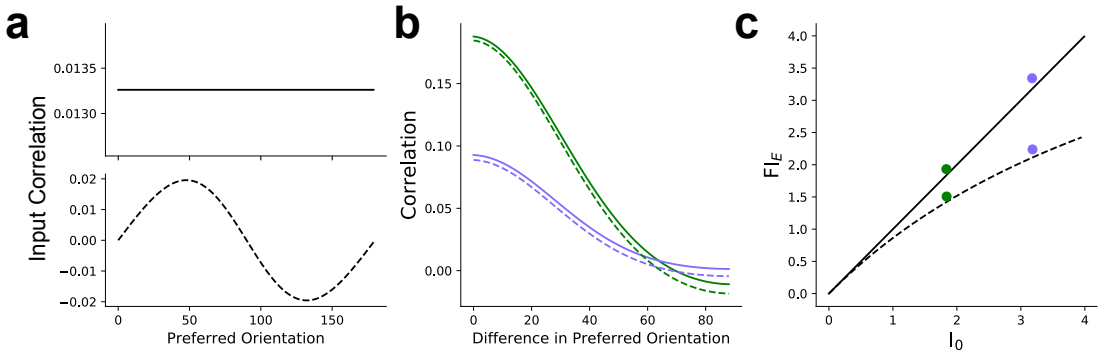

Figure 9: *Differential correlations in subpopulation codes* **a**. Input correlation structures. **b**. Output correlations for two different structures of input variability. Green: unmodulated; purple: modulated. **c**. Corresponding changes in FI_E for correlations in (b). Dashed line: differential correlations; solid line: non-differential correlations. ϵ in equation 28 was 0.3.

Thus, as $I_0 \to \infty$, information in subpopulation codes saturates to $\frac{1}{\epsilon}$ (compare curves in Figure 9C). To see how differential correlations affect modulation of subpopulation codes we introduced $v_\alpha = \Phi_\alpha$ into our model network (Figure 9A, bottom). The same modulation applied to this network resulted in an almost identical change in correlations (Figure 9B, dashed lines, green to magenta). However despite the similarities in the output correlation structures (Figure 9B), the E-population readout, as well as the change in the readout, differs significantly (Figure 9C, dashed line, green to magenta). Thus it is not enough to know how the firing rate statistics of a network change with modulation. Instead, one must understand the nature of the output projections of the network in question, together with its inputs, to get the full picture of how modulation is affecting information flow.

In summary, subpopulation codes can be modulated in the presence of information-limiting correlations as long as the network has not saturated the $\frac{1}{\epsilon}$ bound. Away from saturation, the results of our previous analyses for population codes hold. This is because I_0 is still subject to the modulatory effects described above.

2.6.2 FI_E for low-rank covariance

Here we specify a low-rank component of the input covariance such that $\Sigma_{\text{ext}} = \Sigma_0 + vv^T$ and again denote v_α the elements of v which act on population α. Then

$$\begin{aligned}
\mathrm{FI}_E =& \phi^T[\Sigma_0^E - \Sigma_0^C X^T - X\Sigma_0^C + X\Sigma_0^I X^T \\
&+ \epsilon(v_E v_E^T - v_E v_I^T X^T - X v_I v_E^T + X v_I v_I^T X^T)]^{-1}\phi \qquad (29)\\
=& \phi^T[\Sigma_0 + \epsilon(v_E - Xv_I)(v_E - Xv_I)^T]^{-1}\phi \qquad (30)\\
=& \phi^T\Sigma_0^{-1}\phi - \frac{\epsilon}{1+\epsilon\psi^T\Sigma_0^{-1}\psi}(\psi^T\Sigma_0^{-1}\phi)^2 \qquad (31)\\
=& \langle\phi,\phi\rangle - \frac{\epsilon}{1+\epsilon\langle\psi,\psi\rangle}\langle\psi,\phi\rangle^2 \qquad (32)\\
=& I_0 - \frac{\epsilon}{1+\epsilon\langle\psi,\psi\rangle}(|\phi||\psi|\cos(\theta_{\phi,\psi}))^2 \qquad (33)
\end{aligned}$$

where $\psi = v_E - Xv_I$ and we denoted $\phi = \Phi_E - X\Phi_I$. As above we let $\langle x, y\rangle = x^T\Sigma_0^{-1}y$ and $I_0 = \phi^T\Sigma_0^{-1}\phi$.

With the substitution $v_\alpha = \Phi_\alpha$ we have in the last equation above that $\psi = \phi$ and $\cos(\theta_{\phi,\phi}) = 1$ which results in equation 28.

2.6.3 Analysis of Divergent Excitatory Pathways

Here we expand our discussion of the $E_1 - E_2 - I$ network described in Figure 7. In particular we outline the stability conditions analyzed and limitations on attainable values of x_1 and x_2. We remind the reader of the system dynamics:

$$\delta\dot{\mathbf{r}} = \underbrace{\begin{pmatrix} -1 + L_{E_1}J_{E_1E_1} & L_{E_1}J_{E_1E_2} & -L_{E_1}J_{E_1I} \\ L_{E_2}J_{E_2E_1} & -1 + L_{E_2}J_{E_2E_2} & -L_{E_2}J_{E_2I} \\ L_I J_{E_1E_1} & L_I J_{E_1E_2} & -1 - L_I J_{II} \end{pmatrix}}_{W=-1+LJ} \delta\mathbf{r} + LI. \qquad (34)$$

X for a linear decoder restricted to population E_1 then has components

$$x_{E_1E_2} = \frac{L_{E_2}J_{E_1E_2}w_{II} + L_I J_{E_1I}L_{E_2}J_{IE_2}}{w_{E_2E_2}w_{II} + w_{E_2I}w_{IE_2}}, \qquad (35)$$

$$x_{E_1I} = \frac{L_{E_2}J_{E_1E_2}L_I J_{E_2I} - L_I J_{E_1I}w_{E_2E_2}}{w_{E_2E_2}w_{II} + w_{E_2I}w_{IE_2}}. \qquad (36)$$

	Value	Units
τ_E	20	msec
τ_I	10	msec
τ_n	1	msec
J_{EI}	0.03	
J_{EE}	0.03	
J_{IE}	0.03	
J_{II}	0.029	
σ_E^C	0.515	
σ_E^D	0.5	
σ_I^C	0.515	
σ_I^D	0.5	
σ_{ext}	30^o	
σ_E	30^o	
c_E	0.1	
σ_I	20^o	
c_I	0.5	
σ_C	30^o	
c_C	0.2	

Table 3: Figure 9 Parameters. Variable superscript C denotes constant modulation case, D denotes differential correlation case.

where $w_{\alpha\beta}$ is the element of W given by $w_{\alpha\beta} = -\delta_{\alpha\beta} + L_\alpha J_{\alpha\beta}$ with δ the Kronecker delta function and $\alpha, \beta \in \{E_1, E_2, I\}$. Since we have assumed our system is in a steady-state and linearizable regime we can apply the Routh-Hurwitz stability criteria to our linearized network (see Appendix A.2.1), which requires three conditions be satisfied (note we have not assumed a sign for any of the $w_{\alpha\beta}$'s):

1. $a_2 = -w_{E_1E_1} - w_{E_2E_2} - w_{II} > 0$
2. $a_0 = w_{E_1E_1}(w_{E_2I}w_{IE_2} - w_{E_2E_2}w_{II}) + w_{E_1E_2}(w_{E_2E_1}w_{II} - w_{E_2I}w_{IE_1}) + w_{E_1I}(w_{E_2E_2}w_{IE_1} - w_{E_2E_1}w_{IE_2}) > 0$
3. $a_2[w_{E_1E_1}(w_{E_2E_2} + w_{II}) + w_{E_2E_2}w_{II} - (w_{E_1E_2}w_{E_2E_1} + w_{E_1I}w_{IE_1} + w_{E_2I}w_{IE_2})] > a_0$.

Next we prove bounds on x_1 and x_2 for given parameter sets. We will use the following shorthand: $x_E = L_{E_2}J_{E_1E_2}/w_{E_2E_2}, x_I = L_IJ_{E_1I}/w_{II}$ and $x_{\beta\alpha} = L_\alpha J_{\beta\alpha}/w_{\alpha\alpha}, \alpha, \beta \in \{E_2, I\}$ (i.e. the x's written for the corresponding monosynaptic connections). Finally, to make the notation clearer we write $x_{E_1E_2} = x_1$ and $x_{E_1I} = x_2$. First we prove the general claim that x_1, x_2 cannot both be negative. More precisely, $x_1 < 0 \implies x_2 \geq 0$ and $x_2 < 0 \implies x_1 \geq 0$. Rearranging terms in equations 35 and 36 we can write:

$$x_{E_1E_2} = x_1 = \frac{x_E - x_I x_{IE_2}}{1 - x_{E_2I}x_{IE_2}},$$
$$x_{E_1I} = x_2 = \frac{x_I - x_E x_{E_2I}}{1 - x_{E_2I}x_{IE_2}}.$$

Note that $x_I \geq 0$ since $x_I = \frac{-L_IJ_{E_1I}}{-1-L_IJ_{II}} = \frac{L_IJ_{E_1I}}{1+L_IJ_{II}} \geq 0$, and similarly for x_{E_2I}. By contrast, x_E is unrestricted.

We will consider two cases. For readability we write $x_{E_2I} = a, x_{IE_2} = b$. In this way

$$x_1 = \frac{x_E - x_I b}{1 - ab}, x_2 = \frac{x_I - x_E a}{1 - ab}.$$

Case 1: $1 - ab > 0$. First let $b \geq 0$. Then

$$x_1 < 0 \implies x_E < x_I b \implies x_E a < x_I ab < x_I \implies x_2 > 0$$
$$x_2 < 0 \implies x_I < x_E a \implies x_I b < x_E ab < x_E \implies x_1 > 0$$

where the last inequality follows from the assumption that $1 > ab$. Now suppose $b < 0$. But then $x_1 \propto x_E + x_I|b| < 0 \iff x_E < 0 \implies x_2 \propto x_I - x_E a = x_I + |x_E|a > 0$. Clearly $x_2 < 0$ requires $x_E > 0$ so this condition is immediate.

Case 2: $1 - ab < 0$. Note that this assumption is true iff $a, b > 0$ since $a \geq 0$.

$$x_1 < 0 \implies x_E > x_I b \implies x_E a > x_I ab > x_I \implies x_2 > 0$$
$$x_2 < 0 \implies x_I > x_E a \implies x_I b > x_E ab > x_E \implies x_1 > 0$$

where now the last inequality follows from the assumption that $1 < ab$.

The above arguments illustrate parameter-independent bounds on x_1 and x_2 for this particular system. However there are also parameter-dependent bounds on x_1 and x_2 (Figure 8, variability in the white regions). The main text focused on bounds due to restrictions on the connection (J) values; these arise from the fact that excitation and inhibition necessarily have positive and negative values, respectively. Equations 35 and 36 share the same denominator, which we denote d. In each panel in Figure 7 we varied only $J_{E_1 I}$ and $J_{E_1 E_2}$ (see also Figure 10). Thus equations 35 and 36 can be written in the form:

$$x_1 = \frac{L_{E_2} w_{II}}{d} J_{E_1 E_2} + \frac{L_{E_2} L_I J_{I E_2}}{d} J_{E_1 I},$$
$$x_2 = \frac{L_{E_2} L_I J_{E_2 I}}{d} J_{E_1 E_2} - \frac{L_I w_{E_2 E_2}}{d} J_{E_1 I}.$$

For sake of illustration assume that $d > 0$ (the same argument holds under reversal of signs). Then for $J_{E_1 E_2} = 0, x_1 \geq 0, x_2 \leq 0$ and $x_2 = -\frac{L_I w_{E_2 E_2}}{L_{E_2} w_{I E_2}} x_1$, which is a lower bound for x_2 in the region where $x_1 > 0$. Notice with our assumption that d is positive, a similar argument letting $J_{E_1 I} = 0$ leads only to bounds on $x_1 \leq 0, x_2 \geq 0$ (since $w_{II} \leq 0$). Thus we see how the $E_2 - I$ connectivity places additional constraints on the accessible values of X under modulation of connection strength (observe the white regions in Figure 8 for $x_1 \geq 0, x_2 \leq 0$ and $x_1 \leq 0, x_2 \geq 0$).

Different bounds will result from varying different parameters. For example, since the cellular gain, L_α, is non-negative its minimum value is 0. Letting $L_I = L_{E_2} = \epsilon$ in equations

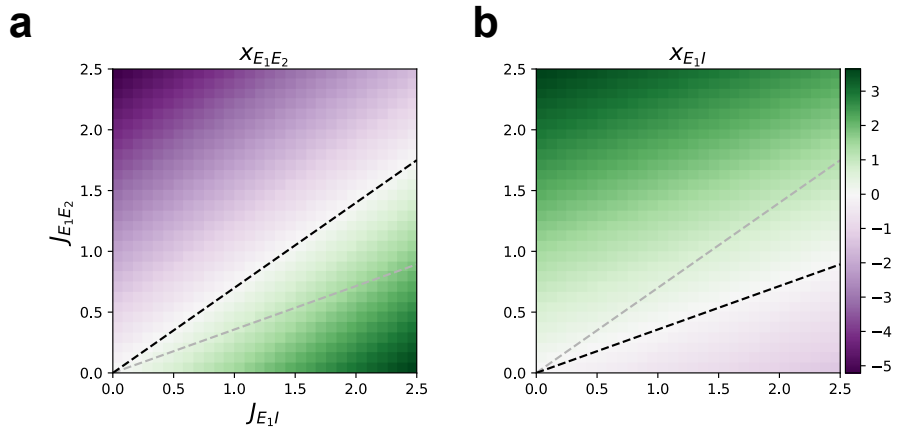

Figure 10: *X components as a function of connectivity.* **a**. $x_{E_1E_2}$ as a function of $J_{E_1E_2}, J_{E_1I}$. Black dashed line corresponds to 0 contour; gray dashed line is 0 contour of panel (b). **b**. Same as (a) for x_{E_1I}; gray dashed line is now 0 contour of panel (a). Region bounded by dashed lines is where $x_{E_1E_2}, x_{E_1I}$ are jointly positive.

35 and 36 it can be shown that $x_1 < 0, \delta > x_2 \geq 0$ for some δ which depends on ϵ in the parameter regime used in Figure 7. In particular, the assumption of L_I, L_{E_2} small leads to the terms in equations 35 and 36 with w's dominating. These values would in turn cross the J-induced boundary, but do not violate our conclusions since L_α was assumed fixed in Figure 7.

2.6.4 X as paths through the network

As discussed above, in full generality $X = \mathbf{J}_{OU} W_{UU}^{-1} \mathbf{L}_U$ where O is the observed (readout) population, U the unobserved population. We recall that $W_{UU} = -\mathbf{1} + \mathbf{L}_U \mathbf{J}_{UU}$. Assuming boundedness of W_{UU} we have the expansion $-(\mathbf{1} - \mathbf{L}_U \mathbf{J}_{UU}) = -\sum_{k=0}^{\infty} (\mathbf{L}_U \mathbf{J}_{UU})^k \implies X = -\mathbf{J}_{OU} \sum_{k=0}^{\infty} (\mathbf{L}_U \mathbf{J}_{UU})^k \mathbf{L}_U$. Since $\mathbf{L}_U \mathbf{J}_{UU}$ are the effective connection strengths in the network, $(\mathbf{L}_U \mathbf{J}_{UU})^k$ denotes the k^{th} step through the unobserved population [128], and consequently X is comprised of all paths through the unobserved population and their projections to the observed population ($\mathbf{J}_{OU}$). This expansion has been used before to decompose the structure

of the response covariance in a neural network [145, 13].

The final panel in Figure 7 additionally derives from this expansion in the following way: consider the ij element of X where i is the index of a decoded unit and j is an undecoded unit. Then $X_{ij} = L_j \sum_k J_{ik} W^{-1}_{kj} = L_j \sum_k J_{ik}(-\sum_p (LJ)^p_{kj})$. In the final sum the kj^{th} element is all the p-step ways to reach element k from element j, and in turn, the readout neuron i.

3.0 Constraints on mechanistic models of attention

3.1 Overview

Attentional modulation of neural activity in visual cortex has been shown to relate to cell- or network-intrinsic properties through the phenomenon of response normalization [101]. In this study we seek a mechanistic description for this relationship. We focus on the correlation between neuron-to-neuron heterogeneity in attentional effects and across-neuron heterogeneity in normalization. Interestingly, we find that, despite being a conspicuous relationship between a cognitive variable (attention) and a circuit-dynamic variable (normalization), it nevertheless is insufficient to adequately constrain circuit models. We instead find that by considering correlations between the heterogeneities in absolute changes in rate with attention and normalization better constrain mechanistic models of attention. In particular, these constraints point to the necessity of strongly recurrent networks in constructing these relationships. We then demonstrate network properties which support this collection of correlated heterogeneities, showing how these depend on the structure of inhibition. This work aims to utilize these known relationships between circuit dynamics (normalization) and a cognitive process (attention) to constrain possible models of attentional effects on circuit activity and the dynamical state of cortex.

3.2 Introduction

Experimental studies of nonhuman primates performing a visual task in which their attention is differentially allocated have provided a wealth of insight into attentional processes. Intracortical electrode recordings in primate visual areas MT and V4 during these tasks have identified changes in neural activity which relate to different attentional states [90]. These changes depend in a nuanced way on the type and spatial organization of the stimuli used, and the nature of attentional cueing [25]. When attention is directed to a spatial location

within a recorded neuron's receptive field, the neural response gain to gabor stimuli presented at that location tends to increase tuning curves in a contrast-invariant manner [91]. When stimuli are presented alone, this results in either a shift or rescaling in the contrast-response function of the neuron [152]. If multiple or overlapping stimuli are presented, however, the neural response becomes a nonlinear function of the inputs. The nature of this nonlinearity has been shown to relate to a property called response normalization.

Cross-orientation suppression provides a prominent example of normalization. A neuron driven by a single gabor at its preferred orientation responds maximally while an orthogonal gabor might elicit very little response from the neuron. Superimposing both (or, presenting both stimuli within the neuron's receptive field at spatially distinct locations) often causes a reduction in firing rate [119]. Cross-orientation suppression can be captured using a descriptive model in which the rates are fit to an equation of the form

$$r = \frac{w_1 c_1 + w_2 c_2}{c_1 + c_2 + \sigma} \tag{37}$$

where c_i is the contrast and w_i weights the relative contribution of the i^{th} stimulus, and σ is a constant which controls the contrast saturation rate [19]. Response normalization of this and related forms is highly successful at fitting responses to visual cortical cells [59, 20].

The observation that attention affects contrast response functions which are well-described by the normalization model led to the suggestion that attention and normalization are related [115]. This idea can be implemented by simply including a term α in equation 37 to differentially weight the attended stimulus; for example attending to the first stimulus can be written $r = (\alpha w_1 c_1 + w_2 c_2)/(\alpha c_1 + c_2 + \sigma)$. Indeed, this extension captures well activity recorded in both MT and V4 in attended and unattended conditions [77, 101, 149]. Modifications to this fundamental descriptive model have been able to explain further attentional effects by tweaking the component parts as a function of attention or stimulus conditions, speaking to the robust generality of this model [100, 122]. While this flexibility is a boon to fitting experimental data, it can limit the model's effectiveness at constraining mechanistic descriptions of attention. Namely, this modeling framework falls short of describing how attention arises from, or interacts with, the fundamental building blocks of circuits: neurons and synapses.

Models attempting to provide an understanding of the mechanism of action of attention in cortical circuits have generally incorporated a knowledge of cortical architecture and responsiveness. There is a rich history of cortical circuit models capable of capturing the neural dynamics of visual cortical areas [10, 146, 80]. Because spatial attention scales tuning curves in a contrast-invariant fashion, these recurrent network models have formed the foundation of attentional models recapitulating this result through modulation of gain parameters [26, 37, 82]. Yet these models also exploit a degree of flexibility in their implementation of attention, incorporating it however necessary to match the desired experimental observations. One potential remedy is akin to a brute-force search: match as many experimental observations as possible with a single model type. This approach has been taken recently in the service of explaining how the normalization model of attention described above could manifest through circuit-specific mechanisms [82]. Recurrent networks have been shown able to intrinsically induce normalization through strong inhibitory feedback together with strong recurrent excitation [5, 118]. This model was then extended to incorporate attention, demonstrating extensive agreement with many experimental observations [82]. However, it is possible that disparate model types could also adequately explain the data. This approach leaves open the question of what neural or circuit properties are required to match these observations.

In an attempt to address this more rigorously, we again focus on the normalization model of attention. A key observation from the studies of Lee and Maunsell [77] and Ni *et al.* [101] was that the effect of both normalization and attention is highly heterogeneous across cells, and positively correlated. This positive correlation implied that neurons strongly affected by normalization were also more strongly affected by attention. The success of the descriptive model (equation 37) in capturing both effects suggests that there is a shared underlying mechanism at either a cellular or circuit level which explains both their heterogeneity and correlation. In contrast to a brute-force approach, we instead start by asking what model constraints can be derived from this single observation that normalization and attention response heterogeneities are correlated.

To this end we begin with a phenomenological model of a neuron in which cellular heterogeneities of the input-output function alone are responsible for the observed correlation,

finding that the normalization-attention relationship is in fact not a good model constraint. Instead, we consider the observation that the absolute magnitude of change in rates with attention is also correlated with normalization, and argue this affords a better set of restrictions on the model. From this, we are able to show that under certain assumptions a saturating input-output function is necessary to capture the data. We take this as evidence that the normalization-attention relationship depends on the network structure since neural transfer functions are expansive in cortex, not saturating [109]. Finally, we consider how this new set of experimental data constrains a network model in which normalization emerges from the circuit structure [5, 118], showing that a model with synaptic scaling best reflects the data. By using correlated response heterogeneities to constrain models, this work thus establishes a circuit framework from which to develop a deeper understanding of the mechanism of action of attention in cortex.

3.3 Results

Experimental studies of attention and normalization have predominantly been conducted in nonhuman primates performing a visual change-detection task. Consistently, the motivation for this work is a key result from Ni et al. [101] in which attention was deployed to either a preferred or null stimulus while both were present in the receptive field (RF) of a visual neuron in area MT (Figure 11). Normalization was measured in this experiment by cueing the monkey to attend to a grating outside the recorded neuron's receptive field oriented between the preferred and null stimuli (Figure 11). These five stimulus conditions enable the defining of two metrics, a normalization modulation index (NMI) reflecting how much a neuron's rate is affected by the presence of a superimposed null stimulus, and an attentional modulation index (AMI), capturing the changing in firing rate associate with attention being cued to the preferred or null stimulus within the neuron's receptive field. Consistent with this study, we define the following:

$$\text{NMI} = \frac{r_p - r_x}{(r_p - r_n) + (r_x - r_n)} \tag{38}$$

where r_p, r_n, and r_x are the firing rates of a given neuron in response to its preferred (θ_p), null (θ_n), and both (θ_x) stimuli together, respectively; and

$$\text{AMI} = \frac{r_x^p - r_x^n}{r_x^p + r_x^n} \tag{39}$$

where r_x^p and r_x^n denote the response of a cell when both stimuli are present (subscript x) and attention was cued to either the preferred (superscript p) or null (superscript n) stimulus in the receptive field, respectively. While not entirely consistent with the literature, for simplicity of exposition we will use the term *plaid* to refer to the situation in which both stimuli are in the RF simultaneously. Strictly speaking use of this term implies overlapping stimuli, but the effect of a true plaid and spatially distinct stimuli presented within a single neuron's RF simultaneously are similar [113]. Additionally, it should be noted that some studies use slightly different metrics to measure normalization and attention; we consider any discrepancies in the Appendix.

3.3.1 The NMI-AMI relationship is insufficient to constrain a simple neural model

Neurons themselves are highly diverse, with phenotypic variation within and across classes [53]. It is therefore plausible that the reason NMI and AMI are correlated is rooted in a cell itself. Probing this possibility, we consider cellular heterogeneities in the context of a phenomenological rate model of a collection of single neural units. To make this precise, for neuron i let its trial-averaged firing rate r be described by

$$\dot{r_i} = -r_i + \phi(b + \mu_A + s_{ext}) \tag{40}$$

where ϕ is any monotonically increasing function, b is a baseline offset and s_{ext} a stimulus input. μ_A is an attentional bias with $A = n$ signifying attention to the null stimulus and $A = p$ denoting attending to the preferred stimulus. Since neural activity in experimental data is computed spanning stimulus presentation we focus on the steady-state solution (i.e., $\dot{r_i} = 0$) of equation 40 [101]. The phenomenological nature of this model allows some freedom to address what properties the model must contain to correlate NMI and AMI.

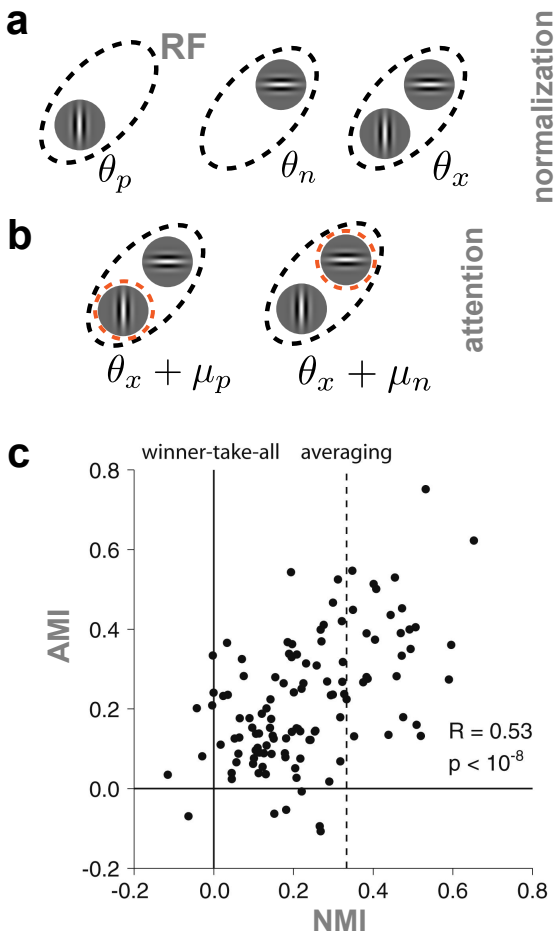

Figure 11: *Experimental measurements of normalization and attention* **a** Schematic illustration of stimulus organization relative to a recorded neuron's receptive field (RF; dashed ellipse). Gabors schematize stimulus positioning relative to the RF. θs indicate stimulus orientations describing a hypothetical neuron's preferred (θ_p) and null (θ_n), as well as the plaid organization (θ_x). **b** As in (a) for attention conditions. Dashed orange circles indicate the locus of attention cued on a given experiment, with μ indicating whether the neuron's preferred (μ_p) or null (μ_n) stimulus is attended. **c** Normalization-attention relationship from single-unit recordings in primate area MT. (c) Reprinted from Ni *et al.* [101] with permission from Elsevier.

We will therefore choose a simple yet flexible function class for $\phi : \phi(I) = k\lfloor I \rfloor_+^m$ where I represents the collective input. Additionally, the exact interpretation of s_{ext} is loose; at this juncture, neuron i may be considered as a sample of a random function from the class of functions defined by the foregoing equations. In this context, s_{ext} may capture feedforward- or recurrent-like connections.

The model needs to fulfill three requirements to capture the data (Figure 11): it must be heterogeneous in both NMI and AMI, and it must positively correlate the two indices. In what follows we will present a largely intuitive argument; mathematical justification can be found in the Methods. First, we must assume normalization is inherited since there is no other mechanism by which this model could induce it. While required in this context, inherited normalization in fact has precedent even in V1 [7]; higher order visual cortex also likely receives normalized inputs from earlier regions since V1 exhibits normalized responses.

Since we are interested in how cellular heterogeneities alone might induce the relationship, we make precise one central assumption: heterogeneities should exist only at the level of the single unit under consideration. If we assume that heterogeneities in attention and normalization are both inherited in this model then they must themselves be correlated; *a priori* this says that attentional processes have some knowledge of the normalization capacity of a cell. In the present context, we find this implausible. Letting the inputs s_p, s_n and s_x map to the respective rates r_p, r_n and r_x, this assumption implies that we are fixing an ordering $s_n < s_x < s_p$ such that $s_x - s_n$ and $s_p - s_x$ are constant for all cells (Figure 12). Similarly, the magnitude of the attentional effects $|\mu_n - s_x|, |\mu_p - s_x|$ are constant across cells.

In order to ease the narrative, we will suppose that the baseline b is the dominant source of heterogeneity across cells with each other parameter varying only slightly. The definition of NMI can be seen to describe a ratio of intervals, or distances, in rate space: $r_p - r_x, r_p - r_n$ and $r_x - r_n$. Consequently, if b is the only heterogeneous model parameter then for NMI to be heterogeneous ϕ cannot be linear ($m = 1$). This follows from the fact that a linear ϕ preserves distances from input to output; as the input distances are assumed constant across cells, this would result in a homogeneous NMI (Figure 12). By contrast, for a nonlinear ϕ ($m \neq 1$) input distances are scaled in a b-dependent manner, resulting in a heterogeneous NMI (Figure 12). For completeness, we consider the implications of m strongly heterogeneous elsewhere,

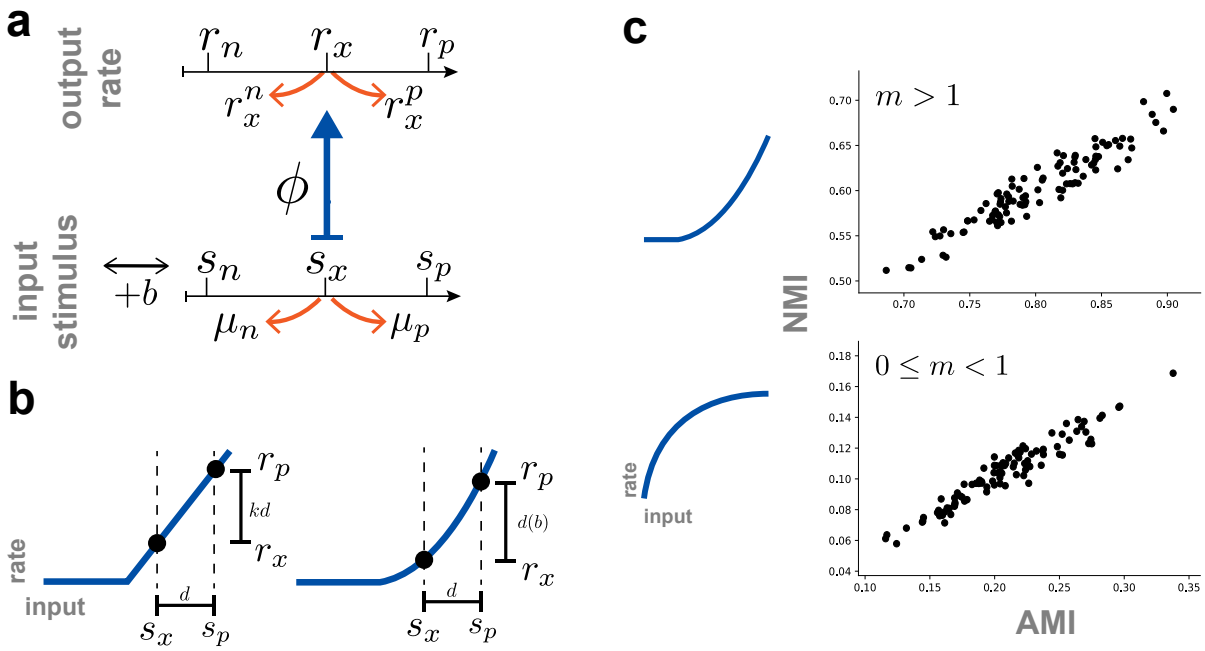

Figure 12: *Phenomenological model of normalization and attention* **a** Schematic illustration of setup. b acts as a left/right translation of the inputs s_{ext}. ϕ defines the mapping from inputs to rates. μ describes the attentional condition, implemented as an additive input bias. r_s^A reports the firing rate, where the subscript denotes the stimulus presentation and the superscript is the attentional target. **b** Intuitive illustration of a linear transfer function's insensitivity to heterogeneity in b (left) as the distances depend only on k and d. In contrast, a nonlinear transfer function differentially scales input distances (right, $d \mapsto d(b)$). This differential scaling enables nonlinear functions to induce heterogeneous responses and by extension, heterogeneities in normalization. **c** Model simulations illustrating that both expansive (top) and saturating (bottom) model regimes recapitulate the desired relationship. All model parameters were sampled from independent Gaussian distributions with $\text{Var}(b) >> \text{Var}(y)$ where $y \in \{m, s_x, s_p\}$. r_n was manually set to zero, in approximation of experimental data.

demonstrating that additional requirements are necessary (Supplemental material).

Ni *et al.* observed that in their experiment, when the subject was attending to the preferred stimulus of a plaid, firing rates of the recorded neuron increased to approximately match the rate when the preferred stimulus is presented alone, whereas attending to the null stimulus decreases rates relative to the plaid alone [101]. Consistent with this, in equation 40 we take $\mu_n < 0$ such that $s_n < \mu_n + s_x < s_x$ and $\mu_p > 0$ such that $\mu_p + s_x \approx s_p$. Following the logic above for heterogeneous b, we can see that in this case as well if ϕ is linear then AMI will be homogeneous whereas if ϕ is nonlinear then AMI will be heterogeneous. Therefore heterogeneous b and $m \neq 1$ together satisfy the first two requirements. By symmetry in the attentional and normalization effects, as well as in their indices, we further contend that NMI and AMI will be positively correlated. Numerical analysis confirms this statement for both $m > 1$ and $m < 1$ (Figure 12).

These results justify at least two conclusions. On the one hand, they show that it is possible to explain the NMI-AMI relationship with cellular heterogeneities alone. On the other, it illustrates the fact that two very different classes of model are capable of satisfying this particular correlated heterogeneity: a model with a saturating nonlinearity, and a model with an expansive nonlinearity. Because of this, the NMI-AMI relationship is in fact not a good constraint on our model.

3.3.2 Absolute changes in rates with attention better constrain a simple model

We next turned to a dataset recorded in visual area V4 with a monkey performing a similar task [150]. In this study a selectivity index was computed for each cell in addition to a normalization index. The selectivity index captured the extent to which a neuron's activity in response to its preferred stimulus differed from its response to the null. It was shown that for sufficiently high selectivity, a gradient of attentional effects exists as a function of normalization (Figure 13). In particular, the change in rates with attention to the preferred (respectively, null) stimulus increases (decreases) in absolute magnitude relative to the plaid alone for increasing normalization index (Figure 13). (This study used a slightly different metric for normalization; we show in the Appendix that it reflects a similar trend to the NMI

used here, hence we continue our analysis with NMI.) Mathematically, this says that the derivative of the input-output function ϕ should increase together with NMI (Figure 13ci). Since NMI decreases with increasing b (Methods), $d\phi/dI$ must decrease with increasing b (Figure 13cii). As this describes a saturating function, it implies $m < 1$ (Figure 13).

This result therefore shows that considering the absolute change in rates with attention provides a good constraint on our model, since it is able to distinguish between two functionally distinct input-output regimes. Interestingly, we find that a saturating response function best captures the observed data. In interpreting this result, we need to make concrete what ϕ represents. A neuron in cortex essentially has two distinct functions through which inputs map to outputs: its neuronal transfer function, which specifies how changes in input current are converted to firing rates, and its stimulus-response function, a more conceptual relationship between a continuous stimulus parameter and the firing rate it evokes. It is known from intracellular recordings that neural transfer functions are expansive whereas neural response functions saturate [109]. For this reason, we interpret our result that the input-output function in our simple model must saturate as evidence that the correlations between attentional effects and normalization are network-determined. We can therefore update our conclusion from the previous section to say that, in the absence of evidence to the contrary, cellular heterogeneities alone are insufficient to explain attention-normalization relationships. We next turn to a full circuit model to explore the implications of these refined experimental constraints.

3.3.3 A heterogeneous network model reveals synaptic constraints

There is evidence that the origin of normalization effects in cortical circuits may come either through feedforward [7] or through recurrent mechanisms [4]. Although not mutually exclusive, recent modeling work has shown that normalization emerges naturally in cortical models placed in an inhibition-stabilized regime (ISN) [5]. Furthermore, experimental evidence supports the claim that cortex is inhibition-stabilized [3, 125]. We use a network framework dubbed the stabilized supralinear network (SSN) [118]. In an SSN, neural transfer functions are expansive while excitatory rates saturate well below any refractory period

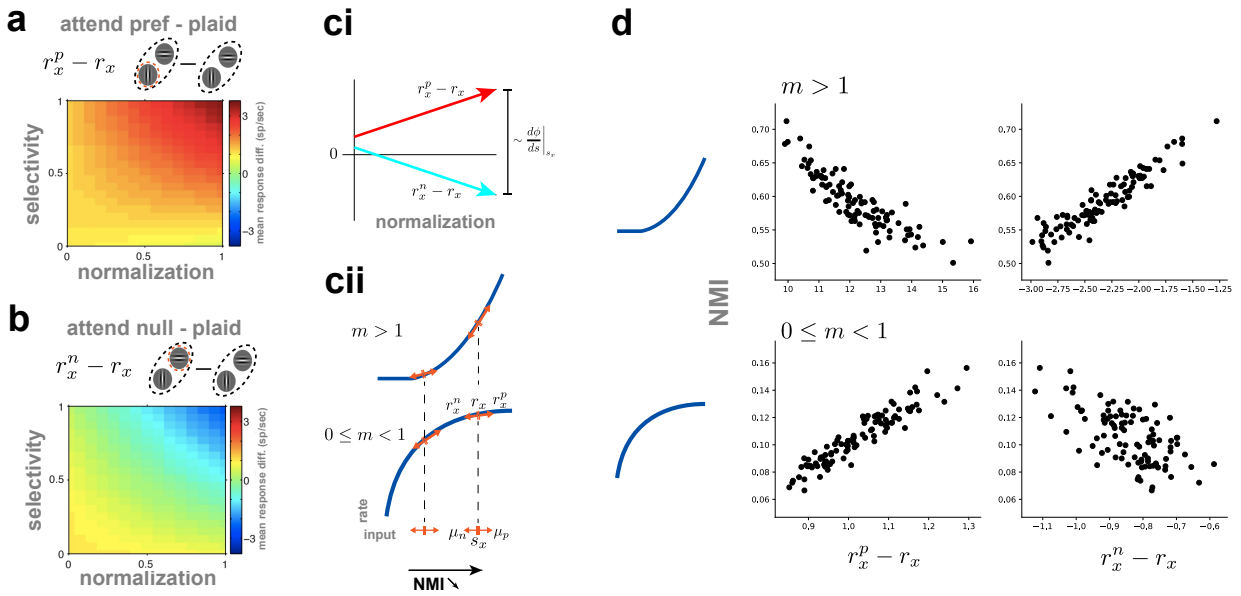

Figure 13: *Absolute changes in attentional firing rates provide a useful model constraint* **a, b** Relationship between a normalization index (Methods) and selectivity showing for sufficiently high selectivity, attending preferred increases rates more (a), and attending null decreases rates more (b) for stronger normalizing cells. **c** (i) Summary schematic of trends in (a), red line, and (b), blue line, collapsed along the selectivity axis. (ii) Schematic demonstrating the relationship between the derivative of ϕ at s_x and the change in rates with additive attention. **d** Model simulation results for the analogous criteria. Values sampled from same distributions as Figure 12. (a,b) Adapted by permission from Springer Nature: Nature Neuroscience *Attention-related changes in correlated neuronal activity arise from normalization mechanisms*, Verhoef and Maunsell (2017) [150].

effects, consistent with both our derived requirement, as well as experimental observations [118]. In this framework connections are sufficiently strong that as the amplitude of feedforward inputs increases, network activity transitions from a feedforward-driven non-ISN regime to one dominated by recurrent activity. This latter regime depends on inhibition stabilization, which induces saturation in the excitatory response [147]. Thus, this model can be read as testing the hypothesis that heterogeneities in circuit architecture account for the normalization-attention relationships.

We therefore consider a circuit model in which N excitatory (E)/inhibitory (I) pairs are organized around a ring (Figure 14). The location of an E/I pair corresponds to their preferred stimulus orientation θ of a visual grating. The dynamics of a single unit on the ring are given by

$$\tau \dot{r_i} = -r_i + f(\mathbb{J}_i r + \mu_\alpha(\theta_A) + s_{ext}(\theta_j)) \tag{41}$$

Stimuli are modeled as a wrapped Gaussian $cg(\theta, \sigma)$ centered at orientation θ_j with width σ. Attention μ_α is modeled as an excitatory bias to E cells and is also a wrapped Gaussian centered at orientation θ_A with $\alpha \in \{E, I\}$. θ_A matches the stimulus being attended ($\sigma_A = \sigma$), with amplitude c_A generally much less than c (Figure 14). Following previous work we take $f(I) = k\lfloor I \rfloor_+^2$ [118].

Synaptic connections follow a Gaussian profile as a function of position θ and weights are chosen to place the network in an ISN [118]. In order to induce heterogeneities in the responses there are in principle a vast number of ways to do so. We chose to sample certain connection weights from a random function thereby testing the hypothesis that the distribution of synaptic inputs is sufficient to explain the data. Additionally, this choice tacitly assumes that relative to cellular heterogeneities, the variability in a neuron's inputs is much larger and thereby determines a neuron's stimulus response properties to a greater degree. Procedurally, we independently sample each $I \to E$ connection ($\mathbb{J}_{EI}$) from a Gaussian distribution with mean centered on the nonrandom connection profile, and threshold at zero such that weights cannot change sign (Figure 14).

NMI and AMI are computed following equations 38, 39 and in a manner consistent with the experimental setup. For example, to calculate a neuron j's response in the attend preferred condition, the plaid stimulus $s_{ext}(\theta) = c(g(\theta_j, \sigma_{ext}) + g(\theta_{j+N/2}, \sigma_{ext}))$ and the atten-

tional bias $\mu_A(\theta_A, \sigma_A)$ are presented simultaneously and neuron j's firing rate r_x^p is recorded. We see that in this network again, NMI and AMI are both heterogeneous and positively correlated (Figure 14). More importantly, we see that the additional constraints are met in this network (Figure 14): for increasing NMI, the increase (respectively, decrease) in rates with attention shows an upward (respectively, downward) trend. Therefore, a sufficient condition to meet the experimental criteria is for inhibitory inputs to be heterogeneous across stimulus orientations.

3.3.4 Synaptic scaling is required for robust agreement with constraints

A naive explanation of normalization in the context of a plaid stimulus is that the orthogonal stimulus in recruiting a source of inhibition that was not driven by the preferred stimulus alone. Neurons which do not normalize would simply not have this same pool of orthogonally tuned inhibitory sources. The results of the preceding section suggest that this could be a plausible explanation. By targeting the excitatory units in the network, the attentional bias we have chosen would serve to further amplify these effects. Attention to the null stimulus would drive further inhibitory recruitment, whereas attending to the preferred would override the additional suppression. We probe these ideas further by reconsidering the nature of the synaptic heterogeneity.

We start by testing the minimal hypothesis that heterogeneity in orthogonal sources of inhibition is sufficient to recapture the data. To do this, we randomly sample the weights $\mathbb{J}_{EI}$ with a modified mean. For each neuron i, the mean is then the sum of the original distance-dependent connectivity profile, plus an additional orthogonal component: $\mathbb{J}_{E_iI} = J_{EI}g(\theta_i, \sigma_{EI}) + \delta g(\theta_i + N/2, \sigma_{ortho})$ with $\sigma_{ortho} < \sigma_{EI}$ (Figure 15). Here δ itself is a random variable scaling the amplitude of the orthogonal component. Each weight is again sampled from a Gaussian centered at the given mean. As a sanity check, this sampling procedure should not induce a correlation in the weights which are proximal to a given neuron and those which are mostly orthogonal. We therefore define values $\mathbb{J}_{E_{\theta_{pref}}I_{\theta_{pref}}}$ and $\mathbb{J}_{E_{\theta_{pref}}I_{\theta_{ortho}}}$ for each neuron to summarize the total amount of inhibition local to an E neuron, and orthogonal to an E neuron (Methods). In the present case, there is indeed no relationship

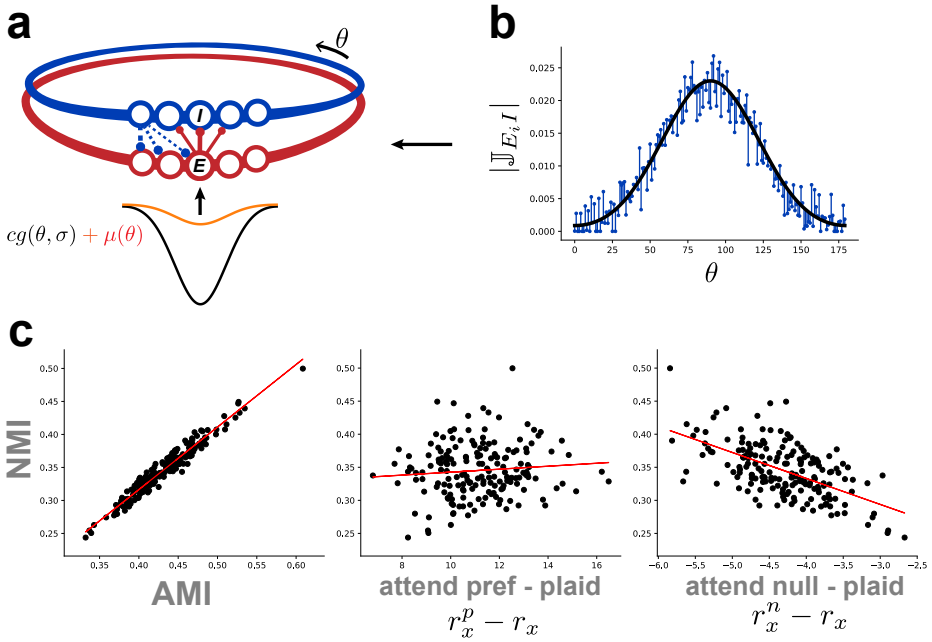

Figure 14: *Network model of normalization and attention* **a** Schematic illustration of ring network setup. Feedforward inputs are illustrated in black, given by $cg(\theta)$ and attentional inputs are schematized in orange ($\mu(\theta)$). Dashed connection lines identify which weights were randomly sampled. **b** Illustration of the sampled inhibitory inputs to neuron E_i. Black line is mean and dots correspond to sampled values. **c** Normalization and attention indices computed from the steady state solutions of the network firing rate recapture the right trends in both NMI-AMI and between normalization and absolute rate changes with attention. Red lines are linear fits to data.

between these two values (Figure 15).

When we calculate NMI and AMI in this minimal case, we see that as before, NMI and AMI are again strongly positively correlated (Figure 15). This would seem to match our intuitive argument. However, when we consider the relationship between NMI and the change in rates with attention to the preferred stimulus, this network produces the wrong relationship, inducing a weak negative correlation (Figure 15). This follows from the results of our sanity check - attending to the preferred stimulus recruits additional sources of excitation localized to the area around the driven neuron (we focus only on the attend preferred case here because the attend null follows from the same arguments). Since we have randomized the $I \rightarrow E$ connections, one can think of this in the complement: less localized inhibition (smaller $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{pref}}}$) will result in larger attentional effects. By contrast, the magnitude of the orthogonal sources of inhibition determines the neuron's normalization effects, with more orthogonal inhibition (larger $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{ortho}}}$) resulting in larger NMI. Since these two values are uncorrelated (Figure 15) we cannot expect there to be a relationship between the magnitude of the attentional effects and NMI. Nevertheless, this argument suggests a way forward.

The most obvious remedy is to anti-correlate $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{pref}}}$ and $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{ortho}}}$. In this way, stronger sources of orthogonal inhibition, which induce larger normalization indices, will be related to weaker sources of local inhibition, resulting in larger attentional effects. We implement this by modifying the sampled mean $I \rightarrow E$ weights with a cosine function whose peak is aligned orthogonally to a given neuron: $\mathbb{J}_{E_i I} = J_{EI} g(\theta_i, \sigma_{EI}) + \delta \cos(2\pi \frac{\theta_i + N/2}{N})$ (Figure 15). The areal inhibition metrics $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{pref}}}$ and $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{ortho}}}$ are now anti-correlated, as desired (Figure 15). Again, NMI and AMI are positively correlated (Figure 15). Yet we now see that there is a strong positive correlation between NMI and the change in rates to attend preferred, as desired (Figure 15). Consistent with this result, if we return to the original heterogeneous network and compute the inhibition metrics $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{pref}}}$ and $\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{ortho}}}$ we also find a negative correlation (not shown). Of course, this relationship occurred by chance through the first sampling procedure, and resampling the weights over instantiations of that network can result in the wrong relationships between NMI and the change in rates with attention.

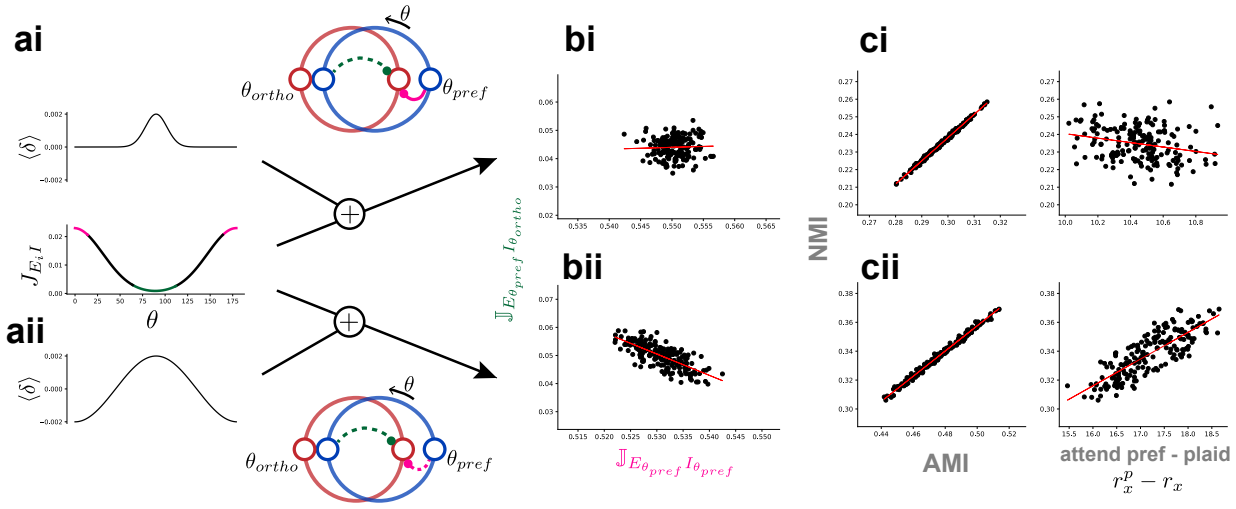

Figure 15: *Network dissection of inhibitory effects on correlating heterogeneities* **a** Schematic illustration of different sampling conditions. A small additional bias δ with variable amplitude was included from orthogonal sources of inhibition alone (i), or anti-correlated between proximal and orthogonal sources (ii). **b** Confirmation and illustration of sampling procedure in (a). Net proximal inhibitory inputs ($\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{pref}}}$; colored pink in (a)) are uncorrelated with net distal inhibitory inputs ($\mathbb{J}_{E_{\theta_{pref}} I_{\theta_{ortho}}}$; colored green in (a)) in the orthogonal alone condition (i) and anti-correlated in the second condition (ii). **c** Normalization and attention indices are correlated in both cases (left), whereas only the anti-correlation condition satisfies the normalization-attentional rates condition (right). Red lines show linear fits to data.

3.4 Discussion

Despite decades of research, the neural circuit mechanisms of attention have largely eluded explanation. Yet experimental evidence is not lacking. Attention has been clearly shown to increase neural response gain, affect oscillatory activity and modulate noise correlations. In turn, each of these effects has been linked to possible neural bases. Indeed, all three can be modulated through neuromodulatory effects [142], potentially through inhibitory subcircuits [93]. The source of these effects has been shown to originate in frontal cortices, though neuromodulators in the midbrain and brainstem have been shown to play an important role as well. What we potentially have, then, is an embarrassment of riches. A means by which to help narrow the space of explanations is required in order to make progress, and identify more targeted hypotheses. Mechanistic circuit modeling is well-positioned to fill this void. The challenge in model development is finding the right experimental data to properly constrain models. This is the problem we sought to address in the present study.

As outlined in the introduction, much previous theoretical work developed models which captured aspects of attention in a general way. Many successfully reproduced important observations. In a pessimistic vein, this could be seen as adding to the problem: a number of plausible successful models. To allude to a popular cliché, the question becomes, which are useful? Indeed, recent computational studies have appealed to the relationship between normalization and attention in support of various models [115, 83]. Undoubtedly this relationship is an intriguing starting point, presenting a clear link between the neural effects of a cognitive variable, and those of a circuit-dynamic variable which in principle need not be related.

As a first step we considered whether a single cell-type model could provide a sufficient level of explanation for the data. In this case we adopted a flexible phenomenological model as opposed to a biophysically rigorous model. This parametric flexibility allowed us to inquire whether the NMI-AMI relationship would serve as a good model constraint while abstracting away other details. Surprisingly, we found that this experimental observation was unable to distinguish between distinct model regimes and is therefore not a good test of model utility. Of course one possibility is that the metrics themselves belie any effective differentiation

between models. In point of fact, we gave an intuitive argument in section 3.3.1 that the symmetry in the definitions of NMI and AMI implied their positive correlation. Another possible obfuscation is that each metric is a ratio. While ratios play an important role in comparing across distinct values, in the context of uncovering mechanisms they can hide the driving factor. This was true in the case of attentional effects on correlations. While it was known that attention decreases correlated variability [24], separate analysis showed that variance changes little while covariance decreases significantly [68]. This was an important insight as the mechanistic implications of changing covariance as opposed to variance can be distinct. Similarly, we found that by considering a separate dataset showing a correlation between the magnitude of attentional effects (related to the numerator of AMI) and the strength of normalization (NMI) that we were able to find constraints on the single cell-type model.

The additional constraints we identified led to the conclusion from the single cell-type model that a saturating input-output function was best able to capture the data. Ultimately, both of the limitations we identified with the NMI-AMI correlation failing as a useful constraint bore out: by focusing on the absolute change in firing rate with attention, we broke the symmetry in the relationship with normalization, uncovering a condition that is satisfied only by saturating functions. It is well known that cortical neurons operating in physiologically normal range saturate far below the point at which refractoriness would play a restricting role. This saturation is due to network effects. Accordingly, our phenomenological analysis revealed that the correlation between attention and normalization is most probably a network effect.

There are naturally a number of limitations in our use of such a simple phenomenological model. The rigid assumptions on the structure of the inputs s_{ext} and the attentional effects μ_A limited the scope of conclusions one is able to draw. For example, rather than incorporating μ_A as an input to ϕ suppose that μ_A scaled the amplitude of ϕ. As discussed in more detail in the Methods, this would lead to AMI being homogeneous. If μ_A itself were allowed to vary, then this issue could be overcome. Similarly, s_{ext} served as a proxy for many different sources of input to a neuron, including feedforward and recurrent connections, which were made explicit in the network model. Clearly, the values s_n, s_p and s_x will in reality vary in

some rich way. However, if we navely allowed for s_{ext} and μ_A to independently vary in this model, then for NMI and AMI to correlate these external sources would need to be correlated. In this context, this would be a uselessly trivial conclusion. In a richer biophysical model which carefully considered how neuromodulators or neurotransmitters affect physiological properties of the cell perhaps this would not be the case, but we did not pursue this here.

Our use of the term "heterogeneous" was rather loose in the model. Naturally this implies that there is some underlying probability distribution for a given parameter. In this work, we always assumed it to be Gaussian. It is reasonable to assume that careful estimates of the true distributions in data may differ from our Gaussian assumption, but we don't expect this alone to significantly change our results. The stronger assumption in this work was joint independence between parameters, which is not likely true in reality (though see Supplemental material). Nevertheless, we accepted this as a simplifying assumption since a complete characterization of possible fits to neural response profiles was beyond the scope of this work but would be a useful direction for future inquiry.

Turning then to a network model, we showed that by introducing heterogeneity in the synaptic connections between I and E we were able to recapture the three experimental observations. We then further dissected this result by identifying the distribution of heterogeneity which was required to robustly capture the data. In this regard, we found that inhibitory inputs proximal to a given E cell needed to be anti-correlated with those orthogonal to it.

There are a number of shortcomings in our choice of a ring network model [10]. Principally, the organization of E/I pairs confounds the physical location of neurons in cortex with their stimulus tuning properties. However, it has been shown that neurons preferentially connect as a function of both physical space and shared tuning preference, suggesting this is a reasonable approximation [79]. Our implementation of attention in this model deserves some scrutiny. As mentioned in the introduction, attention to stimulus features and attention to regions of space function differently within the brain. By the nature of our circuit model, attending to the preferred stimulus was indistinguishable from attending to the region of space in which the neural population preferring that stimulus is located. Given that primate cortex is organized in a columnar fashion with similarly tuned cells nearby one-another, this

nevertheless appears again to be a reasonable approximation. We note that the motivating experiments were careful to control against feature attention, instead restricting the attentional paradigm to one of space [101]. Thus a more faithful model would dissociate these two components. One means of achieving this is to organize neurons on a two-dimensional sheet and assign each a tuning preference through some map, such as a pinwheel [71]. Future work could implement this spatially extended model and confirm whether the present results still hold.

One question prompted by the network analysis is whether the anti-correlation required of the inhibitory synaptic inputs to E is plausible. Studies of synaptic plasticity mechanisms suggest that excitatory cells exhibit synaptic scaling and homeostatic mechanisms which can regulate the strength and number of connections with inhibitory interneurons. If certain connections proximal to a neuron are potentiated or increase in number, say, it is conceivable that orthogonal inputs may be down-regulated as a result. In this way an E cell could experience anti-correlated inhibitory inputs between local and distal sources, consistent with our prediction.

There are many natural extensions to this work. As we touched upon briefly above, a prominent result in attentional studies is that correlated variability within a cortical region decreases with attention [24, 150]. While we did not explore it here, we expect this to be consistent with our conclusions as well. For if we consider again the phenomenological model and linearize about a given operating point, the expansive input-output function will amplify input variability with increases in rates whereas the saturating function will quench variability. A recent study showed this held true in an analogous ring model to the one we used here, in the context of Fano factor reduction with stimulus onset [60]. However, it is unclear what additional model constraints may arise from incorporating this additional piece of data. Conceivably, it could help narrow the space of viable attentional perturbations and be a test on the one we chose here, namely, an excitatory input bias to E cells alone.

Modern techniques such as optogenetics have enabled researchers to probe the dynamics of neural circuits through targeted manipulations of activity. If the network results derived here are robust across datasets, they might suggest that attention is (and by extension other cognitive variables are) useful for inferring network structure and dynamics. In this way,

attention could serve as a natural perturbation experiment. More generally, the analysis of correlated heterogeneities between dynamic and cognitive variables may provide a useful, general avenue forward in linking these two regimes.

3.5 Methods

In this section we provide full details of each model described in the main text, together with derivations and claims therein. Additional analyses and figures are provided in Supplemental material.

3.5.1 Single unit phenomenological model

Requirements for heterogeneity in normalization

Recall the definition of the noiseless firing rate r for neuron i, without attention:

$$\dot{r_i} = -r_i + \phi(b + s_{ext})$$

There are two potential sources of cellular heterogeneity: b, denoting the cellular baseline activity (or threshold) and ϕ, the input-output function. We take $\phi(I) = k\lfloor I \rfloor_+^m$ where I denotes the collective inputs. Heterogeneity in ϕ is thus parameterized by the distributions over k and m. k and m are both bounded below by zero. s_{ext} denotes the net inputs to the neuron. Consistent with our assumptions that normalization is inherited and that stimulus inputs are constant across all units implies three stimulus input terms: $s_n \leq s_x \leq s_p$ representing input drive from the null, plaid (or, concurrently presented preferred and null stimuli in a neuron's receptive field, superimposed [119] or disjoint [101]), and preferred stimulus, respectively.

We begin by fixing ϕ and considering the effect of variability in b on NMI, defined in equation 38. Denote the NMI of a single unit NMI_i:

1. ϕ is linear ($m = 1$): $r_i = k(b_i + s_{ext})$. Hence $\text{NMI}_i = \frac{(r_i)_p - r_x}{(r_p - r_n)+(r_x - r_n)}$ $= \frac{k}{k}\frac{(b_i+s_p)-(b_i+s_x)}{(b_i+s_p-(b_i+s_n))+(b_i+s_x-(b_i+s_n))} = \frac{s_p - s_x}{(s_p - s_n)+(s_x - s_n)}$ for all i. This last term has no dependence on b_i and is therefore the same for every neuron.

2. ϕ is threshold quadratic ($m = 2$): $r_i = k(b_i + s_{ext})^2$. We now have that $\text{NMI}_i = \frac{(b_i+s_p)^2-(b_i+s_x)^2}{((b_i+s_p)^2-(b_i+s_n)^2)+((b_i+s_x)^2-(b_i+s_n)^2)}$ which does depend on b_i. For a given collection of inputs $s_n \leq s_x \leq s_p$ we ask how NMI changes in the limit as $b \to \infty$ (we have dropped the subscript i for notational convenience). For $b = 0$ NMI$= \frac{(r_p - r_n)-(r_x - r_n)}{(r_p - r_n)+(r_x - r_n)} = \frac{((s_p)^m-(s_n)^m)-((s_x)^m-(s_n)^m)}{((s_p)^m-(s_n)^m)+((s_x)^m-(s_n)^m)} \geq 0$ with equality only when $s_p = s_x$. Hence we let $s_p > s_x \implies \text{NMI}|_{b=0} > 0$. For $b > 0$ we consider

$$\lim_{b\to\infty} \text{NMI} = \lim_{b\to\infty} \frac{((b + s_p)^m - (b + s_n)^m) - ((b + s_x)^m - (b + s_n)^m)}{((b + s_p)^m - (b + s_n)^m) + ((b + s_x)^m - (b + s_n)^m)} = \lim_{b\to\infty} \frac{\Delta_p - \Delta_x}{\Delta_p + \Delta_x}$$

where we have defined $\Delta_p = (b + s_p)^m - (b + s_n)^m$ and similarly for Δ_x. We will show that for b increasing, NMI is decreasing for all $m > 1$. Notice that $\Delta_p + \Delta_x > \Delta_p - \Delta_x$ since $\Delta_p, \Delta_x > 0$. Further, $\frac{d}{db}\Delta_p = m[(b + s_p)^{m-1} - (b + s_n)^{m-1}]$ and similarly for Δ_x. Hence we also have that $\frac{d}{db}(\Delta_p + \Delta_x) > \frac{d}{db}(\Delta_p - \Delta_x)$. As ϕ is monotonically increasing, NMI varies monotonically with $b \geq 0$. From the foregoing arguments we have that the denominator is larger, and grows faster than, the numerator, hence, NMI is decreasing.

If $0 \leq m < 1$ we must be more careful. For fixed m we have that $\frac{d}{db}(b+s_p)^m = \frac{m}{(b+s_p)^{1-m}} \leq \frac{d}{db}(b+s_x)^m = \frac{m}{(b+s_x)^{1-m}} \leq \frac{m}{(b+s_n)^{1-m}} = \frac{d}{db}(b+s_n)^m$ by the ordering of the inputs. What this then implies is that in the full expression for NMI the denominator decreases faster than the numerator. This can be seen more easily by rewriting $\text{NMI} = \frac{(b+s_p)^m-(b+s_x)^m}{(b+s_p)^m+(b+s_x)^m-2(b+s_n)^m}$ and hence NMI is now increasing for increasing b.

If instead r_n is kept near zero then for any b $\lim_{b\to\infty}\text{NMI} = \lim_{b\to\infty} \frac{(b+s_p)^m-(b+s_x)^m}{(b+s_p)^m+(b+s_x)^m} \to 0$ for all m. This condition appears to be in general agreement with experimental data and for that reason we implemented this condition in the main text.

Thus we have that ϕ must be nonlinear. Notice that if ϕ is threshold linear, $\phi = k\lfloor b_i+s_{ext}\rfloor_+$, then provided $b_i+s_{ext} < 0$ for some s_{ext} we have that NMI will be heterogeneous. However for $b_i + s_{ext} > 0$ the results in example 1 hold. We will generally consider this to be the case.

The other source of potential heterogeneity is ϕ. If we now fix the threshold b constant and allow ϕ to vary across cells, we have from the argument in example 1 above that if ϕ is linear ($m = 1$) and k varies we have the same result (replace all k's in the examples with neuron-specific k_i's and the conclusion follows). Hence, provided ϕ is nonlinear then heterogeneities in ϕ will reflect as heterogeneities in NMI. The nature of this variation can be extrapolated from a simple example. Consider m in the range [0, 1]. We have shown that m near 1 will approximately preserve the inherited NMI while m close to 0 will result in $r_p - r_x \approx 0$ for sufficiently large inputs, hence very low NMI. Taken together, NMI should be proportional to m, and should induce a reduction in NMI relative to the inherited value. We next show this holds in general. Without loss of generality set $b = 0$, and assume the null response is 0 Hz ($r_n = 0$). Then

$$\begin{aligned} \text{NMI} = \frac{r_p - r_x}{r_p + r_x} &= \frac{ks_p^m - ks_x^m}{ks_p^m + ks_x^m} = \frac{s_p^m - s_x^m}{s_p^m + s_x^m} \\ \implies \frac{d}{dm}\text{NMI} &= \frac{(ln(s_p)s_p^m - ln(s_x)s_x^m)(s_p^m + s_x^m) - (s_p^m - s_x^m)(ln(s_p)s_p^m + ln(s_x)s_x^m)}{(s_p^m + s_x^m)^2} \\ &= \frac{2s_p^m s_x^m (ln(s_p) - ln(s_x))}{(s_p^m + s_x^m)^2} \geq 0 \end{aligned} \tag{42}$$

with equality if $s_p = s_x$ or $s_x = 0$. Hence, NMI grows as a function of m for fixed inputs, as claimed.

Requirements for attentional heterogeneity

We now extend the previous rate model to include a term reflecting the effects of attention. Since attentional processes have been shown to affect firing rates of neurons, we modify the rate expression for neuron i in one of two ways, indicating the attentional modifier with the term μ_A:

1. *additive attention* $\dot{r_i} = -r_i + \phi(b + \mu_A + s_{ext})$ in which attention μ_A acts like an additional stimulus input
2. *multiplicative attention* $\dot{r_i} = -r_i + \mu_A \cdot \phi(b + s_{ext})$ in which attention μ_A scales the input-output function

There are additionally two attentional conditions to consider: attend preferred (μ_p), in which attention is cued to the preferred stimulus while both preferred and null stimuli (plaid) are present in a neuron's receptive field, and attend null (μ_n), in which attention is cued to the null stimulus instead. From experimental data we have that attending the preferred stimulus increases firing rates relative to the plaid, on average returning them to the rate of firing to the preferred stimulus alone [101]. Conversely attending null decreases firing rates on average relative to the plaid, but is not so suppressive as to reduce firing rates to those when the null stimulus is present alone [101, 150].

Recall that we are treating attention (like the stimulus parameters s_{ext}) as homogeneous across cells. We now take as given that ϕ is nonlinear and consider two examples:

1. Suppose attention is multiplicative. For fixed ϕ and variable b we have

$$\mathrm{AMI}_i = \frac{\mu_p \cdot \phi(b_i + s_x) - \mu_n \cdot \phi(b_i + s_x)}{\mu_p \cdot \phi(b_i + s_x) + \mu_n \cdot \phi(b_i + s_x)} = \frac{\mu_p - \mu_n}{\mu_p + \mu_n} \cdot \frac{\phi(b_i + s_x)}{\phi(b_i + s_x)} = \frac{\mu_p - \mu_n}{\mu_p + \mu_n}$$

   for all i. This last term has no dependence on b_i and AMI is therefore homogeneous across cells. If we instead assume that ϕ is heterogeneous we can see from the same argument that AMI will be invariant as well. From this, we conclude that attention cannot be incorporated into this model in this form, given the assumptions of constant inputs and attentional effects.
2. Now consider the additive attention model. With a heterogeneous b it follows immediately from the definition that AMI will be heterogeneous since

$$\mathrm{AMI}_i = \frac{\phi(b_i + \mu_p + s_x) - \phi(b_i + \mu_n + s_x)}{\phi(b_i + \mu_p + s_x) + \phi(b_i + \mu_n + s_x)}$$

   and the relationship between $(r_i)^p_x = \phi(b_i + \mu_p + s_x)$ and $(r_i)^n_x = \phi(b_i + \mu_n + s_x)$ depends on the choice of b_i. Analogously, variability in ϕ will also induce AMI to be heterogeneous. This can be seen by making the substitution $\phi \mapsto \phi_i$ in the foregoing arguments.

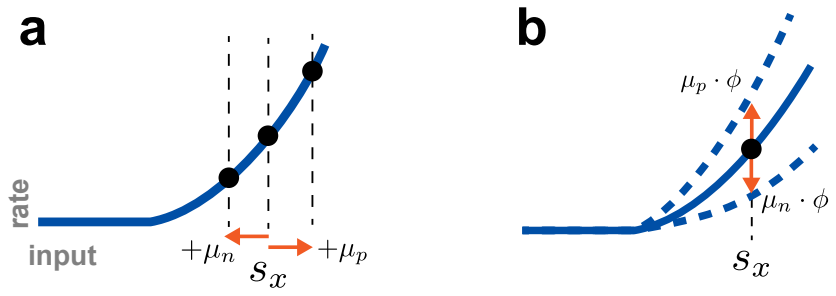

Figure 16: *Single neuron heterogeneities in attentional models* **a** Schematic illustration of additive attention. **b** Schematic illustration of multiplicative attention.

The above arguments show that if we incorporate attention as some additive input process (or, e.g., as a contrast change) then the previously derived nonlinear transfer function will insure that AMI is also heterogeneous. We note that the homogeneity of attention as a gain modulator rests of the rigidness of our assumptions; any heterogeneity in attentional processes themselves could induce heterogeneity in AMI. While likely true in reality, our focus here has been on uncovering necessary cellular mechanisms which may give rise to variable attentional responses without assuming any inherited variability in attentional processes or signals.

Requirements for NMI and AMI to be positively correlated

Given that we have conditions under which the single cell model will exhibit heterogeneity in normalization and attention we next ask what conditions are necessary for the two to be positively correlated [101]. In particular, given our assumption that ϕ is monotonic and non-decreasing it suffices to ask whether NMI and AMI change in the same way for a given ordering of ϕ or b.

If b is the dominant heterogeneity, we have already shown above that NMI decreases with increasing b if we keep r_n is close to zero. The argument for AMI is much more straightforward. Assuming as before that μ_p is a positive perturbation and μ_n a negative

perturbation with respect to s_x

$$\lim_{b\to\infty} \text{AMI} = \frac{(b+\mu_p+s_x)^m - (b+\mu_n+s_x)^m}{(b+\mu_p+s_x)^m + (b+\mu_n+s_x)^m} \approx \frac{b^m - b^m}{b^m + b^m} \to 0$$

and since again for $b = 0$, AMI > 0, AMI too is monotonically decreasing with increasing b. Thus, AMI and NMI are positively correlated.

The above also holds true if we relax the constraint on $r_n \approx 0$ but fix $m > 1$. By contrast if we let $0 \leq m < 1$ then NMI and AMI will be anti-correlated in this scenario. Hence, for the additive attentional model with variable baseline and polynomial input/output function NMI and AMI are strictly positively correlated for $m > 1$, whereas we can in principle see either a positive or negative relationship between NMI and AMI with $m < 1$, however experimental evidence suggests r_n close to zero and thus in this condition as well we see a positive correlation.

Now consider the model with variable ϕ, in particular, m heterogeneous. As we showed above, NMI increases for increasing m. The same argument goes through for AMI since (setting $b = 0$ for convenience)

$$\frac{d}{dm}\text{AMI} = \frac{2(\mu_p+s_x)^m(\mu_n+s_x)^m(ln(\mu_p+s_x) - ln(\mu_n+s_x))}{((\mu_p+s_x)^m + (\mu_n+s_x)^m)^2} \geq 0$$

Thus NMI and AMI will be positively correlated in this model as well.

Calibrating constraints from Ni et al. *and Verhoef and Maunsell*

Verhoef and Maunsell [150] reported rate changes as a function of neuron selectivity for a given stimulus and attention condition (Figure 13). Roughly, the selectivity metric used in that study is meant to capture the responsiveness of a neuron to its preferred stimulus relative to the null stimulus. To simplify our treatment in this section, unless otherwise noted we will consider firing rate changes with respect to the preferred stimulus (e.g. Figure 13a). For a selectivity index above approximately 0.25, attending preferred causes an increase in rates relative to the plaid for all normalization indices [150].

In Verhoef and Maunsell [150] normalization was quantified with a metric distinct from NMI. In order to relate these results to the model we're studying, we first need conditions for proportionality of the two metrics of interest: NMI [101] and nonpreferred suppression

[150] (we'll call this new metric SI for *suppression index*). The latter is derived from fits of the following equation to measured firing rates:

$$r_x = \frac{L_1 + L_2}{\alpha_1 + \alpha_2 + \sigma} \tag{43}$$

with r_p defined by setting $\alpha_2, L_2 = 0$ and r_n defined by setting $\alpha_1, L_1 = 0$. After fitting these parameters to the data, SI was defined

$$\text{SI} = \frac{\alpha_2}{\alpha_1 + \alpha_2}. \tag{44}$$

We claim that this follows the same trend as NMI and thus using the latter is 'good enough' to uncover the right relationships. Assume σ is sufficiently small and hence ignorable [149]. (Experimental results show that while this is a reasonable assumption there is a small but significant improvement in model fits obtained by incorporating σ [149].) Then $L_1 = r_p\alpha_1, L_2 = r_n\alpha_2 \implies r_x(\alpha_1 + \alpha_2) = r_p\alpha_1 + r_n\alpha_2$ by the above definition of r_x. Write $c_i = \frac{\alpha_i}{\alpha_1+\alpha_2}$. Then by definition $c_2 :=$ SI. It follows from these definitions that

$$r_x = c_1 r_p + c_2 r_n = (1 - c_2)r_p + c_2 r_n.$$

Because of the assumed ordering of s_{ext} in section 3.3.1, $r_x \in [r_n, r_p]$ from which it follows that $c_2 \in [0, 1]$. Intuitively we see that as c_2 approaches 1, r_x approaches r_n, which corresponds to greater normalization. Similarly, this same limit also appears as an increase in NMI towards 1. To be more precise we rearrange and solve for c_2:

$$c_2 = \frac{r_p - r_x}{r_p - r_n}.$$

If we compare this to the definition of NMI in equation 38, we see that while this expression is linear in r_x NMI is nonlinear in r_x. In spite of this, both follow the same trend. Note that both metrics will be zero ($r_x = r_p$) and one ($r_x = r_n$) at approximately the same inputs, the main difference being in the scaling at intermediate values. For example, response averaging ($r_x = (r_p + r_n)/2$) yields $\frac{1}{3}$ with NMI and $\frac{1}{2}$ with SI. Hence, going forward, we will consider NMI a general metric of normalization, and use the acronym interchangeably with the word "normalization" with the understanding that some published data may be reported with a slightly different metric.

Incorporating constraints from Verhoef and Maunsell

The firing rate results reported in Verhoef and Maunsell [150] can be summarized in three general points. For sufficiently high selectivity (as defined in the previous section):

1. Firing rates in response to simultaneously presented preferred and null stimuli are lower for neurons with stronger normalization (not shown; see [150])
2. The increase in firing rate with attention to the preferred stimulus is greater for neurons with stronger normalization (Figure 13a)
3. The change in firing rate with attention to the null stimulus is greater for neurons with stronger normalization (Figure 13b)

In the framework of our model, the first point says that r_x should be inversely proportional to NMI. If b is the dominant source of heterogeneity then this is true for all m since NMI decreases with increasing b. Hence, this observation is not a good constraint in the context of variable b. We consider the case of variable m in the Supplemental.

The second point says that $r_x^n - r_x \propto$ NMI, and the third point can be approximated by $|r_x^n - r_x| \propto$ NMI. Taken together, non-normalizing (low NMI) cells have higher rates and a smaller derivative with respect to attentional changes while strongly normalizing units have lower rates but a larger derivative with respect to attentional changes. We again look at two cases:

1. Let ϕ be threshold quadratic (a specific case of $m > 1$) and b heterogeneous with other parameters fixed. Then $\frac{\mathrm{d}\phi}{\mathrm{d}I} = 2\lfloor I \rfloor_+$ where I is again a generic input. Since NMI decreases with increasing input, this implies that a smaller NMI will correspond to a bigger changes in rates (since changes with attention are modeled as perturbations around r_x, these will be larger when the derivative is greater), which in turn implies $r_x^p - r_x \propto 1/\text{NMI}$. This contradicts the second data relationship and hence $m > 1$ fails to capture the data.
2. Now let $0 < m < 1$. ϕ thus describes a saturating function. We now have that the derivative $\frac{\mathrm{d}\phi}{\mathrm{d}I} = \lfloor \frac{m}{I^{1-m}} \rfloor_+$ clearly decreases with increasing rate (input). Hence, $r_x^p - r_x \propto$ NMI and so a saturating function is capable of capturing the correct relationship between normalization and rate changes with attention.

We consider the case of m strongly heterogeneous in the Supplemental (section 3.6), where we show that additional assumptions are required to match the experimental data.

3.5.2 Heterogeneous ring network model

We use a modified version of a previously published ring network in which contrast response functions saturate due to strong inhibitory connections [118]. In particular, we structure an E/I network around a ring such that each unit occupies a position which corresponds to its preferred value of a tuning variable θ (Figure 18A). The dynamics of a single unit are given by (rewriting equation 41)

$$\tau_\alpha \dot{r_\alpha} = -r_\alpha + \kappa[I_{ext,\alpha} + \mathbb{J}_{\alpha\alpha} r_\alpha + \mathbb{J}_{\alpha\beta} r_\beta]_+^2 \tag{45}$$

where $\alpha, \beta \in \{E, I\}$. $\mathbb{J}_{\alpha\beta}$ is a connectivity (weight) matrix from population β to population α, the elements of which are given by $J_{\alpha_i\beta_j}$. $I_{ext,\alpha} = b + cg(\theta) + \mu_\alpha$ is the external input with a constant baseline b for all cells, $cg(\theta)$ an external stimulus with contrast c, and μ_α captures attentional effects (here attention is applied to E or I, as described below, by contrast to the single unit model in which attention was defined relative to the stimulus). $g(\theta)$ is modeled as a wrapped Gaussian of width σ centered at θ_0:

$$g(\theta; \theta_0, \sigma) = c_0 \sum_{i=-\infty}^{\infty} e^{-(\theta - \theta_0 + iN)^2 / 2\sigma^2} \tag{46}$$

c_0 is a constant which normalizes the peak of the curve to 1. Attention is delivered only to E cells with a spatial profile which matches that of the stimulus; thus, $\mu_I = 0$ and $\mu_E = d\mu_E \cdot g(\theta)$ where $d\mu_\alpha$ scales the magnitude of the attentional signal [83]. In the unattended state, $\mu_\alpha = 0$. Changes in cortical state with attention are thus captured by a change in the operating point of the system.

Connectivity is structured in a distance-dependent manner also described by equation 46 of width $\sigma_{\alpha\beta}$. In order to induce heterogeneous response profiles in the E network, $I \to E$ weights are randomly drawn from a Gaussian probability distribution with mean equal to the non-variable network, that is, the connection $J_{E_i I_j}$ is drawn according to $\mathcal{N}(g(\theta_{I_j}; \theta_{E_i}, \sigma_{EI}), \sigma_\eta)$. Weights which change sign are set to zero to obey Dale's law. This

will inevitably affect the sampled mean of the connection weights. By contrast, all other connections are symmetric about the ring and therefore do not vary from cell to cell. The modifications we introduced to this procedure were described in section 3.3.4.

In section 3.3.4 we introduced two metrics to quantify inhibition localized to an E unit, $\mathbb{J}_{E_{i,\theta_{pref}} I_{\theta_{pref}}}$, and distal to it (i.e., from orthogonally-tuned sources), $\mathbb{J}_{E_{i,\theta_{pref}} I_{\theta_{ortho}}}$. Precisely, each metric is the dot product of a weight vector with the inhibitory inputs to a given E unit:

$$\mathbb{J}_{E_{i,\theta_{pref}} I_{\theta_{pref}}} = w(i) \cdot J_{E_i I} \tag{47}$$

and

$$\mathbb{J}_{E_{i,\theta_{pref}} I_{\theta_{ortho}}} = w(i + N/2) \cdot J_{E_i I}. \tag{48}$$

where $w(i)$ is a weighting function (vector) defined as a Gaussian centered at the i^{th} location on the ring: $w(i) = g(\theta; \theta_i, \sigma_w)$. In the first equation this corresponds to the E unit and in the latter equation orthogonal to the relevant E unit.

3.5.3 Model and Simulation Parameters

Figure 2, 3: $(m < 1)$ $b \sim \mathcal{N}(0, 0.25^2)$, $s_n = 0, s_x \sim \mathcal{N}(2.5, 0.1^2), s_p \sim \mathcal{N}(5, 0.05^2), \mu_p = s_p - s_x + r_1, \mu_n = (s_n - s_x)/2 + r_2$ where r_1, r_2 are random variables with distribution $\mathcal{N}(0, 0.05^2)$. r_n was set equal to zero. Otherwise, rates were computed on a per-unit basis with m sampled from $\mathcal{N}(0.3, 0.05^2)$, $k = 1$.
$(m > 1)$ Same as above with the following exceptions: $s_x \sim \mathcal{N}(2, 0.01^2), s_p \sim \mathcal{N}(4, 0.01^2), \mu_n = (s_n - s_x)/3 + r_2$ where r_2 is as above. r_n was again set equal to zero. $m \sim \mathcal{N}(2, 0.01^2)$.

Figure 4: Baseline model parameters follow Rubin *et al.* [118]. The number of E and I units, respectively, was $N_E = N_I = 180 := N$. The scaling term of the transfer function $\kappa = 0.04$, and $\tau_E = 20$ms, $\tau_I = 10$ms. Inputs were given to both E and I units, with baseline $b = 5$, contrast 30 (a.u.) and width 30^o. The attentional signal had amplitude $d\mu_E \in [1.5, 2.5]$, chosen such that E rates approximated that when the preferred stimulus alone drove the network. Connectivity parameters were given by $J_{EE} = 0.044, J_{EI} = 0.023, J_{IE} = 0.042, J_{II} = 0.018$. The connectivity width $\sigma_{\alpha\beta} = 32^o$ was the same for all class pairs. Heterogeneity was introduced as described above with sample variance $\sigma_\eta = 0.0025^2$ regardless of the sample mean.

Parameters were ultimately chosen to insure that the system settled down to a steady state solution.

Figure 5: Same baseline parameters as before, with the following modifications to the randomization procedure. In section 3.3.4 we defined the orthogonal perturbation (rewritten in terms of single connections) $J_{E_iI_j} = J_{EI}g(\theta_j;\theta_i,\sigma_{EI}) + \delta g(\theta_j;\theta_i + N/2,\sigma_{ortho}), \delta \sim \mathcal{N}(0.001, 0.001^2), \sigma_{ortho} = 12$. This function defined the mean of the sample distribution, from which individual connections were drawn as $\mathcal{N}(J_{E_iI_j}, 0.001^2)$. The mean of the anti-correlation function was defined similarly: $J_{E_iI_j} = J_{EI}g(\theta_j;\theta_i,\sigma_{EI}) + \delta\cos(2\pi\frac{(\theta_j-\theta_i)+N/2}{N}), \delta \sim \mathcal{N}(0, 0.0015^2)$ and individual connections were drawn as $\mathcal{N}(J_{E_iI_j}, 0.0005^2)$.

3.6 Supplemental Material

3.6.1 Alternative solutions to phenomenological model constraints

Satisfying all constraints with heterogeneous ϕ

The addition of the constraints from Verhoef and Maunsell [150] placed further requirements on how parameters of ϕ would need to vary in order to satisfy them. Suppose m varies. The first stated constraint from this study was that r_x goes like the inverse of NMI. We observed in the Methods that NMI is proportional to m and thus r_x is increasing with NMI. We could fix this by allowing k to vary inversely with m. Then $k(m) = k_0^{-m}$ and

$$\frac{d}{dm}r_x = \frac{d}{dm}k_0^{-m}s_x^m = k_0^{-m}s_x^m(ln(s_x) - ln(k_0))$$

where we omitted the baseline term b. So if $k_0 > s_x$ then r_x will decay as a function of m and thus be inversely related to NMI. Notice this would not affect the base case relationship between NMI and AMI because they are independent of k.

To then match the attentional conditions we require that the derivative of the input-output (I/O) function evaluated at s_x grows with m (thereby correlating with NMI). Choose

$m_1 < m_2$ and label the associated rate responses r_1, r_2 respectively. This implies (taking $b = 0$) we require

$$\frac{d}{ds_x}(r_1)_x = \frac{d}{ds_x}k_0^{-m_1}s_x^{m_1} = m_1 k_0^{-m_1}s_x^{m_1-1} < m_2 k_0^{-m_2}s_x^{m_2-1} = \frac{d}{ds_x}(r_2)_x \tag{49}$$

$$\implies k_0 < \left(\frac{m_2}{m_1}\right)^{\frac{1}{m_2-m_1}} s_x \tag{50}$$

Since k_0 is bounded below by s_x we now have an additional upper bound on k_0. Therefore while it is in principle possible to satisfy each of the constraints with dominant heterogeneity in m, due to the finely tuned nature of the result under this model, we did not consider it further.

m *and* b *jointly heterogeneous*

A natural third possibility in the single unit model is that m and b vary jointly. It therefore remains to determine whether the right balance of heterogeneity in b and m could produce the correct collection of correlated heterogeneities for any m. To this end we numerically explore $m - b$ and $m - s_x$ space for $m > 1$ and $0 < m < 1$. We seek paths through $m - b$ or $m - s_x$ space in which the local gradient will obey the correct rate-normalization relationships. We perform this latter manipulation to insure that our choice of s_x when searching over b space isn't drastically affecting the nature of our conclusions.

We partition $m - b$ space into a discrete grid (the same procedure is applied to $m - s_x$ space, under the change $b \mapsto s_x$) such that m increments in steps of Δm and b in steps of Δb. For a given location on the grid (b_k, m_i), we compute the effect of a perturbation in the direction of each of the vectors $(\Delta b, 0), (0, \Delta m), (\Delta b, \Delta m)$ on the values r_x, NMI, and the derivative $\frac{d}{ds}\phi|_{s_x}$. In particular we focus only on the sign of the change for each of the three computed values.

We focus on fitting the three criteria relating absolute rate changes to normalization. In requiring that $r_x \propto 1/\text{NMI}$ we check whether $\text{sign}(\Delta r_x) \neq \text{sign}(\Delta\text{NMI})$. To assess whether $|r_x^p - r_x| \propto \text{NMI}$ and $|r_x^n - r_x| \propto \text{NMI}$, we check whether the condition $\text{sign}(\Delta \frac{d}{ds}\phi|_{s_x}) = \text{sign}(\Delta\text{NMI})$.

We begin by searching through $m - b$ space. As done previously, we set $r_n = 0$ and fix s_x, s_p while varying m and b. We therefore compute NMI and the attention conditions from

the rates $r_x = (b + s_x)^m, r_p = (b + s_p)^m$. We see that in no case are the two relationships jointly satisfied for $m > 1$ but for $m < 1$, consistent with the calculations in section 3.5.1, all perturbations in the b direction satisfy the constraints (Figure 17, red lines). In other words: if b is the dominant heterogeneity, then the correct relationships will be satisfied.

If instead we let the x-axis correspond to the input variable s_x, this describes the situation in which the input normalization to a cell is variable. Here we fix s_p, such that for fixed m, r_p is constant. On the one hand nearly all of the $m - s_x$ space for $m < 1$ is permissible. In contrast, there is a narrow range in which the rate-normalization relationships are obeyed for $m > 1$ provided m is the dominant heterogeneity. This is in fact consistent with the bounds calculated in the preceding section. To see this, we note that since we have chosen $k_0 = 1$, $s_x = 1$ is an upper bound (Figure 17). To calculate the lower bound, we take the inequality given by equation 50, writing $m_2 = m + \Delta m$, from which it follows that

$$k_0 < \left(\frac{m + \Delta m}{m}\right)^{\frac{1}{\Delta m}} s_x. \tag{51}$$

Rewriting as an equality, substituting in $k_0 = 1$ and rearranging, we find

$$m = \frac{s_x^{\Delta m} \Delta m}{1 - s_x^{\Delta m}} \tag{52}$$

defines the lower boundary, which agrees well with the numerical calculations (Figure 17, green line). Hence, this region is completely described by our previous calculations. These results show again that $m < 1$ is the more plausible model, since uncovering the correct relationships for $m > 1$ would require fine-tuning. As a consequence, this fine-tuning could also limit the range of NMI attainable to a narrow region, which would contradict the data. These results thus establish more completely the generality of the result that $0 \leq m < 1$.

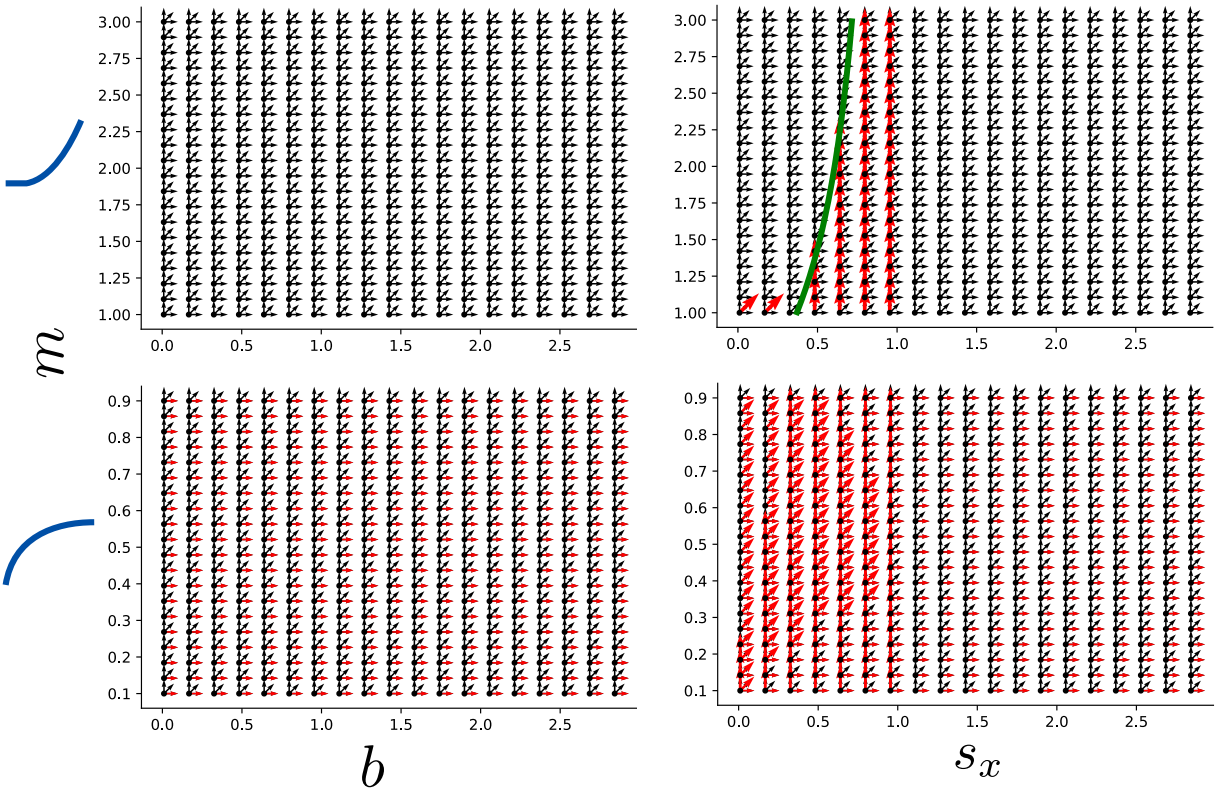

Figure 17: Quiver plots of valid gradients: black arrows fail at least one of the two directional relationships $r_x \propto 1/\text{NMI}$, $\frac{d}{ds}\phi|_{s_x} \propto \text{NMI}$; red arrows satisfy both. Green line in upper-right panel shows constraint boundary calculated from equation 52. Parameter ranges: $(0 < m < 1) 0.1 < m < 0.9; (m > 1) 1 < m < 3$; for b variable, $0.01 < b < 3; s_x = 1, s_p = 4$; for s_x variable, $0.01 < s_x < 3, s_p = 4$.

3.6.2 A naïve analysis of inhibitory connectivity effects supports our circuit dissection procedure

In section 3.3.4 we showed that by choosing the underlying statistical structure of the $I \to E$ connectivity wisely we could uncover a network which satisfied the experimental constraints. In this section we briefly summarize results from a much simpler, nave approach showing that the distribution of inhibition around the ring, rather than the absolute magnitude of inhibitory inputs to an E unit, is critical to uncovering the correct circuit architecture to support the experimental evidence.

We first compute the total $I \to E$ input, testing the hypothesis that the net amount of inhibitory input to an E unit determines the observed effects across normalization and attention. Indeed, calculating the total $I \to E$ connectivity for the i^{th} E unit as $\sum_j J_{E_i I_j}$ we see that this value is, in fact, a strong predictor of NMI and AMI (Figure 18b), thereby implicating the heterogeneities in a neuron's inhibitory field in its attentional and normalizing signatures.

Now consider a proportional metric, the normalized dot product $\frac{\mathbb{J}_{E_i I} \cdot \vec{1}}{||\mathbb{J}_{E_i I}|| \cdot ||\vec{1}||}$, where $\vec{1}$ is the vector of all ones. We again see a strong relationship between this value and both NMI and AMI (Figure 18c). This says something more: NMI and AMI are strongly correlated with a broader spread of inhibition around the ring. This normalized metric is the cosine angle between the constant vector $\vec{1}$ and the weight vector $\mathbb{J}_{E_i I}$. Intuitively, the more peaked $\mathbb{J}_{E_i I}$, the more restricted its connectivity is to nearby excitatory units, by the distance-dependence construction of the weights. This further implies that this vector and the constant vector are closer to being orthogonal, producing a smaller value of the normalized dot product. Conversely, a more broadly connected $\mathbb{J}_{E_i I}$ will have a profile that more closely resembles a constant function around the ring, hence, a smaller angle with the vector $\vec{1}$ and consequently a larger value of the normalized dot product.

If we recall our circuit dissection analysis, we showed that a perturbation (a cosine function) to the underlying sample mean of the inhibitory inputs to a given excitatory unit which anti-correlates the local and distal sources of inhibition would best capture the data. If this perturbation was large in magnitude, it served to flatten out the distribution of inhibition

around the ring, whereas if the perturbation was small, inhibition would remain peaked. This observation is consistent with the conclusions in this section, thereby supporting our hand-designed approach.

3.6.3 Alternative sources of synaptic heterogeneity and E/I balance

There is a long history of models incorporating subtraction or division to explain normalization mechanisms [19]. This largely motivated our exploration of monosynaptic inhibitory synaptic sources to explain the correlated heterogeneities. Another possible explanation is through recurrent excitation itself.

We test this by implementing the same procedure used to introduce heterogeneity in the $I \rightarrow E$ weights in an anticorrelated fashion between local and distal and applying it instead to the $E \rightarrow E$ weights (section 3.3.4 in the main text). Hence the inputs from all excitatory units to the i^{th} E unit are given by $\mathbb{J}_{E_iE} = J_{EE}g(\theta_i, \sigma_{EE}) + \delta\cos(\theta_i + N/2)$ where J_{EE} scales the magnitude the $E \rightarrow E$ weights, $g(\theta, \sigma)$ is defined in equation 46, σ_{EE} is the width of the $E \rightarrow E$ connectivity and δ is a random variable which scales the amplitude of the anticorrelation function. Again, this sampling procedure produces the correct relationships as observed in data (Figure 19).

This result suggests that the driving force behind the agreement of our network model and data is not necessarily the absolute magnitude of inhibition but the distribution of E/I balance around the ring, though the two are often related (see section 3.6.2).

We note that the nature of our ring model is not well-suited to studying second-order and greater synaptic effects (e.g., J_{IE}, J_{II}). Beyond first-order connections the network smooths out heterogeneities of the size we considered here making expression of the relationship between different metrics difficult to discern. A larger, spatially extended model which can better cope with large variability across all weights may be a better test of the extent to which second-order and higher connections affect our results.

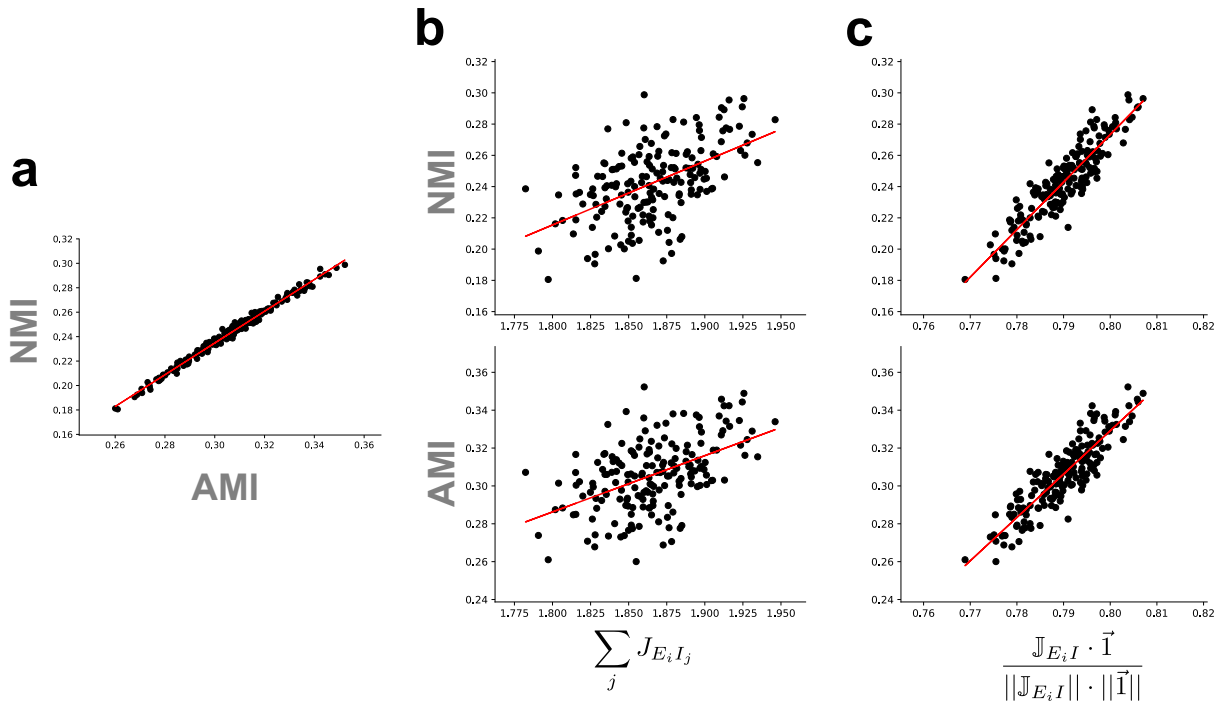

Figure 18: *Naïve analysis of inhibitory connectivity effects* **a** NMI and AMI are strongly correlated. **b** The unweighted sum of inhibition to an E unit correlates with both NMI (top) and AMI (bottom). **c** The normalized sum of inhibition to an E unit (cosine angle with the constant vector) is more strongly correlated with both NMI and AMI. Parameters are as in Figure 14. Red lines are linear fits to data.

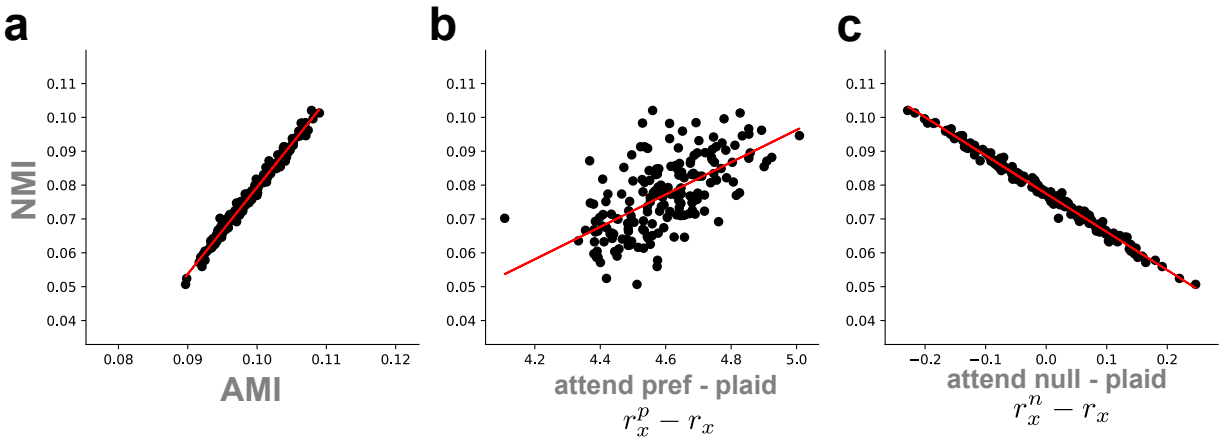

Figure 19: *Anticorrelated J_{EE} satisfies constraints* **a** NMI and AMI. **b** Change in rates for attention to the preferred stimulus location relative to the plaid vs NMI **c** Same as (b) for attention to the null stimulus location. Note many values are positive. Red lines show linear fits to the data.

4.0 Conclusions

The brain displays an incredible ability to perform myriad complex actions subserving an organism's chosen behaviors. In service of these goals, the central nervous system is adjusting itself, or being adjusted by bodily mechanisms, across many spatial and temporal scales. This may seem to suggest amazement at the ability of neurons to encode and represent information in spite of all this variability; seen differently, these various cortical knobs are a positive feature of the system, allowing for the tuning of cortical circuits across varying contexts to subserve the rich repertoire of behavior and cognition available to an organism. The question then is, how do these knobs interact with neural circuits, and how can we understand the resulting effects?

In Chapter 2 we observed that changes in cortical state often accompany changes in an animal's ability to perform an action. In attentional conditions or for moderate levels of arousal, this means an improved sensory perceptual ability. We represented this change in ability with Fisher information (FI), under the assumption that information flow along the cortical hierarchy should relate in a consistent fashion with perceptual capacity. Using linear FI, we then formally showed that, within a circuit, there needs to be a distinction between readout and non-readout cells for information to change due to modulation of a circuit. Reassuringly, this coincides with known anatomy in which cortical cells may be subdivided into locally and long-range projecting subclasses. Since locally-projecting interneurons are largely inhibitory, this analysis further revealed the unexpected result that inhibition is the key modulator of information flow through excitatory subnetworks. Ultimately, this work identified circuit components whose activity needs to be studied across cortical states to understand how information is being affected, and, by extension, how an animal's perceptual capacity is changing. Yet, given the general nature of this work, we could not make further deductions about how a specific state change might be affecting the circuit. For this, we chose to study attention.

Despite decades of intense research, the precise neural mechanisms of attention remain unknown. In Chapter 3 we considered a compelling result in attentional research which linked

attentional effects on cortical neurons to normalization effects of those same cells. Unfortunately, we found that the normalization-attention relationship was not a good constraint on mechanistic models of attention. Instead, we showed that observing the relationship between the absolute change in activity with attention and a cell's degree of normalization better constrained models. This result established the necessity of a circuit framework to link normalization and attention. Constructing a neural network model which obeyed the identified constraints revealed that inhibitory synaptic strengths from distal and proximal sources onto a given E unit need to be anti-correlated to match the data.

An overriding principle which has emerged from the results presented in this work is the central role that inhibition plays in regulating neural activity and information flow in neural circuits as a function of cortical state. Chapter 2 revealed that FI_E formally depended only on the gain and connectivity of the inhibitory cells when considering an excitatory-inhibitory network. Chapter 3 demonstrated the central role inhibition plays in governing the response properties of cortical cells under attention. Why might this be a good solution for the brain, especially given the critical role inhibition plays in preventing excess excitability?

One explanation could come from balanced networks. The theory of *balance* in E/I networks was originally derived to explain asynchronous activity in cortex [148]. In it, inhibition tracks excitation either in a precise, correlated fashion (tight balance) [111] or on a slower timescale with uncorrelated fast fluctuations (loose balance) [148]. When in a tightly balanced regime, networks are capable of highly efficient coding due to the quick cancellation of excitatory signals by inhibition. In fact, tightly balanced, efficient spike-coding networks have been shown capable of outperforming rate coding frameworks [33]. Experimental evidence supports the synaptic scaling required by balanced networks [8], suggesting that computations in the brain depends on inhibitory tracking of excitation.

In Chapter 3 we implemented an SSN model because of its ability to capture the dissociation between expansive neural transfer functions and saturating contrast response functions. A clear computational benefit of this type of transfer function in early sensory systems is the amplification of weak signals. Stabilizing this activity depends on the existence of strong recurrent inhibition which is a hallmark of the SSN framework [5, 118]. The SSN was originally devised to explain normalization through circuit mechanisms. In this regard, inhibition

stabilized (ISN) regimes inherit the computational benefits of normalization, such as scaling the dynamic range of firing rates [59] and removing statistical dependencies in natural signals [127]. Additionally, while theoretically justified, there is growing experimental support that neocortex lives in an ISN regime [3, 125].

Given the central role of inhibition in our results, it is important to point out that there are many inhibitory interneuron subtypes in the brain [143]. One natural extension of the present work would be to determine to what extent these different inhibitory species play a role in the results we have described. A previous network-style model of attention argued for the role of multiple inhibitory interneuron subtypes to explain attentional data; one subtype received top-down projections, the other was more strongly coupled to the excitatory units in a feedforward manner [18]. The top-down subtype was the target of feature attention, whereas the feedforward subtype was affected by spatial attention. This dichotomy is perhaps related to somatostatin-positive (SST) and parvalbumin-positive (PV) subtypes. The former largely targets pyramidal cell dendrites and plays a more modulatory role whereas PV synapses more proximally to the soma and appears to be a key stabilizer of excitatory activity [13].

Another class of dendrite-targeting inhibitory interneurons expresses vasoactive intestinal peptide (VIP). These neurons also inhibit SST cells. It has been shown that VIP cells in mouse V1 are strongly activated by cholinergic projections during locomotion. Their activation leads to inhibition of SST cells, thereby removing a source of inhibition from pyramidal neurons [45]. In this way, VIP cells play a central role in modulating the gain of excitatory cells during changes in arousal state. Yet later work complicated this simple picture by exploring a richer space of visual stimuli. It was found that locomotion affects VIP and SST cells within V1 in a nonlinear, stimulus-dependent fashion, suggesting further studies are needed to clarify the computational role of these species across arousal states [34].

Overall what these examples makes clear is that different classes of inhibitory interneuron play an important and complex role in affecting neural activity across cortical states. The key findings of this dissertation highlighted the important role of inhibition in relating cortical state to behavior. Understanding the neural basis of cognition will depend on our ability to

understand inhibition.

Appendix Mathematical techniques and derivations

We include some mathematical details and derivations underlying the machinery used in the results. Much of it is by now quite standard, but collected here for completeness and coherence.

A.1 Firing rate models

In this section we review derivations of classical firing rate models illustrating two different views of the model: a heuristic derivation from spiking activity, and a derivation based on averaging activity across a homogeneous population. We can think of the former as what we have in mind in our single-unit-like model in Chapter 3, and the latter as representative of the network models in Chapters 2 and 3.

A.1.1 Heuristic derivation from spikes

The unit of information encoded by neurons is the action potential, or spike, which, given their fast timescale, can be characterized as discrete events at the times which they occur. A neuron's activity can thus be represented as a series of delta functions, capturing its activity in terms of a spike count up to some time point t:

$$n(t) = \sum_i \delta(t - t_i). \tag{53}$$

While exact, this representation is not often useful. A temporally discrete representation is that of spike counts, in which time is discretized and spikes are summed within fixed time intervals. The spike count sc over an interval T ending at time t would thus be given by

$$sc(t) = \sum_{t-T<t_i\leq t} \delta(t - t_i). \tag{54}$$

Alternatively, one may construct a continuous variable $r(t)$ to be the firing rate, capturing the underlying frequency with which a neuron is emitting spikes. Following Abbott [2], the time-continuous variable $r(t)$ may be defined as the convolution of the spike times with some kernel $K(t)$:

$$r(t) = \int_{-\infty}^{t} K(t - t')n(t')dt'. \tag{55}$$

The choice of kernel will depend on the assumptions about the nature of a neuron's firing rate. A common choice for $K(t)$ is the exponential function $K(t) = \frac{1}{\tau}e^{-t/\tau}$ [2]; this function defines a low-pass filter over the spikes with decay constant τ. If one assumes that the timescale of $K(t)$ relative to the neural and network dynamics is slow, the measured rate will relax to an approximate steady-state rate for a reasonably slowly-varying stimulus [2]. In this way we can replace $n(t - t')$ with a response function $h(\cdot)$ of the stimulus $s(t)$, such that equation 55 becomes

$$r(t) = \int_{-\infty}^{t} K(t - t')h(s(t'))dt' \tag{56}$$

which, using the exponential kernel, can be written in a differential form

$$\tau\frac{dr(t)}{dt} = -r(t) + h(s(t)). \tag{57}$$

A neuron i within a network will receive inputs from other connected cells j through connections of strength J_{ij}. Assuming inputs to a cell incorporate linearly, the total input from all other neurons in a network is thus given by

$$a_i = \sum_j J_{ij} \int_{-\infty}^{t} K(t - t')n_j(t')dt'. \tag{58}$$

With the same choice of kernel, the activity of a cell now becomes

$$\tau\frac{dr_i(t)}{dt} = -r_i(t) + h_i(s(t)) + \sum_j J_{ij}r_j(t). \tag{59}$$

This equation is more typically written in an analogous form in terms of the postsynaptic voltage v. The rate can then be thought of as a filtered version of v. In this context, it is the presynaptic rates which would affect the postsynaptic voltage. Hence we define the function

$f(v) = r$ as the input-output function governing the relationship between a cell's membrane potential and its firing rate. Then we can write

$$\tau \frac{dv_i(t)}{dt} = -v_i(t) + h_i(s(t)) + \sum_j J_{ij} f_j(v_j(t)). \tag{60}$$

Alternatively (and more consistent with the original work of Wilson and Cowan [153]), one may incorporate the network input to a neuron inside the input-output function f:

$$\tau \frac{dr_i(t)}{dt} = -r_i(t) + f_i\left(\sum_j J_{ij} r_j(t) + h_i(s(t))\right). \tag{61}$$

This is perhaps the more common form of a rate equation, and the one we use in the majority of this thesis. Despite their mathematical equivalence (see: Miller and Fumarola [95]), these two forms may represent different assumptions about the dominant timescales of the dynamics if one is a bit more careful about distinguishing them. In this regard we highlight an argument by Ermentrout and Terman [38], and give the kernel K some biophysical significance as the response of a passive membrane to an input. Writing V instead of K we have $\tau_m \dot{V} + V = I(t)$ (we have taken the membrane resistance $R = 1$ for simplicity), we take the input $I(t)$ to be a decaying exponential with time constant τ_d. This input could capture presynaptic activity with commensurate transmitter release, for example. Taking $V(0) = 0$ the solution is then $V = \frac{\tau_d}{\tau_d - \tau_m}(e^{-t/\tau_d} - e^{-t/\tau_m})$. Now in principle τ_d could vary across input cells, which would require one to define a unique V_{ij} for each neuron pair. Instead, we make one of two assumptions. If the postsynaptic cell alone determines the response then V is independent of j. Consistently, if $\tau_m >> \tau_d$ then we have the approximation $V \approx \frac{1}{\tau_m} e^{-t/\tau_m}$ which is independent of the input parameter τ_d. One then arrives at equation 60 (see Ermentrout and Terman [38] for details).

By contrast suppose that the response of the postsynaptic cell is dictated only by the presynaptic cell. Then by a similar argument, if we assume the timescale $\tau_m << \tau_d$ we can approximate the potential response $V \approx \frac{1}{\tau_d} e^{-t/\tau_d}$. From this we get equation 61 (with $\tau \mapsto \tau_d$) in that presynaptic inputs are combined before being converted to a postsynaptic rate through the function f.

A.1.2 Markov process derivation

Here we follow an approach taken by Bressloff [14]. Consider a homogeneous population of neurons each of which can be in one of two states: active or silent (i.e., spiking or not spiking). Let n define the number of active neurons in the population of size $N \geq n$. For some small time interval dt define transition rates $n \to n-1 : T_-(n) = n$ and $n \to n+1 : T_+(n) = Nf(n/N)$. Thus the change in probability P of n active units at time t is given by

$$\frac{d}{dt}P(n,t) = T_+(n-1)P(n-1,t) + T_-(n+1)P(n+1,t) - (T_+(n) + T_-(n))P(n,t) \quad (62)$$

with boundary condition $P(-1,t) = 0$ (activity can only increase if no cells are firing). We now define the firing rate at time t as the expected value of the activity over the population:

$$r(t) = E[n/N] = \sum_{n=0}^{N} \frac{n}{N} P(n,t). \quad (63)$$

Therefore letting $N \to \infty$ and dropping the t for notational convenience,

$$\sum_{n=0}^{\infty} n\frac{d}{dt}P(n) = \sum_{n=0}^{\infty} n[T_+(n-1)P(n-1) + T_-(n+1)P(n+1) - (T_+(n) + T_-(n))P(n)] \quad (64)$$

$$= \sum_{n=0}^{\infty} n[(n+1)P(n+1) - nP(n)] \quad (65)$$
$$+ \sum_{n=0}^{\infty} n[Nf((n-1)/N)P(n-1) - Nf(n/N)P(n)]$$
$$= \sum_{n=0}^{\infty} n(n+1)P(n+1) - \sum_{n=0}^{\infty} n^2P(n) \pm \sum_{n=0}^{\infty}(n+1)P(n+1) \quad (66)$$
$$+ \sum_{n=0}^{\infty} nNf((n-1)/N)P(n-1) - \sum_{n=0}^{\infty} nNf(n/N)P(n)$$
$$\pm \sum_{n=0}^{\infty} Nf((n-1)/N)P(n-1)$$

$$= \sum_{n=0}^{\infty}(n+1)^2P(n+1) - \sum_{n=0}^{\infty}n^2P(n) - \sum_{n=0}^{\infty}(n+1)P(n+1)$$
$$+ \sum_{n=0}^{\infty}(n-1)Nf((n-1)/N)P(n-1) - \sum_{n=0}^{\infty}nNf(n/N)P(n) \quad (67)$$
$$+ \sum_{n=0}^{\infty}Nf((n-1)/N)P(n-1)$$

$$= -\sum_{n=0}^{\infty}nP(n) + \sum_{n=0}^{\infty}Nf(n/N)P(n) \quad (68)$$

$$= -E[n] + E[Nf(n/N)]. \quad (69)$$

Assuming that the population is in a stable asynchronous state, we interchange expectations with f [14] to arrive at the desired equation:

$$\frac{d}{dt}r(t) = \frac{d}{dt}E[n] = -E[n] + f(E[n]) = -r(t) + f(r(t)). \quad (70)$$

The extension to M interacting homogenous populations of size N_k, $k = 1, ..., M$ is very similar [15, 14], with a vector now denoting the number of active neurons within one of the M populations: $\vec{n} = (n_1, ..., n_M)$. Let w_{ij} connote the strength of interaction between population j and i. Define the transition rates to be $T_{k-}(\vec{n}) = n_k, T_{k+}(\vec{n}) = Nf(\sum_j w_{ij}n_j/N + h_k)$ where we have also now explicitly included external inputs h_k to population k. The probability distribution again evolves similarly to equation 62:

$$\frac{d}{dt}P(\vec{n},t) = \sum_{k=1}^{M}[T_+(\vec{n}-e_k)P(\vec{n}-e_k,t) + T_-(\vec{n}+e_k)P(\vec{n}+e_k,t) - (T_+(\vec{n}) + T_-(\vec{n}))P(\vec{n},t)] \quad (71)$$

where e_k is the M-dimensional vector whose k^{th} element equals one with all other elements zero (or: the k^{th} vector of the standard basis in M dimensions). The modified boundary condition is $P(\vec{n},t) = 0$ if any $n_i = -1$. This then leads to the master equation

$$\frac{d}{dt}P(\vec{n},t) = \sum_{k=1}^{M}[Nf\big(\sum_j w_{ij}(n_j + \delta_{ij})/N + h_k\big)P(\vec{n} - e_k) + (\vec{n_k} + 1)P(\vec{n} + e_k)$$
$$-(Nf\big(\sum_j w_{ij}n_j/N + h_k\big) + \vec{n_k})P(\vec{n})] \quad (72)$$

where δ_{ij} is the Kronecker delta. By again taking expectations as described above we arrive at the more general form of the rate equations:

$$\frac{d}{dt}r_i(t) = -r_i(t) + f\left(\sum_j w_{ij}r_j(t) + h_i(t)\right). \tag{73}$$

A.2 Linear theory of stochastic dynamics

A.2.1 Linear stability analysis in a deterministic system

Consider a dynamical system of the form $\dot{x} = f(x)$ with a fixed point at x^*. A linear approximation of the system near the fixed point is given by $\frac{d}{dt}(x - x^*) = \dot{x} \approx f'(x^*)(x - x^*)$. In Chapter 2 we used the Routh-Hurwitz criterion (RHC) to analyze the stability of the $E_1/E_2/I$ network which is valid for a linear system. Using the linearized form of the network response, the RHC allowed us to rewrite the stability criteria in terms of the variables we were surfing over [94], namely the effective connectivity weights w_{ij}. For an equation of the form $\dot{x} = Wx + I$ where $x \in \mathbb{R}^3$, $W \in \mathbb{R}^3 \times \mathbb{R}^3$, write the characteristic polynomial $p(\lambda) = \det(I\lambda - W) := a_0\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0$ where I is the identity matrix. The RHC says that $p(\lambda)$ will have all zeros in the left half plane if all coefficients $a_i > 0$ and if, for this 3×3 case, $a_1a_2 > a_0a_3$.

A.2.2 Stochastic processes in one dimension

Consider a random variable X, samples of which are drawn at time points $t_1, t_2, ...$ Then $Y(t) = f(X, t)$ defines a stochastic process, that is, a function of the random variable X and t, which here we take to be time.

Brownian motion (also known as white noise) is defined as a stochastic process $w(X, t)$ continuous in time in which for each time step δt the evolution of the state variable w: is Gaussian with mean 0 and variance δt, is independent of the previous step, and has initial condition $w(X, 0) = 0$ for all X. This process is nowhere differentiable. To see this, following

Chorin and Hald [23] we define the variable

$$dw = \frac{w(X, t+\delta t) - w(X,t)}{\delta t}. \tag{74}$$

By the definition of Brownian motion dw is Gaussian with expectation $\langle dw \rangle = 0$ and variance $\mathrm{Var}(dw) = \mathrm{Var}\left(\frac{w(X,t+\delta t)-w(X,t)}{\delta t}\right) = \frac{1}{\delta t^2} \cdot \mathrm{Var}(w(X,t+\delta t) - w(X,t)) = \frac{1}{\delta t}$. Hence $\mathrm{Var}(dw) \to \infty$ as $\delta t \to 0$ and thus for a fixed X the derivative exists nowhere. In what follows we will use the expression dw to refer to the increment of Brownian motion.

A stochastic differential equation (SDE) is, informally, a differential equation together with a stochastic part. Consider an SDE of the form

$$\tau_x dx = -x(t)dt + \sigma dw \tag{75}$$

where w is Brownian motion (Wiener process). This is the Langevin equation in physics. Integrating through from 0 to t gives

$$x(t) - x(0) = \frac{1}{\tau_x} \int_0^t x(s)ds + \frac{\sigma}{\tau_x} \int_0^t dw. \tag{76}$$

The first expression $\int_0^t dw$ is called a stochastic integral. Again following Chorin and Hald [23], the way to evaluate a stochastic integral of the form $\int_a^b g(t)dw$ for an arbitrary function $g(t)$ is to partition the interval of integration into discrete steps $a = t_0 < t_1 < \cdots < t_n = b$ and approximate $g(t)$ with a piecewise-constant function:

$$\int_a^b g(t)\tau dw \approx \sum_{i=0}^{n-1} \gamma_i (w(t_{i+1} - w(t_i)) \tag{77}$$

where γ_i is a constant. One then takes the limit as the width of the largest interval in the partition goes to zero. It remains to determine how to choose γ_i. The Ito solution evaluates γ_i at the left end of the interval, such that $\gamma_i = g(t_i)$. Alternatively, the Stratonovich solution takes the average of the values at the endpoints of the interval: $\gamma_i = \frac{1}{2}[g(t_i) + g(t_{i+1})]$. In general these interpretations need not agree [23], but because $g(t)$ is constant throughout this work, there is no concern.

An alternative derivation given by Lindner [81] sidesteps these issues by making use of the assumptions on w. We write the Langevin equation less formally as

$$\tau_x \dot{x} = -x(t) + \sigma w(t). \tag{78}$$

In the absence of noise this is a linear, first-order ordinary differential equation (ODE) whose solution is an exponentially decaying process with time constant $\tau : x(t) = x_0 e^{-t/\tau_x}$. Hence,

$$x(t) = x_0 e^{-t/\tau_x} + \frac{\sigma}{\tau_x} \int_0^t e^{-(t-s)/\tau_x} w(s) ds. \tag{79}$$

The stationary mean value can be calculated by first recalling that $E[w(t)] = 0$. Then $E[x(t)]_{t\to\infty} = \lim_{t\to\infty} x_0 e^{-t/\tau_x} + \frac{\sigma}{\tau_x} \int_0^t e^{-(t-s)/\tau_x} \langle w(s) \rangle = \lim_{t\to\infty} x_0 e^{-t/\tau_x} = 0$, where we let $t \to \infty$ to remove any dependence on initial conditions.

In order to measure the variability appropriately, we use the autocorrelation function, given by $G(\tau) = E[x(t)x(t+\tau)]$. Because neural data acquired in experiments is often pooled across trials and summed over discrete time windows which are long relative to the intrinsic timescale of neural dynamics, we compute the integrated variance over all time $\int_{-\infty}^{\infty} G(\tau) d\tau$. We again follow the exposition in Lindner [81]. Since a white noise process evolves in independent increments, $E[w(t)w(t+\tau)] = \delta(\tau)$ where $\delta(\tau)$ is the Dirac delta function. From the exposition following equation 79 we see that in the $t \to \infty$ limit the expectation $E[x(t)x(t+\tau)]$ will only retains terms which don't involve the initial condition term. Hence,

$$E[x(t)x(t+\tau)] = \left(\frac{\sigma}{\tau_x}\right)^2 \int_0^t e^{-(t-s)/\tau_x} ds \int_0^{t+\tau} e^{-(t+\tau-s')/\tau_x} E[w(s)w(s')] ds' \tag{80}$$

$$= \left(\frac{\sigma}{\tau_x}\right)^2 \int_0^t e^{-(t-s)/\tau_x} ds \int_0^{t+\tau} e^{-(t+\tau-s')/\tau_x} \delta(s'-s) ds' \tag{81}$$

$$= \left(\frac{\sigma}{\tau_x}\right)^2 \int_0^t e^{-(2t+\tau-2s)/\tau_x} ds \tag{82}$$

$$= \left(\frac{\sigma}{\tau_x}\right)^2 \frac{\tau_x}{2} [e^{-\tau/\tau_x} - e^{-(2t+\tau)/\tau_x}] \xrightarrow[t\to\infty]{} \frac{\sigma^2}{2\tau_x} e^{-\tau/\tau_x}. \tag{83}$$

Carrying out the same calculations for negative time lags $E[x(t-\tau)x(t)]$ gives the same result, such that we can arrive at

$$G(\tau) = \frac{\sigma^2}{2\tau_x} e^{-|\tau|/\tau_x} \tag{84}$$

from which we can compute the integral over all time lags.

A.2.3 N-dimensional stochastic process

Such that we may model networks of neurons we extend the previous concepts to N-dimensions. This section follows Gardiner [46]. Therefore consider a network of N units (or, homogeneous populations) whose (possibly linearized) dynamics are given formally by

$$d\vec{x}(t) = -M\vec{x}(t)dt + DdW \tag{85}$$

where $\vec{x} \in \mathbb{R}^n, M, D \in \mathbb{R}^n \times \mathbb{R}^n$ are constant and dW is an n-dimensional Wiener process. Assume that the system is stable and stationary (that is, $E[\vec{x}(t)] = 0$).

The main object we want to calculate is the covariance matrix. The solution to these N SDEs is the N-dimensional analogue of equation 76:

$$\vec{x}(t) = exp(-Mt)\vec{x}(0) + \int_{-\infty}^{t} exp(-M(t-t'))DdW. \tag{86}$$

We can compute the covariance matrix at zero time lag, $E[\vec{x}(t)\vec{x}^T(t)] = \sigma$ as follows. Write

$$\begin{aligned}
M\sigma + \sigma M^T &= \int_{-\infty}^{t} Mexp(-M(t-t'))DD^T exp(-M^T(t-t'))dt' && (87)\\
&\quad + \int_{-\infty}^{t} exp(-M(t-t'))DD^T exp(-M^T(t-t'))M^T dt' && (88)\\
&= \int_{-\infty}^{t} \frac{d}{dt'}[exp(-M(t-t'))DD^T exp(-M^T(t-t'))]dt' && (89)\\
&= DD^T - \lim_{t'\to-\infty}[exp(-M(t-t'))DD^T exp(-M^T(t-t'))] && (90)\\
&= DD^T && (91)
\end{aligned}$$

where the last line follows by the assumption that the system is stable such that the limit vanishes. Ultimately we want the integrated covariance at all time lags, as discussed previously, so similarly to the 1-dimensional case we define the cross-covariance $G(\tau) = E[\vec{x}(t)\vec{x}^T(t+\tau)]$ where for $\tau < 0$,

$$\begin{aligned}
G(\tau) &= \int_{-\infty}^{t+\tau} exp(-M(t-t'))DD^T exp(-M^T(t+\tau-t'))dt' && (92)\\
&= exp(-M\tau)\int_{-\infty}^{t+\tau} exp(-M(t+\tau-t'))DD^T exp(-M^T(t+\tau-t'))dt' && (93)\\
&= exp(-M\tau)\sigma && (94)
\end{aligned}$$

and for $\tau > 0$,

$$G(\tau) = \int_{-\infty}^{t} exp(-M(t-t'))DD^T exp(-M^T(t+\tau-t'))dt' \tag{95}$$

$$= \left[\int_{-\infty}^{t} exp(-M(t-t'))DD^T exp(-M^T(t-t'))dt'\right] exp(-M^T\tau) \tag{96}$$

$$= \sigma exp(-M^T\tau) \tag{97}$$

Now we could stop here and integrate over τ to get the full covariance. But with a little more work, a much simpler and more intuitive expression results. We now apply the Wiener-Khinchin theorem to compute the spectrum matrix from the cross-correlation function computed above:

$$S(\omega) = \int_{-\infty}^{\infty} exp(-i\omega\tau)G(\tau)d\tau \tag{98}$$

$$= \int_{0}^{\infty} \sigma exp(-(i\omega + M^T)\tau)d\tau + \int_{-\infty}^{0} exp((-i\omega + M)\tau)\sigma d\tau \tag{99}$$

$$= (M - i\omega)^{-1}\sigma + \sigma(M^T + i\omega)^{-1} \tag{100}$$

Multiplying through by $M - i\omega$ on the left and $M^T + i\omega$ on the right, we have

$$(M - i\omega)S(\omega)(M^T + i\omega) = M\sigma + \sigma M^T. \tag{101}$$

Substituting the result in equation 91 and rearranging we have that the spectrum matrix in stationary state becomes

$$S(\omega) = \frac{1}{2\pi}(M - i\omega)^{-1}DD^T(M^T + i\omega)^{-1} \tag{102}$$

Using the fact that the long-time covariance can be computed by evaluating the spectrum at zero frequency we get the that the N-dimensional covariance matrix is given by

$$\Sigma = M^{-1}DD^T(M^T)^{-1} = M^{-1}D(M^{-1}D)^T \tag{103}$$

which is the expression we have used throughout this work.

Recall that in Chapters 2, 3 we were interested in neurons with nonlinear transfer functions. Fluctuations in the voltage will thus be scaled nonlinearly into fluctuations in the

rates [1]. Since in this work we were interested only in steady state responses, our system permitted linearization of the rates as described previously. In order to make use of the foregoing theory, we needed to assume that the noise in the system was sufficiently small as to not perturb the system to a new steady state. In this way we can write down an equation of the form

$$\delta\dot{r}_{ss} = -\delta r_{ss}(t) + L(J\delta r_{ss}(t) + I_\xi(t)) \tag{104}$$

where L is a diagonal matrix of the linearization terms and J is the weight matrix, as defined in Chapter 2; I_0 is a signal input, $I_\xi(t) := \Sigma_D dW$ is some noisy input, and $\delta r_{ss}(t) = r(t) - r_{ss}$ describes the evolution of r near the point $r_{ss} = (1 - LJ)^{-1}LI_0$, obtained by solving the deterministic system $\dot{r} = -r(t) + L(Jr(t) + I_0) = 0$. With the substitutions $1 - LJ \mapsto M$ and $L\Sigma_d \mapsto D$ we recover equation 85 from equation 104.

A.3 Information-theoretic analyses

Given its noisiness, neural activity can be described as a conditional probability density in firing rate space, conditioned on a variable of interest. In the present case our variable of interest is a stimulus. For a population of N neurons we define *rate space* as the N-dimensional space of positive reals $\mathbb{R}_+^N$ with a point in this space determined by the firing rates of each neuron at a given time or under a particular condition. Repeated samples of points over trials across a fixed stimulus presentation therefore define the probability distribution of the rates $\vec{r}$ conditioned on that stimulus θ, $P(\vec{r}|\theta)$. As discussed in section 1.4, an abstraction of an animal's behavioral capacity on a discrimination task is the decodability of upstream population activity (carrying signal information) by downstream cortical regions (mediating active report). In this way downstream regions must solve an estimation problem over the upstream activity.

An estimator is a function $T : X^n \to \Theta$ from the sample space of n observations to the parameter set Θ [27]. We will mostly concern ourselves with linear estimators such that the

[1] In performing simulations, we consider a low-pass filtered noise term $\tau_{noise}\dot{\xi} = -\xi(t) + \sigma_{noise}dW$ such that the equation for x is now $\tau\dot{x} = -x(t) + \sigma\xi$. In the limit as $\tau_{noise} \to 0$ this is exactly a Wiener process and hence the preceding results hold. Practically, τ_{noise} should be sufficiently small relative to τ.

function T can be written as a weighted sum over the n observations x_i (see below). If θ is the true parameter to be estimated (say, a stimulus orientation) and $\hat{\theta}$ is the estimated value (that is, the image of $T(X_1, ..., X_n)$) then the *error* of the estimate is given by $\hat{\theta} - \theta$ and the *bias* is defined $E[\hat{\theta} - \theta]$. An unbiased estimator is one in which this expectation is zero. It turns out (for proof see, e.g., Cover and Thomas [27]) that a lower bound on the mean squared error of an unbiased estimator is given by the reciprocal of the Fisher information (FI) $I_X(\theta)$:

$$var(T) := E[(\hat{\theta} - E[\hat{\theta}])^2] \geq \frac{1}{I_X(\theta)}. \tag{105}$$

FI is defined for a continuous variable θ as

$$I_X(\theta) = -\int P(\vec{r}|\theta) \frac{\partial^2}{\partial \theta^2} \log P(\vec{r}|\theta) d\vec{r} \tag{106}$$

We now make two observations which demonstrate the relevance of this result. First, we show the relationship between FI and a measure of discriminability, which establishes the utility of FI in estimating behavioral capacity. Second, we describe how a linear component of FI (linear FI) provides a biologically plausible instantiation through synaptic weights while providing a reasonable approximation to the full FI.

Consider a fine-discrimination task in which an animal needs to differentiate between two stimuli, θ and $\theta + \Delta\theta$ where $\Delta\theta$ is a small shift along some stimulus dimension. Assume a neural population response to each stimulus is gaussian with mean responses μ_1, μ_2 and equivalent variance σ^2. Then the ability to tell these two distributions apart can be captured by the discriminability [30]

$$d' = \frac{\mu_2 - \mu_1}{\sigma}. \tag{107}$$

For an unbiased estimator, the expected difference in the means $\mu_2 - \mu_1 = \Delta\theta$. Moreover, if the decoder is optimal, its variance is given by equation 105, with equality. Hence, we can write the discriminability as

$$d' = \Delta\theta \sqrt{I_X(\theta)}. \tag{108}$$

For an N-dimensional population, if we assume the response to a stimulus θ is gaussian with mean $\vec{r}$ and covariance Σ then equation 106 becomes [1]

$$I_X(\theta) = f'(\theta)^T \Sigma(\theta)^{-1} f'(\theta) + \frac{1}{2} \mathrm{Tr}\left(\Sigma'(\theta)\Sigma^{-1}(\theta)\Sigma'(\theta)\Sigma^{-1}(\theta)\right) \tag{109}$$

where $f(\theta)$ is the population response to θ and Σ is the population covariance. We refer to this expression as the full FI, while the first term on the right-hand side is the linear FI. If the covariance matrix is independent of the stimulus, then the full FI exactly equals the linear FI. More generally, Seriés *et al.* [131] used a locally optimal linear decoder to bound estimates of FI. That is, they trained a decoder to estimate the orientations of two bars oriented $\Delta\theta = 1^o$ apart in a network simulation of early visual cortex. Thus for a given set of output rates $\vec{r}$ they estimated the angle $\hat{\theta} = \vec{w}^T\vec{r} + b$ where $\vec{w}$ and b are the parameters to be fit. This is also a reasonable instantiation of a downstream neuron attempting to estimate the stimulus angle from synaptic inputs, with the synaptic strengths represented by $\vec{w}$, the weights over the presynaptic activity. The linear FI was then approximated from the means $\hat{\theta}_1, \hat{\theta}_2$ and variances σ_1^2, σ_2^2 of the estimates:

$$I_{estimated} = \frac{((\hat{\theta}_2 - \hat{\theta}_1)/\Delta\theta)^2}{\sqrt{(\sigma_1^2)^2 + (\sigma_2^2)^2}}. \tag{110}$$

The authors found that the majority of nonlinear methods did not outperform this estimate in the context of their model, suggesting that linear FI is an appropriate representation of the system's information.

Bibliography

[1] Larry F Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

[2] LF Abbott. Decoding neuronal firing and modelling neural networks. *Quarterly reviews of biophysics*, 27(3):291–331, 1994.

[3] Hillel Adesnik. Synaptic mechanisms of feature coding in the visual cortex of awake mice. *Neuron*, 95(5):1147–1159, 2017.

[4] Hillel Adesnik, William Bruns, Hiroki Taniguchi, Z Josh Huang, and Massimo Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–231, 2012.

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

[6] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006.

[7] Dylan Barbera, Nicholas J Priebe, and Lindsey L Glickfeld. Feedforward mechanisms of cross-orientation interactions in mouse v1. *Neuron*, 110(2):297–311, 2022.

[8] Jérémie Barral and Alex D Reyes. Synaptic scaling rule preserves excitatory–inhibitory balance and salient neuronal network dynamics. *Nature neuroscience*, 19(12):1690–1696, 2016.

[9] Jeffrey Beck, Vikranth R Bejjanki, and Alexandre Pouget. Insights from a simple expression for linear fisher information in a recurrently connected population of spiking neurons. *Neural computation*, 23(6):1484–1502, 2011.

[10] Rani Ben-Yishai, R Lev Bar-Or, and Haim Sompolinsky. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences*, 92(9):3844–3848, 1995.

[11] Corbett Bennett, Sergio Arroyo, and Shaul Hestrin. Controlling brain states. *Neuron*, 83(2):260–261, 2014.

[12] Richard T Born and David C Bradley. Structure and function of visual area mt. *Annu. Rev. Neurosci.*, 28:157–189, 2005.

[13] Hannah Bos, Anne-Marie Oswald, and Brent Doiron. Untangling stability and gain modulation in cortical circuits with multiple interneuron classes. *bioRxiv*, 2020.

[14] Paul C Bressloff. Metastable states and quasicycles in a stochastic wilson-cowan model of neuronal population dynamics. *Physical Review E*, 82(5):051903, 2010.

[15] Paul C Bressloff. Stochastic neural field theory and the system-size expansion. *SIAM Journal on Applied Mathematics*, 70(5):1488–1521, 2010.

[16] Kenneth H Britten, Michael N Shadlen, William T Newsome, and J Anthony Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–4765, 1992.

[17] Emery N Brown, Patrick L Purdon, and Christa J Van Dort. General anesthesia and altered states of arousal: a systems neuroscience analysis. *Annual review of neuroscience*, 34:601, 2011.

[18] Calin I Buia and Paul H Tiesinga. Role of interneuron diversity in the cortical microcircuit for attention. *Journal of neurophysiology*, 99(5):2158–2182, 2008.

[19] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[20] Matteo Carandini, David J Heeger, and J Anthony Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644, 1997.

[21] Marisa Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[22] James R Cavanaugh, Wyeth Bair, and J Anthony Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of neurophysiology*, 88(5):2530–2546, 2002.

[23] Alexandre Joel Chorin and Ole H Hald. *Stochastic tools in mathematics and science*, volume 3. Springer, 2009.

[24] Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600, 2009.

[25] Marlene R Cohen and John HR Maunsell. Using neuronal populations to study the mechanisms underlying spatial and feature attention. *Neuron*, 70(6):1192–1204, 2011.

[26] Albert Compte and Xiao-Jing Wang. Tuning curve shift by attention modulation in cortical neurons: a computational study of its mechanisms. *Cerebral Cortex*, 16(6):761–778, 2006.

[27] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[28] Benjamin R Cowley, Adam C Snyder, Katerina Acar, Ryan C Williamson, M Yu Byron, and Matthew A Smith. Slow drift of neural activity as a signature of impulsivity in macaque visual and prefrontal cortex. *Neuron*, 108(3):551–567, 2020.

[29] Rajan Dasgupta, Frederik Seibt, and Michael Beierlein. Synaptic release of acetylcholine rapidly suppresses cortical activity by recruiting muscarinic receptors in layer 4. *Journal of Neuroscience*, 38(23):5338–5350, 2018.

[30] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

[31] Jaime De La Rocha, Brent Doiron, Eric Shea-Brown, Krešimir Josić, and Alex Reyes. Correlation between neural spike trains increases with firing rate. *Nature*, 448(7155):802–806, 2007.

[32] Sophie Deneve, Peter E Latham, and Alexandre Pouget. Reading population codes: a neural implementation of ideal observers. *Nature neuroscience*, 2(8):740–745, 1999.

[33] Sophie Denève and Christian K Machens. Efficient codes and balanced networks. *Nature neuroscience*, 19(3):375–382, 2016.

[34] Mario Dipoppa, Adam Ranson, Michael Krumin, Marius Pachitariu, Matteo Carandini, and Kenneth D Harris. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron*, 98(3):602–615, 2018.

[35] Brent Doiron, Ashok Litwin-Kumar, Robert Rosenbaum, Gabriel K Ocker, and Krešimir Josić. The mechanics of state-dependent neural correlations. *Nature neuroscience*, 19(3):383–393, 2016.

[36] Alexander S Ecker, Philipp Berens, R James Cotton, Manivannan Subramaniyan, George H Denfield, Cathryn R Cadwell, Stelios M Smirnakis, Matthias Bethge, and Andreas S Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.

[37] Alexander S Ecker, George H Denfield, Matthias Bethge, and Andreas S Tolias. On the structure of neuronal population activity under fluctuations in attentional state. *Journal of Neuroscience*, 36(5):1775–1789, 2016.

[38] Bard Ermentrout and David H Terman. *Mathematical foundations of neuroscience*, volume 35. Springer, 2010.

[39] Sarah R Erwin, Brianna N Bristow, Kaitlin E Sullivan, Rennie M Kendrick, Brian Marriott, Lihua Wang, Jody Clements, Andrew L Lemire, Jesse Jackson, and Mark S Cembrowski. Spatially patterned excitatory neuron subtypes and projections of the claustrum. *Elife*, 10, 2021.

[40] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

[41] Katie A Ferguson and Jessica A Cardin. Mechanisms underlying gain modulation in the cortex. *Nature Reviews Neuroscience*, 21(2):80–92, 2020.

[42] Ian C Fiebelkorn and Sabine Kastner. A rhythmic theory of attention. *Trends in cognitive sciences*, 23(2):87–101, 2019.

[43] Ian C Fiebelkorn and Sabine Kastner. Functional specialization in the attention network. *Annual review of psychology*, 71:221, 2020.

[44] Alfredo Fontanini and Donald B Katz. Behavioral states, network states, and sensory response variability. *Journal of neurophysiology*, 100(3):1160–1168, 2008.

[45] Yu Fu, Jason M Tucciarone, J Sebastian Espinosa, Nengyin Sheng, Daniel P Darcy, Roger A Nicoll, Z Josh Huang, and Michael P Stryker. A cortical circuit for gain control by behavioral state. *Cell*, 156(6):1139–1152, 2014.

[46] Crispin Gardiner. *Stochastic methods*, volume 4. springer Berlin, 2009.

[47] Apostolos P Georgopoulos, Joseph T Lurito, Michael Petrides, Andrew B Schwartz, and Joe T Massey. Mental rotation of the neuronal population vector. *Science*, 243(4888):234–236, 1989.

[48] Apostolos P Georgopoulos, Andrew B Schwartz, and Ronald E Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, 1986.

[49] Matthew Getz, Chengcheng Huang, Jeffrey Dunworth, Marlene R Cohen, and Brent Doiron. Attentional modulation of neural covariability in a distributed circuit-based population model. *Cosyne Abstracts, Denver, CO*, 2018.

[50] Matthew P Getz, Chengcheng Huang, and Brent Doiron. Subpopulation codes permit information modulation across cortical states. *bioRxiv*, 2022.

[51] Lisa M Giocomo and Michael E Hasselmo. Neuromodulation by glutamate and acetylcholine can change circuit dynamics by regulating the relative influence of afferent input and excitatory feedback. *Molecular neurobiology*, 36(2):184–200, 2007.

[52] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[53] Nathan W Gouwens, Jim Berg, David Feng, Staci A Sorensen, Hongkui Zeng, Michael J Hawrylycz, Christof Koch, and Anton Arkhipov. Systematic generation of biophysically detailed models for diverse cortical neuron types. *Nature communications*, 9(1):1–13, 2018.

[54] David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

[55] Kenneth D Harris and Thomas D Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474):51–58, 2013.

[56] Kenneth D Harris and Gordon MG Shepherd. The neocortical circuit: themes and variations. *Nature neuroscience*, 18(2):170–181, 2015.

[57] Kenneth D Harris and Alexander Thiele. Cortical state and attention. *Nature reviews neuroscience*, 12(9):509–523, 2011.

[58] Michael E Hasselmo. Neuromodulation and cortical function: modeling the physiological basis of behavior. *Behavioural brain research*, 67(1):1–27, 1995.

[59] David J Heeger. Normalization of cell responses in cat striate cortex. *Visual neuroscience*, 9(2):181–197, 1992.

[60] Guillaume Hennequin, Yashar Ahmadian, Daniel B Rubin, Máté Lengyel, and Kenneth D Miller. The dynamical regime of sensory cortex: Stable dynamics around a single stimulus-tuned attractor account for patterns of noise variability. *Neuron*, 98(4):846–860, 2018.

[61] Lotte J Herstel and Corette J Wierenga. Network control through coordinated inhibition. *Current Opinion in Neurobiology*, 67:34–41, 2021.

[62] Chengcheng Huang. Modulation of the dynamical state in cortical network models. *Current opinion in neurobiology*, 70:43–50, 2021.

[63] Chengcheng Huang, Douglas A Ruff, Ryan Pyle, Robert Rosenbaum, Marlene R Cohen, and Brent Doiron. Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101(2):337–348, 2019.

[64] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[65] Krešimir Josić, Eric Shea-Brown, Brent Doiron, and Jaime de la Rocha. Stimulus-dependent correlations and population codes. *Neural computation*, 21(10):2774–2804, 2009.

[66] MohammadMehdi Kafashan, Anna Jaffe, Selmaan N Chettih, Ramon Nogueira, Iñigo Arandia-Romero, Christopher D Harvey, Rubén Moreno-Bote, and Jan Drugowitsch. Scaling of information in large neural populations reveals signatures of information-limiting correlations. *bioRxiv*, 2020.

[67] MohammadMehdi Kafashan, Anna W Jaffe, Selmaan N Chettih, Ramon Nogueira, Iñigo Arandia-Romero, Christopher D Harvey, Rubén Moreno-Bote, and Jan Drugowitsch. Scaling of sensory information in large neural populations shows signatures of information-limiting correlations. *Nature communications*, 12(1):1–16, 2021.

[68] Tatjana Kanashiro, Gabriel Koch Ocker, Marlene R Cohen, and Brent Doiron. Attentional modulation of neuronal variability in circuit models of cortex. *Elife*, 6:e23978, 2017.

[69] Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015.

[70] Jessleen K Kanwal, Emma Coddington, Rachel Frazer, Daniela Limbania, Grace Turner, Karla J Davila, Michael A Givens, Valarie Williams, Sandeep Robert Datta, and Sara Wasserman. Internal state: dynamic, interconnected communication loops distributed across body, brain, and time. *Integrative and Comparative Biology*, 61(3):867–886, 2021.

[71] Matthias Kaschube, Michael Schnabel, Siegrid Löwel, David M Coppola, Leonard E White, and Fred Wolf. Universality in the evolution of orientation columns in the visual cortex. *science*, 330(6007):1113–1116, 2010.

[72] Andreas J Keller, Morgane M Roth, and Massimo Scanziani. Feedback generates a second receptive field in neurons of the visual cortex. *Nature*, 582(7813):545–549, 2020.

[73] Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual review of neuroscience*, 39:237–256, 2016.

[74] Adam Kohn, Anna I Jasper, João D Semedo, Evren Gokcen, Christian K Machens, and M Yu Byron. Principles of corticocortical communication: proposed schemes and design considerations. *Trends in Neurosciences*, 43(9):725–737, 2020.

[75] Richard Komuniecki, Vera Hapiak, Gareth Harris, and Bruce Bamber. Context-dependent modulation reconfigures interactive sensory-mediated microcircuits in caenorhabditis elegans. *Current opinion in neurobiology*, 29:17–24, 2014.

[76] John W Krakauer, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017.

[77] Joonyeol Lee and John HR Maunsell. A normalization model of attentional modulation of single unit responses. *PloS one*, 4(2):e4651, 2009.

[78] Joonyeol Lee, Tori Williford, and John HR Maunsell. Spatial attention and the latency of neuronal responses in macaque area v4. *Journal of Neuroscience*, 27(36):9632–9637, 2007.

[79] Wei-Chung Allen Lee, Vincent Bonin, Michael Reed, Brett J Graham, Greg Hood, Katie Glattfelder, and R Clay Reid. Anatomy and function of an excitatory network in the visual cortex. *Nature*, 532(7599):370–374, 2016.

[80] Zhaoping Li. Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. *Neural computation*, 13(8):1749–1780, 2001.

[81] Benjamin Lindner. A brief introduction to some simple stochastic processes. *Stochastic Methods in Neuroscience*, 1, 2009.

[82] Grace W Lindsay, Daniel B Rubin, and Kenneth D Miller. A simple circuit model of visual cortex explains neural and behavioral aspects of attention. *bioRxiv*, 2019.

[83] Grace W Lindsay, Daniel B Rubin, and Kenneth D Miller. A unified circuit model of attention: neural and behavioral effects. *bioRxiv*, pages 2019–12, 2020.

[84] Ashok Litwin-Kumar and Brent Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature neuroscience*, 15(11):1498–1505, 2012.

[85] Simona Lodato, Caroline Rouaux, Kathleen B Quast, Chanati Jantrachotechatchawan, Michèle Studer, Takao K Hensch, and Paola Arlotta. Excitatory projection neuron subtypes control the distribution of local inhibitory interneurons in the cerebral cortex. *Neuron*, 69(4):763–779, 2011.

[86] Cheng Ly, Jason W Middleton, and Brent Doiron. Cellular and circuit mechanisms maintain low spike co-variability and enhance population coding in somatosensory cortex. *Frontiers in Computational Neuroscience*, 6:7, 2012.

[87] Eve Marder. Neuromodulation of neuronal circuits: back to the future. *Neuron*, 76(1):1–11, 2012.

[88] Eve Marder, Timothy O'Leary, and Sonal Shruti. Neuromodulation of circuits with variable parameters: single neurons and small circuits reveal principles of state-dependent and robust neuromodulation. *Annual review of neuroscience*, 37:329–346, 2014.

[89] Julio C Martinez-Trujillo and Stefan Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current biology*, 14(9):744–751, 2004.

[90] John HR Maunsell. Neuronal mechanisms of visual attention. *Annual Review of Vision Science*, 1:373–391, 2015.

[91] Carrie J McAdams and John HR Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience*, 19(1):431–441, 1999.

[92] Matthew J McGinley, Stephen V David, and David A McCormick. Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron*, 87(1):179–192, 2015.

[93] Matthew J McGinley, Martin Vinck, Jacob Reimer, Renata Batista-Brito, Edward Zagha, Cathryn R Cadwell, Andreas S Tolias, Jessica A Cardin, and David A McCormick. Waking state: rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161, 2015.

[94] Gjerrit Meinsma. Elementary proof of the routh-hurwitz test. *Systems & Control Letters*, 25(4):237–242, 1995.

[95] Kenneth D Miller and Francesco Fumarola. Mathematical equivalence of two common forms of firing rate models of neural networks. *Neural computation*, 24(1):25–31, 2012.

[96] Jorrit Steven Montijn, Rex G Liu, Amir Aschner, Adam Kohn, Peter E Latham, and Alexandre Pouget. Strong information-limiting correlations in early visual areas. *bioRxiv*, page 842724, 2019.

[97] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

[98] J Anthony Movshon, Ian D Thompson, and David J Tolhurst. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *The Journal of physiology*, 283(1):53–77, 1978.

[99] Farzaneh Najafi, Gamaleldin F Elsayed, Robin Cao, Eftychios Pnevmatikakis, Peter E Latham, John P Cunningham, and Anne K Churchland. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. *Neuron*, 105(1):165–179, 2020.

[100] Amy M Ni and John HR Maunsell. Spatially tuned normalization explains attention modulation variance within neurons. *Journal of neurophysiology*, 118(3):1903–1913, 2017.

[101] Amy M Ni, Supratim Ray, and John HR Maunsell. Tuned normalization explains the size of attention modulations. *Neuron*, 73(4):803–813, 2012.

[102] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.

[103] MA Paradiso. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological cybernetics*, 58(1):35–49, 1988.

[104] Donald H Perkel and Theodore H Bullock. Neural coding. *Neurosciences Research Program Bulletin*, 1968.

[105] Erez Persi, David Hansel, Lionel Nowak, Pascal Barone, and Carl van Vreeswijk. Power-law input-output transfer functions explain the contrast-response and tuning properties of neurons in visual cortex. *PLoS computational biology*, 7(2):e1001078, 2011.

[106] Pierre-Olivier Polack, Jonathan Friedman, and Peyman Golshani. Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature neuroscience*, 16(9):1331–1339, 2013.

[107] Michael I Posner. *Cognitive Neuroscience of Attention*. Guilford Press, 2012.

[108] Alexandre Pouget, Kechen Zhang, Sophie Deneve, and Peter E Latham. Statistically efficient estimation using population coding. *Neural computation*, 10(2):373–401, 1998.

[109] Nicholas J Priebe and David Ferster. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, 57(4):482–497, 2008.

[110] Neil C Rabinowitz, Robbe L Goris, Marlene Cohen, and Eero P Simoncelli. Attention stabilizes the shared gain of v4 populations. *Elife*, 4:e08998, 2015.

[111] Alfonso Renart, Jaime De La Rocha, Peter Bartho, Liad Hollender, Néstor Parga, Alex Reyes, and Kenneth D Harris. The asynchronous state in cortical circuits. *science*, 327(5965):587–590, 2010.

[112] Alfonso Renart and Mark CW van Rossum. Transmission of population-coded information. *Neural computation*, 24(2):391–407, 2012.

[113] JH Reynolds and L Chelazzi. Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27:611–647, 2004.

[114] John H Reynolds, Leonardo Chelazzi, and Robert Desimone. Competitive mechanisms subserve attention in macaque areas v2 and v4. *Journal of Neuroscience*, 19(5):1736–1753, 1999.

[115] John H Reynolds and David J Heeger. The normalization model of attention. *Neuron*, 61(2):168–185, 2009.

[116] John H Reynolds, Tatiana Pasternak, and Robert Desimone. Attention increases sensitivity of v4 neurons. *Neuron*, 26(3):703–714, 2000.

[117] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

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

[119] Douglas A Ruff, Joshua J Alberts, and Marlene R Cohen. Relating normalization to neuronal populations across cortical areas. *Journal of Neurophysiology*, 116(3):1375–1386, 2016.

[120] Douglas A Ruff and Marlene R Cohen. Attention can either increase or decrease spike count correlations in visual cortex. *Nature neuroscience*, 17(11):1591–1597, 2014.

[121] Douglas A Ruff and Marlene R Cohen. Attention increases spike count correlations between visual cortical areas. *Journal of Neuroscience*, 36(28):7523–7534, 2016.

[122] Douglas A Ruff and Marlene R Cohen. A normalization model suggests that attention changes the weighting of inputs between visual areas. *Proceedings of the National Academy of Sciences*, 114(20):E4085–E4094, 2017.

[123] Oleg I Rumyantsev, Jérôme A Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105, 2020.

[124] Paul-Antoine Salin and Jean Bullier. Corticocortical connections in the visual system: structure and function. *Physiological reviews*, 75(1):107–154, 1995.

[125] Alessandro Sanzeni, Bradley Akitake, Hannah C Goldbach, Caitlin E Leedy, Nicolas Brunel, and Mark H Histed. Inhibition stabilization is a widespread property of cortical networks. *Elife*, 9:e54875, 2020.

[126] Shreya Saxena and John P Cunningham. Towards the neural population doctrine. *Current opinion in neurobiology*, 55:103–111, 2019.

[127] Odelia Schwartz and Eero P Simoncelli. Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8):819–825, 2001.

[128] Alice C Schwarze and Mason A Porter. Motifs for processes on networks. *arXiv preprint arXiv:2007.07447*, 2020.

[129] Gary Sclar, John HR Maunsell, and Peter Lennie. Coding of image contrast in central visual pathways of the macaque monkey. *Vision research*, 30(1):1–10, 1990.

[130] Terrence J Sejnowski and Patricia Smith Churchland. *Brain and Cognition*, pages 1–47. The MIT Press, 1993.

[131] Peggy Seriès, Peter E Latham, and Alexandre Pouget. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature neuroscience*, 7(10):1129–1135, 2004.

[132] H Sebastian Seung and Haim Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753, 1993.

[133] Gordon MG Shepherd. Corticostriatal connectivity and its role in disease. *Nature Reviews Neuroscience*, 14(4):278–291, 2013.

[134] Herbert A Simon and Craig A Kaplan. *Foundations of cognitive science*, pages 1–47. The MIT Press, 1993.

[135] Samuel G Solomon and Adam Kohn. Moving sensory adaptation beyond suppressive effects in single neurons. *Current Biology*, 24(20):R1012–R1022, 2014.

[136] Hedva Spitzer, Robert Desimone, and Jeffrey Moran. Increased attention enhances both behavioral and neuronal performance. *Science*, 240(4850):338–340, 1988.

[137] Carsen Stringer, Michalis Michaelos, Dmitri Tsyboulski, Sarah E Lindo, and Marius Pachitariu. High-precision coding in visual cortex. *Cell*, 184(10):2767–2778, 2021.

[138] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

[139] Jie Tang and Nobuo Suga. Modulation of auditory processing by cortico-cortical feedforward and feedback projections. *Proceedings of the National Academy of Sciences*, 105(21):7600–7605, 2008.

[140] Gaia Tavoni, David E Chen Kersen, and Vijay Balasubramanian. Cortical feedback and gating in odor discrimination and generalization. *PLoS computational biology*, 17(10):e1009479, 2021.

[141] Alexander Thiele. Muscarinic signaling in the brain. *Annual review of neuroscience*, 36:271–294, 2013.

[142] Alexander Thiele and Mark A Bellgrove. Neuromodulation of attention. *Neuron*, 97(4):769–785, 2018.

[143] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. Gabaergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

[144] Stefan Treue and Julio C Martinez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, 1999.

[145] James Trousdale, Yu Hu, Eric Shea-Brown, and Krešimir Josić. Impact of network structure and cellular response on spike time correlations. *PLoS computational biology*, 8(3):e1002408, 2012.

[146] Todd W Troyer, Anton E Krukowski, Nicholas J Priebe, and Kenneth D Miller. Contrast-invariant orientation tuning in cat visual cortex: thalamocortical input tuning and correlation-based intracortical connectivity. *Journal of Neuroscience*, 18(15):5908–5927, 1998.

[147] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11):4382–4388, 1997.

[148] Carl van Vreeswijk and Haim Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural computation*, 10(6):1321–1371, 1998.

[149] Bram-Ernst Verhoef and John HR Maunsell. Attention operates uniformly throughout the classical receptive field and the surround. *Elife*, 5:e17256, 2016.

[150] Bram-Ernst Verhoef and John HR Maunsell. Attention-related changes in correlated neuronal activity arise from normalization mechanisms. *Nature neuroscience*, 20(7):969, 2017.

[151] Ross S Williamson and Daniel B Polley. Parallel pathways for sound processing and functional connectivity among layer 5 and 6 auditory corticofugal neurons. *Elife*, 8:e42974, 2019.

[152] Tori Williford and John HR Maunsell. Effects of spatial attention on contrast response functions in macaque area v4. *Journal of neurophysiology*, 96(1):40–54, 2006.

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

[154] Rafael Yuste. From the neuron doctrine to neural networks. *Nature reviews neuroscience*, 16(8):487–497, 2015.

[155] Edward Zagha, Amanda E Casale, Robert NS Sachdev, Matthew J McGinley, and David A McCormick. Motor cortex feedback influences sensory processing by modulating network state. *Neuron*, 79(3):567–578, 2013.

[156] Ehud Zohary, Michael N Shadlen, and William T Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, 1994.