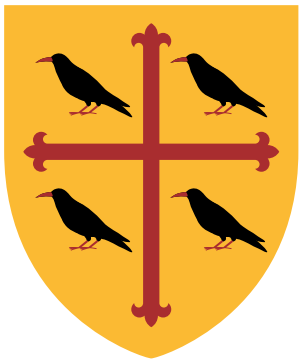
Neural Mechanisms of State-Dependent Sensory Processing





Thijs L. van der Plas

Department of Physiology, Anatomy and Genetics

St Edmund Hall

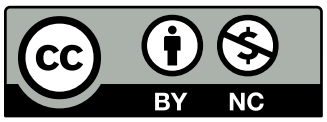
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

This work is licensed under a Creative Commons "Attribution-NonCommercial 4.0 International" license.



Preferred citation: Van der Plas, TL. *Neural mechanisms of state-dependent sensory processing*. (2023) D.Phil. thesis. University of Oxford.

Abstract

A key purpose of the central nervous system is to transform sensation into action. During this process, identical stimuli can cause variable behavioural responses because of changes in the condition – or, state – of the brain. This response variability is a common phenomenon, yet the underlying neural mechanisms of how sensory information differentially interacts with varying network states remain obscure. In this thesis, we consider three such mechanisms in depth: inhibition stabilisation, neural variability and predictive learning. First, we analyse all-optical experiments where we were able to simultaneously record and insert neural activity in somatosensory cortex *in vivo*. Our results show that population activity is stabilised by inhibition, thereby allowing faint sensory stimulation to be detected and propagated downstream. In mice trained to detect these stimuli, we further observe that the neural variability across neurons modulated their perceptual performance, whereas this effect did not occur during passive stimulation. Therefore, maximising the cortical signal-to-noise ratio of behaviourally relevant stimuli is critical to facilitate activity propagation and perception. Separately, in a study of artificial neural networks, we demonstrate that it is advantageous for networks to continuously predict upcoming sensory information, as this facilitates learning of more complex working memory tasks. This could explain recent experimental findings showing that cortical neurons engage in predictive learning even when reward is absent. In sum, this thesis explores how sensory information interacts with ongoing neural activity in both biological and artificial neural networks, and explains mechanisms that underpin the variability of our behavioural response.

Acknowledgements

As an aspiring theoretical neuroscientist starting my PhD, who would have guessed I would end up in a cutting-edge experimental lab (or, closer to the truth, at a desk just outside the lab/in my bedroom)? I was easily convinced, once I met Adam Packer. Our meetings usually ended with Adam synthesising a way forward from an array of computational studies I did not know, or by the loud-and-clear instruction to “do something smart!”. Equally effective, and always inspiring. Adam, you are a trusted mentor and inspiring scientist, encouraging me to follow my passions and do the science that “gets me out of bed in the morning" – a motto that I will take with me wherever I go.

I am also indebted to Tim Vogels, who routinely went out of his way to supervise me. Our collaboration was a string of 8 AM meetings, midnight editing sessions and Whatsapp groups (with no doubt the most efficient means of communication), which we were somehow able to transform into a paper. And to Sanjay Manohar, whose insight in all facets of neuroscience has been so helpful and inspiring. I have learnt enormously from working with you both, and our discussions always re-ignited my passion for neuroscience. Further, I want to thank Andrew Saxe for his brief period of supervision, which helped me develop the ideas in Chapter 4 of this thesis, as well as Armin Lak and Tim Behrens for their curiosity and advice throughout my PhD.

My PhD research would not have been possible without my lab mates extraordinaire – Jimmy, Rob and Matthias – who meticulously acquired the finest data anyone could ever wish for. Being able to work on this project together has undoubtedly been one of the highlights of my PhD. A special shout-out goes to Sarah and Sam – who would agree that (some?) neuroscience is best studied in theory.! Also thanks to Hamish, Prajay, Huriye, Ivan, Gabi, Caitlin, Blake, Marko, David and Andrew, who together created a brilliant atmosphere in the lab. And when I was fortunate enough to visit Tim’s lab in Oxford, Klosterneuburg or Lisbon, I could always count on the warm welcome from Chai, Basile and Panos, and later Sebastian, James and Chris and everyone else in the ever-expanding Vogels lab.

Doing a PhD gave me the freedom to work on my interests and own ideas, even when they often changed. This freedom meant I usually worked on multiple projects at the same time, which would not have been possible without the understanding and patience of my colleagues, in particular Jérôme, Georges, Volker and Bernhard, David and Simon, Tim, Sanjay and of course Adam. Not all projects I worked on could make it into this thesis, regrettably, as I am incredibly proud of the science we produced.

I was fortunate to undertake my doctoral studies in the stimulating academic environment that is Oxford, not in the least because of the company provided by my cohort at Teddy Hall and the DTC. Being around these great scientific minds was a privilege, including Jack, Jacques, Jai, Nic, Andrew, Hank and Jochem – who I could always draw on for late-night scientific discussions. The All-Day-Symposium that was organised in the summer of 2023 was a particular highlight for us all.

Much of my PhD was done 'locked-down', where old friends Freek, Lennaert, Uphoff, Dab and Jos provided much-needed distraction and morale boosts from across the channel (*als de Waal uit zicht is ..*), as did Jacob – who inspired me to apply for Oxford – on our weekly bike rides. At all times I could count on the support of pap, mam, Luka and Fons, encouraging me to adventure the unknown. And who better to adventure with than Julia, who has taught me to love all the good things in life. Thank you for helping me realise my dreams and ideas, and for supporting me in everything I do.

Author contributions

I declare that this thesis is my own and that attribution has been given where information was derived from other sources. I designed and implemented the data analyses, modelling studies, and performed all data visualisation. The project as a whole has relied on key contributions – in particular the data acquisition – from others, which is further detailed below. Throughout, the text uses the pronoun 'we' indicating myself and the team below.

Chapter 1: Section 1.5.1 was adapted, with permission, from Van der Plas et al. (2022).

Chapters 2 and 3: Data acquisition and pre-processing by **James Rowland**, **Joshua Keeling**, **Robert Lees** and **Adam Packer**. Most of the work presented in Chapter 2 and sections 3.1.1 & 3.1.2 has been published in Rowland*, Van der Plas*, Loidolt* et al. (2023), which I have re-used with permission, and only the data analyses and visualisations performed by me are presented in this thesis. Parts of the text in these sections were adapted from our paper, and therefore co-written with **James Rowland**, **Matthias Loidolt** and **Adam Packer**.

The second set of experiments presented in section 3.1.3 was acquired and pre-processed by **Robert Lees**. Both projects were supervised by **Adam Packer**.

Chapter 4: Most of the work presented in this Chapter has been published in Van der Plas et al. (2022), which I have re-used with permission, with supervision from **Tim Vogels** and **Sanjay Manohar**.

Chapter 5: Parts of the text were adapted, with permission, from Rowland*, Van der Plas*, Loidolt* et al. (2023), which I co-wrote with **James Rowland**, **Matthias Loidolt** and **Adam Packer**, and Van der Plas et al. (2022).

Contents

List of Selected Abbreviations x

1 Introduction 1

1.1 State of the brain . . . 1
1.2 Bypassing the (somato)sensory pathway with optogenetic stimulation . . . 3
1.3 Inhibition stabilisation and stimulus detection . . . 4
1.3.1 Inhibitory cell types . . . 4
1.3.2 Stimulus detection in inhibition-stabilised networks . . . 5
1.3.3 Using optogenetics to study stimulus-detection *in vivo* . . . 6
1.3.4 Measuring the propagation of neural activity . . . 9
1.4 Quantifying neural variability at different time scales . . . 9
1.4.1 Changes in neural variability across experimental conditions . . . 11
1.4.2 Changes in neural variability at single-trial resolution . . . 11
1.4.3 Dynamically quantifying neural variability across trials . . . 13
1.4.4 Dynamically quantifying neural variability instantaneously . . . 15
1.5 Predictive learning in recurrent neural networks . . . 15
1.5.1 Predictive learning . . . 15
1.5.2 Studying neural mechanisms with recurrent neural networks . . . 17

2 Inhibition Stabilisation 18

2.1 Results . . . 19
2.1.1 Causally driving behaviour with all-optical interrogation across cortical areas . . . 19
2.1.2 Two-photon photostimulation of S1 drives balanced activity that propagates to S2 . . . 20
2.1.3 S1 and S2 populations encode representations of perceived photostimulation 25
2.2 Methods . . . 31
2.2.1 Experimental procedures . . . 31
2.2.2 Behavioural data analysis . . . 36
2.2.3 Normalisation and sorting of post-stimulus neural activity . . . 36

2.2.4 Definition of targets, photostimulation responsiveness, and excited/inhibited neurons . . . 37
2.2.5 Dynamic decoding of stimulus representation . . . 37
2.2.6 Data and code availability . . . 39

3 Neural Variability **40**
3.1 Results . . . 41
3.1.1 Pre-stimulus S1 activity predicts trial outcome . . . 41
3.1.2 Stimulus strength and neural population variance determine probability of perception . . . 42
3.1.3 Passive processing of optogenetic stimulation . . . 49
3.2 Methods . . . 55
3.2.1 Pre-stimulus population metrics . . . 55
3.2.2 Experimental details of section 3.1.3 . . . 57
3.2.3 Suprathreshold cluster size test . . . 58
3.2.4 Population decoder . . . 58
3.2.5 Data and code availability . . . 59

4 Predictive Learning **60**
4.1 Results . . . 61
4.1.1 Recurrent neural networks use eavesdropping to bootstrap learning . . . 61
4.1.2 Quantification of the eavesdropping effect . . . 62
4.1.3 Matching task eavesdrops on the stimulus memory of the prediction task 65
4.1.4 Increasing matching task complexity reduces eavesdropping effect . . . 68
4.1.5 Simulated annealing enables RNNs to learn the matching task with uncorrelated stimuli . . . 70
4.2 Methods . . . 72
4.2.1 Terminology . . . 72
4.2.2 Tasks . . . 72
4.2.3 Synthetic data . . . 73
4.2.4 Recurrent neural network training . . . 74
4.2.5 Quantifying stimulus representation strength . . . 75
4.2.6 Software & code availability . . . 76

5 Discussion **77**
5.1 Optogenetic stimulation . . . 77
5.1.1 Measuring the propagation of neural activity . . . 77
5.1.2 Technical limitations . . . 78
5.2 Inhibition stabilisation . . . 79

5.2.1 Inhibition stabilisation in a stimulus-detection task . . . 79
5.2.2 Generalised or random neural stimulus-response . . . 80
5.3 Neural variability . . . 81
5.3.1 Signal-to-noise ratio as a mechanism for stimulus detection . . . 81
5.3.2 Measures of Neural Variance . . . 81
5.4 Predictive learning . . . 83
5.4.1 Summary and limitations . . . 83
5.4.2 Predictive learning and eavesdropping in natural learning . . . 84
5.4.3 Experimental predictions . . . 86
5.5 Conclusion . . . 86

Appendices

A Neural variability estimation methods in neuroscience **88**

B Derivation of optimal loss for RNN prediction task **92**

Bibliography **95**

List of Selected Abbreviations

$\Delta F/F$	Normalised fluorescence activity (from the calcium reporter).
1D, 2D, ..	1-dimensional, 2-dimensional, ..
AUC	Area under the curve.
CR	Correct rejection.
d'	Discriminability index
DCM	Delayed category match.
DSM	Delayed sample match.
FDR	False discovery rate.
E/I	Excitation/inhibition
FP	False positive.
IQR	Interquartile range.
L1	Least absolute deviations loss function (1-norm).
L2	Least squares loss function (2-norm).
M/NM	Match/non-match.
PC(A)	Principal component (analysis).
PR	Participation ratio.
R^2	Coefficient of determination.
RNN	Recurrent neural network.
S1	Primary somatosensory cortex (Chapters 2, 3).
S_1	Stimulus 1 (Chapter 4).
S2	Secondary somatosensory cortex (Chapters 2, 3).
S_2	Stimulus 2 (Chapter 4).
SEM	Standard error of the mean.
SNR	Signal-to-noise ratio.

I have got something no one else has; my brain, which I use to my advantage, when advantageous.

— Andy Bernard (The Office)

1 Introduction

1.1 State of the brain

How we process sensory information determines how we perceive the world. First, information about our environment is received by the receptors that make up our senses and electrically relayed to our brain. Sensory information is then compressed, combined, compared against memory, judged, and finally, transformed into action (Meredith and Stein, 1986; Schneidman et al., 2006; Carpenter and Reddi, 2012; Mante et al., 2013; Migault et al., 2018; Ebrahimi et al., 2022). However, two identical pieces of sensory information – or, stimuli – do not necessarily yield the same action (Waschke et al., 2021). We may have learnt to respond differently under different circumstances (Mante et al., 2013), or simply happen to not be paying attention (Cohen and Maunsell, 2009). Hence, *sensory processing* can depend on the *state* of the brain. This can include changes in brain state after learning, its state of attention, or (importantly) its caffeine state. And although we are often able to describe the phenomena that can alter a state in simple terms (*e.g.*, 'paying attention'), the *neural mechanisms* that instantiate these effects at the scale of networks of neurons – in other words, exactly *how* this state mechanistically affects sensory processing – are more difficult to infer. In this thesis, we investigate three of such mechanisms: inhibition stabilisation, neural variability and predictive learning.

This chapter serves to introduce each of these topics, starting with an introduction to the somatosensory pathway. Then, we review how inhibitory neurons stabilise the neural activity of cortical networks, an essential component of neural computation. Previous work has shown that this inhibition-stabilised state facilitates the detection of small, experimentally induced stimulation of cortical neurons (Marshel et al., 2019; Dalgleish et al., 2020). In Chapter 2, we will build on this research and show that this mechanism is linked to the propagation of neural activity across brain regions.

Next, we consider how fluctuations in the neural variability state can account for variable task performance. As changes in neural variability have previously been linked to changes in perceptual performance (Cohen and Maunsell, 2009; Ni et al., 2018; Waschke et al., 2019; Huang et al., 2019), we first review the existing methods to quantify neural variability, including correlation, entropy and effective dimensionality metrics. We observe that the choice of metric confines the analysis to a particular time scale, unable to capture changes in neural variability at faster rates. In Chapter 3, we define a metric of neural variability that can capture instantaneous variability, and demonstrate how it predicts perceptual performance on a stimulus-detection task.

Finally, we switch gears and investigate the utility of predictive learning using a multitask learning paradigm. Predictive learning – learning to predict upcoming sensory information – has been observed as a passive-learning process in cortex (*i.e.*, without reward incentives) (Libby and Buschman, 2021). What could be the purpose of acquiring knowledge using predictive learning? In Chapter 4, we study artificial neural networks to precisely understand how stimulus representations learnt through predictive learning can be re-used across tasks.

Taken together, this thesis probes into several key mechanisms that influence sensory processing, using a combination of experimental and theoretical work (Figure 1.1).

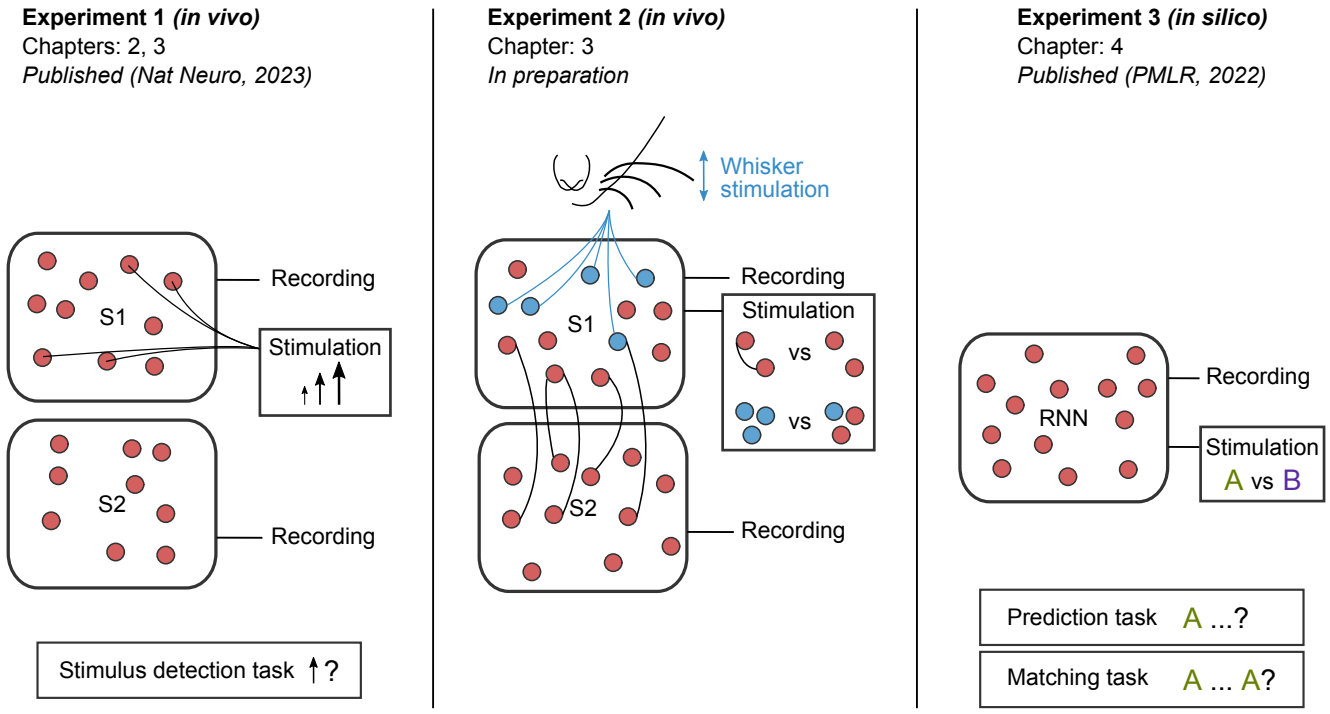


Figure 1.1: Overview experiments analysed in this thesis.
Schematics of experiment set-ups. Each box illustrates a neural network, where neurons are depicted by the circles. S1: primary somatosensory cortex, S2: secondary somatosensory cortex, RNN: recurrent neural network.

1.2 Bypassing the (somato)sensory pathway with optogenetic stimulation

Sensory pathways describe the trajectory of sensory information from the first contact with our sensory organs to the brain. For example, mice use their whiskers to explore their environment via somatosensation. Stimulation of whiskers elicits a neural response in trigeminal ganglion nerves, which propagates to the brain stem and thalamus, before reaching the somatosensory cortex, which is divided into primary (S1) and secondary (S2) somatosensory cortex (Diamond et al., 2008). In these two cortical regions, neurons are tuned to respond to specific whiskers, which gives rise to a somatotopic mapping of 'barrels' that correspond to individual whiskers (Woolsey and Van der Loos, 1970; Diamond et al., 2008). This barrel structure is especially clearly visible in S1, while less clearly separated in S2, a region that has also been implicated in higher-order cognitive functioning (Ni and Chen, 2017). This difference likely stems in part from the fact that most sensory pathways reach S1 first, which relays its information to downstream

S2 (Ni and Chen, 2017). S2 projects back to S1, creating a recurrently connected network, and additionally, both S1 and S2 project back to the thalamus, creating a thalamo-cortical loop that ultimately is involved in decision-making and motor action (Yang et al., 2016; Ni and Chen, 2017). These inter-regional S1-S2 connections are thought to facilitate perception (Kwon et al., 2016; Chen et al., 2016; Condylis et al., 2020), but exactly how this propagation of activity across regions is instantiated remains unclear.

It is challenging to study this propagation of natural stimulus information in detail because, in the case of somatosensory cortex, natural whisker stimulation evokes strong, multi-region neural responses that reach S1 and S2 through multiple, recurrent sensory pathways (Diamond et al., 2008; El-Boustani et al., 2020), hence making it almost impossible to disambiguate the origin (and thus propagation) of these different components of neural activity.

In Chapters 2 and 3, we overcome these constraints by using an all-optical approach that simultaneously records neuronal calcium activity and optogenetically stimulates pre-programmed sets of neurons in sensory cortex (Packer et al., 2015). This allowed us to directly initiate the stimulation of multiple S1 neurons, while recording the ensuing S1 and S2 neural response, without the variability and complexity that are associated with natural whisker stimulation. This precise control of the stimulus in S1 allowed us to interrogate how stimulus information is maintained, propagated and processed by the S1-S2 circuit. In Chapter 4, we use an artificial neural network that simulates a cortical circuit (Barak, 2017), where the stimuli also directly target the observed neurons. These two approaches, *in vivo* and *in silico*, hence bypass the early stages of a sensory pathway, and allow us to precisely analyse the neural representations of stimulation.

1.3 Inhibition stabilisation and stimulus detection

1.3.1 Inhibitory cell types

Our cognitive ability stems from the ability of neurons to connect, forming large biological neural networks that perform complex computations. Connections are made via synapses, where electric cellular activity is relayed between neurons via chemical neurotransmitters.

Neurons generally release the same neurotransmitter at all their synaptic connections (known as Dale's principle), which, simply put, either excites or inhibits the post-synaptic cell, thereby labelling the pre-synaptic neuron as either excitatory or inhibitory (Carpenter and Reddi, 2012). Inhibitory neurons serve various functions, such as muscle relaxation (Carpenter and Reddi, 2012), but here we will focus on their role in large, cortical networks of neurons, which has been studied extensively using theoretical models. Specifically, inhibitory cells have been determined to be essential for preventing unstable network dynamics (Van Vreeswijk and Sompolinsky, 1996; Vogels et al., 2005; Renart et al., 2010; Denève and Machens, 2016; Rosenbaum et al., 2017; Sanzeni et al., 2020; Sadeh and Clopath, 2021). Further, (tuning) inhibition can facilitate a wide range of cognitive functions, including attention (Kanashiro et al., 2017; Huang et al., 2019), memory (Vogels et al., 2011; Hennequin et al., 2017), signal propagation (Vogels et al., 2005; Huang et al., 2022), motor control (Hennequin et al., 2014; Kao et al., 2021) and stimulus detection (Marshel et al., 2019). Lastly, experimental work has implicated deficits in inhibition – and more generally, the balance between excitation and inhibition – in various neurological disorders including Alzheimer's Disease (Maestú et al., 2021), autism (Yizhar et al., 2011; Nelson and Valakh, 2015) and schizophrenia (Yizhar et al., 2011).

1.3.2 Stimulus detection in inhibition-stabilised networks

Theoretical neuroscience has thus assigned a stabilising role to inhibitory neurons (Ozeki et al., 2009). Here, using recurrent connections (*i.e.*, feedback connections within a network), inhibitory neurons can rapidly react to the activation of excitatory neurons, and thereby dampen the network dynamics before these amplify out of control to seizure-like activity (Denève and Machens, 2016; Sadeh and Clopath, 2021). However, as reviewed in section 1.3.1, the function of inhibition goes beyond stabilising neural dynamics. By tracking and cancelling out excitation, inhibition can effectively desynchronise network activity (Renart et al., 2010), which generally enhances perceptual performance (Waschke et al., 2019; Reato et al., 2023). Here we focus specifically on the role of inhibition in stimulus detection: networks stabilised by inhibition are typically in a 'high-gain' state, where the activation of just a relatively small set of excitatory neurons can elicit a network response without inducing a seizure-like state (Sadeh and Clopath,

2021).

To exemplify both of these functions (avoiding unstable dynamics and facilitating stimulus detection), we replicated the simple mean-field model from Marshel et al. (2019), see Figure 1.2. Here, the mean activity of all excitatory neurons (in a network) is modelled by m^E, and the mean activity of all inhibitory neurons by m^I. These two populations are connected to each other with weights $w^{E,I}$ (excitatory to inhibitory) and $w^{I,E}$ (vice versa), and feed back within each population with weights $w^{E,E}$ and $w^{I,I}$. In short (see Marshel et al. (2019, Supplementary Information VII) for the derivation), the dynamic activity (at time t) of these two mean activities is given by:

$$m_t^k = H\left(T - (w^{E,k} \cdot m_{t-1}^E - w^{I,k} \cdot m_{t-1}^I)\right), \quad k \in \{E, I\}, \quad H(x) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{x}{\sigma}\right) \tag{1.1}$$

(where standard deviation $\sigma = \frac{3\sqrt{2}}{4}$, threshold $T = 1$, erfc is $1 -$ error function and the connecting weights are $w^{E,E} = 3$, $w^{E,I} = w^{I,E} = 2.25$ and $w^{I,I} = 0.15$ (Marshel et al., 2019)). Using these default values, this system has one stable fixed point, at low activity values ($m^E_{FP_l} \sim 0.1$, $m^I_{FP_l} \sim 0.15$), and any (excitatory) perturbation away from this fixed point is quickly stabilised by the recruited inhibition (Figure 1.2a). Instead, when inhibition is weak ($w_{I,E} = 0.7$), another stable fixed point occurs at high activity values ($m^E_{FP_h} \sim 0.9$, $m^I_{FP_h} \sim 0.3$). While small perturbations away from the low fixed point converge back to that same fixed point, stronger perturbations push the system into the undesirable, high-activity state (Figure 1.2b). Hence, strong inhibition is required for the network activity to be stable (at low activity rates). Secondly, in this strong-inhibition state, small perturbations can still elicit a large, transient network response. To illustrate this, we decreased the excitatory drive to inhibitory neurons ($w^{E,I} = 0.94$, and $w^{I,E}$ was increased to 2.4 for stability): here, a small perturbation away from the only fixed point first recruits more excitatory neurons, until inhibition catches up and slowly brings the system back to normal (Figure 1.2c).

1.3.3 Using optogenetics to study stimulus-detection *in vivo*

As previously detailed in section 1.2, optogenetic stimulation is crucial to study the propagation of stimulus information. Further, this needs to be combined with neural activity recordings

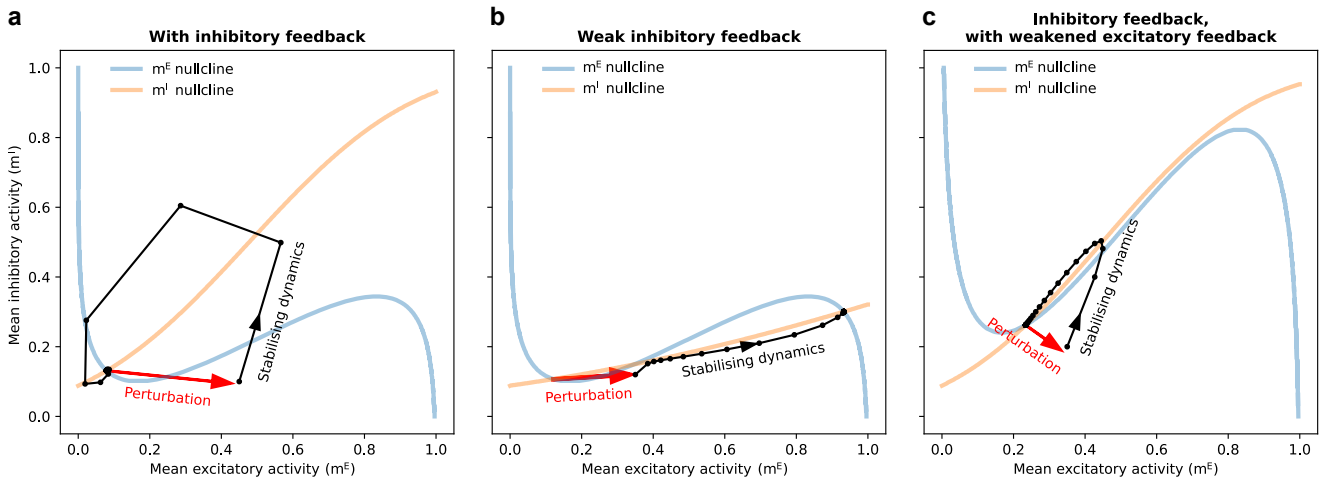


Figure 1.2: Mean-field model of excitatory/inhibitory response to perturbation.
The mean-field model of Marshel et al. (2019, Supplementary Information VII) was replicated (Equation 1.1), showing the equivalent of Marshel et al. (2019, SI Figure 17) (but note that the nullclines are incorrectly labelled in the original paper). Panel **a)** used all original parameters (see caption of Marshel et al. (2019, SI Figure 17)), including inhibitory-to-excitatory feedback weight $w^{I,E} = 2.25$, while panel **b)** weakened this feedback ($w^{I,E} = 0.7$). **c)** Another simulation, but with the excitatory-to-inhibitory weight decreased from 2.25 to $w^{E,I} = 0.94$, and $w^{I,E} = 2.4$. In all panels, dynamics were initiated away from the fixed point (red lines), and the mean-field equations were subsequently run to evaluate the dynamics of the excitatory (m^E) and inhibitory (m^I) means (black lines). The code of our implementation is available online.

in order to simultaneously measure the excited and inhibited activity (of both optogenetically targeted and non-targeted cells) (Packer et al., 2015). Next, optogenetic stimulation must be at single-cell resolution to be able to distinguish between targeted and non-targeted cells, and to allow precise control of the stimulus (strength). However, although neural responses to single-cell perturbations can be observed, they are generally not strong enough to elicit a full network response (Chettih and Harvey, 2019). To overcome this, spatial light modulators can split the (two-photon) optogenetics laser beam into multiple beams, in order to target multiple cells simultaneously (Nikolenko et al., 2008; Packer et al., 2015).

Using a spatial light modulator, Marshel et al. (2019) were able to study the network and behavioural response to optogenetic perturbation of a group of cells with single-cell precision. In their experiments, mice had to perform an orientation-detection task, and responses to visual stimuli (of gratings) were compared to responses to optogenetic stimulation of orientation-tuned neurons in primary visual cortex. In summary, they found that mice reliably and correctly

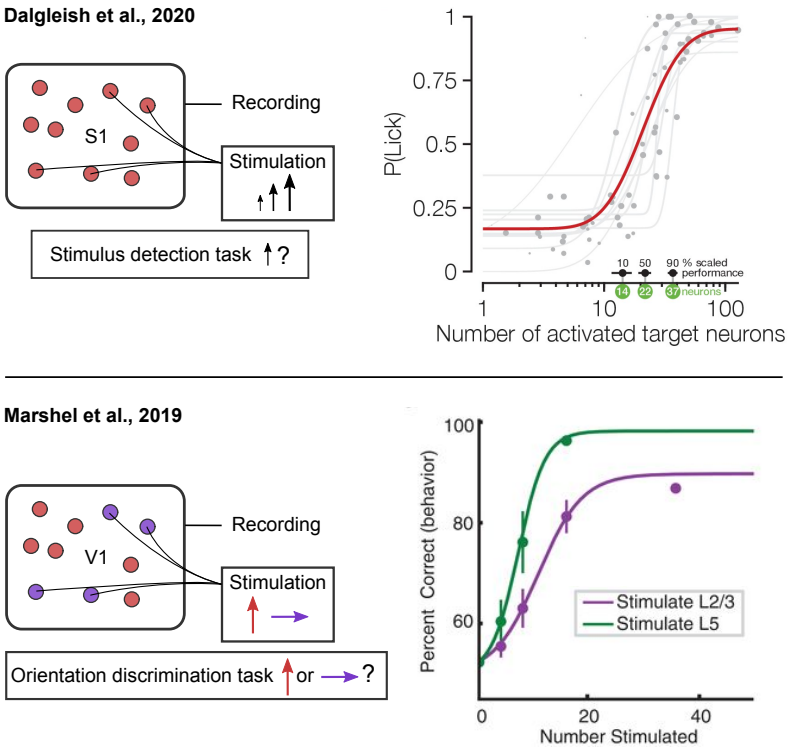


Figure 1.3: Perceptual performance as a function of stimulus strength. Figures adapted from Dalgleish et al. (2020, Figure 2d) (top) and Marshel et al. (2019, Figure 6g) (bottom). The left panels show simplified schematics of the tasks, while the right panels show the perceptual performance as a function of increasing number of optogenetically stimulated neurons.

responded to the optogenetic stimuli of similarly orientation-tuned neurons, even if they had learnt to perform the stimulus-detection task using visual stimuli and photostimulation of other neurons only. Also, stimulating more neurons yielded a higher probability that mice responded correctly (Figure 1.3, bottom). Further, the presence of optogenetic stimulation could also be decoded from the non-targeted population activity in the same brain region. To explain these findings, Marshel et al. (2019) turned to the mean-field model above, arguing that strong inhibition was critical to facilitate detection of the excitation of a small fraction of the excitatory neurons.

Dalgleish et al. (2020) then showed that in addition to being able to respond to the optogenetic stimulation of stimulus-selective neurons (as by Marshel et al. (2019)), mice could also learn to respond to the stimulation of any cortical neurons. Here, mice were trained to respond to optogenetic stimulation of any set of S1 neurons, and successful responses were found to depend on the number of stimulated cells (Figure 1.3, top). Further, Dalgleish et al. (2020) found that the photostimulation induced both excitatory and inhibitory responses within S1, which both increased as a function of number of cells targeted such that they remained in balance.

1.3.4 Measuring the propagation of neural activity

The previous section reviewed how optogenetic stimulation experiments revealed that perceptual performance depends on the number of cells stimulated, and how strong inhibition enables a low perceptual threshold. At the same time, we know that the sensory processing of *natural* stimuli generally involves the propagation of neural activity across the nervous system (section 1.2). In Chapter 2 we will therefore explore whether perception of optogenetic stimulation also depends on propagation of activity, in the context of an inhibition-stabilised system. To do so, we developed an experiment where we stimulated groups of S1 excitatory neurons, while recording both excitatory and inhibitory neurons in S1 and S2, a downstream region (Figure 1.1, left). We measured excitatory and inhibitory responses of a large population of neurons, while mice performed a stimulus-detection task of the optogenetic S1 stimulation. Hence, although we cannot distinguish between excitatory and inhibitory cells in our neural recordings, we were able to disambiguate *excited* from *inhibited* responses (like Dalgleish et al. (2020)).

In Chapter 3, we continue the analysis of this experiment, focusing on the role of neural variability for perception. In the next section, we further introduce this topic.

1.4 Quantifying neural variability at different time scales

Neural variability is the variability in neural responses, and has been implicated as a key driver of variable brain states (Waschke et al., 2021). In this section, we will review how changes in neural variability can correlate to changes in neural mechanisms or behaviour, succinctly summarised by Waschke et al. (2021, p10) as:

> "Behaviorally relevant differences in neural variability within and between individuals may capture neural processes that (1) reflect specific, time-locked aspects of the current task (*e.g.*, stimulus onset; Cohen and Maunsell (2009)); (2) index experimental conditions in general (*e.g.*, aggregated condition-wise differences in rest versus task; Grady and Garrett (2018)), or (3) are not directly related to the task, but nevertheless affect behavior (*e.g.*, spontaneous arousal states; McGinley et al. (2015b))."

Interestingly, these changes can occur at different time scales (ranging from milliseconds to hours) (and possibly simultaneously (Ni et al., 2018; Waschke et al., 2019)), creating a demand

for multiple methods operating at different temporal resolutions. Partly because of this, there now exists a wide range of established metrics to quantify neural variability, including variance, correlations, synchrony and effective dimensionality. Here, we review and systematically categorise these metrics into three key time scale categories: experimental condition, single-trial and instantaneous (Figure 1.4). The time scale of interest depends on the mechanism studied, because of a trade-off between temporal resolution and data availability. Higher temporal resolution enables the study of faster-changing neural mechanisms, while higher data availability allows more complex metrics to be used. This leads to complementing analyses, as it is generally not possible to derive a precise relation between different metrics, especially not when based on data acquired using different experimental techniques (Waschke et al., 2021). For example, the relationships between noise correlations and both neural variability and dimensionality are complex and not monotonic (Umakantha et al., 2021; Waschke et al., 2021).

In complement to the summary below, Appendix A provides more detail and lists all studies and metrics that we considered. We will argue that to understand the influence of fast changes in neural variability, metrics are required that can operate on short time scales. In Chapter 3, we apply this analysis to study the effect of neural variability in our optogenetic stimulation experiments.

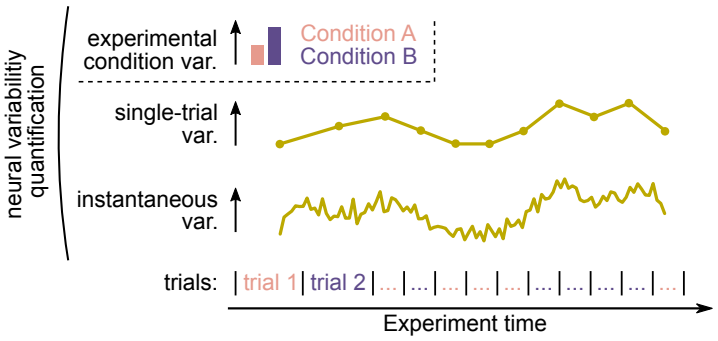


Figure 1.4: Neural variability can be measured at different timescales.
First, given an experiment with trials divided into two experimental conditions, data from each condition can be aggregated and analysed to give a condition-dependent estimate of neural variability (first, from left to right). Second, data can be aggregated across time per trial, yielding a single measurement per trial (second), or across trials to compute variability dynamically (third). Lastly, data can be analysed instantaneously, yielding a dynamic estimate of neural variability (fourth).

1.4.1 Changes in neural variability across experimental conditions

Experimental condition-based analysis can be a powerful tool to study the effect of an isolated variable on neural variability and behaviour (Doiron et al., 2016). In its simplest form, this means that experimental trials can be divided into 2 (or more) conditions (Figure 1.5A), and that the neural variability is calculated for each condition separately (Figure 1.4, top). Generally, this yields sufficient data to calculate covariance-based metrics, such as correlations or (Principal Component Analysis (PCA)-derived) effective dimensionality (Doiron et al., 2016; Umakantha et al., 2021). One notable field of study is (spatial) attention, which is known to have strong influence on neural correlations, variance and dimensionality (Cohen and Maunsell, 2009; Ni et al., 2018; Huang et al., 2019). Spatial attention has been shown to decorrelate the neural population – increasing the effective dimensionality, which boosts performance of orientation change-detection tasks (Cohen and Maunsell, 2009; Huang et al., 2019). This difference in neural variability is typically quantified by using all attended trials to obtain a correlation/dimensionality estimate for the attended condition, and comparing this to the unattended condition (Cohen and Maunsell, 2009; Huang et al., 2019), *i.e.*, at the temporal resolution of experimental conditions.

1.4.2 Changes in neural variability at single-trial resolution

Condition-based analysis is useful to assess the influence of a single (or a few) task variables (such as attended versus unattended trials). However, other, gradual processes that slowly change neural variability require a finer (trial-by-trial) temporal resolution to be observed (Figure 1.6). In particular, learning paradigms concurrent with neural recordings have shown that the structure of population variability can undergo long-term changes. For example, as correlated variability (noise correlations) can limit information encoding by neural populations (Averbeck et al., 2006; Rumyantsev et al., 2020; Bartolo et al., 2020), perceptual learning has been shown to be able to decrease these correlations, which improved performance of a change-detection task in monkeys (Ni et al., 2018). This effect was quantified by measuring the Fano factor and spike count correlations per trial, and assessing the gradual trend during learning (over the course of the experiment). In other words, this analysis was performed with the resolution of single

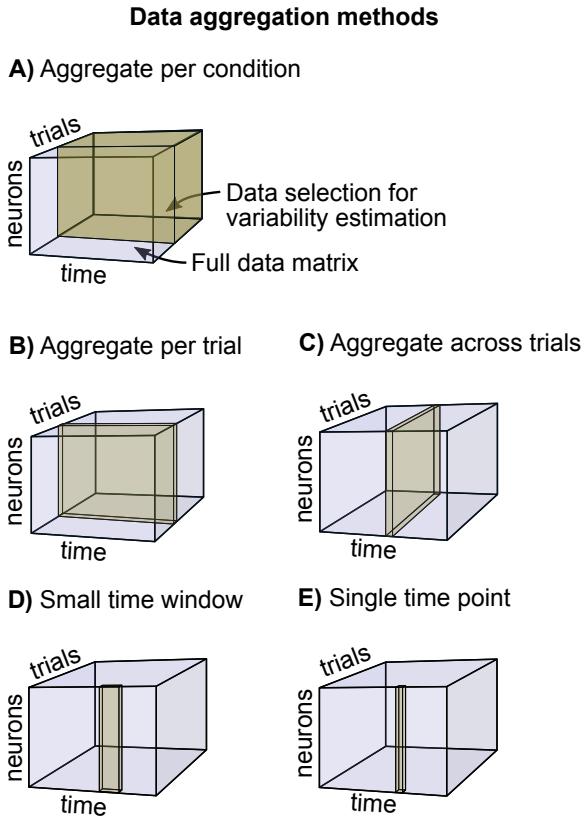


Figure 1.5: Different methods have been employed to aggregate data to quantify its neural variability. **A)** To obtain large data selections, one can average per condition (*e.g.*, (Cohen and Maunsell, 2009)), here schematically depicted by the second half of trials – although typically conditions would intersperse during the experiment. **B)** A single-trial estimate can be obtained from the 2D neurons × time points matrix, per trial (Ni et al., 2018; Reato et al., 2023). **C)** The variability across trials can be dynamically determined using the time-varying 2D matrix of neurons × trials (Churchland et al., 2010; Javadzadeh and Hofer, 2022; Farrell et al., 2022). **D)** For single-trial measurements, multiple time points can be grouped to compute a metric, which still enables methods that require multiple time points per neuron to operate (Waschke et al., 2019). This can potentially be combined with a running-window approach. **E)** Lastly, instantaneous measurements can be taken from the 1D array of neurons at a given trial and time point, which limits the range of metrics that can be used, but includes population variance (Rowland*, Van der Plas*, Loidolt* et al., 2023).

trials (Figure 1.4, middle), and is necessary to track gradual processes (such as learning). By performing this analysis on both the experimental condition and single-trial resolutions, Ni et al. (2018) were then able to show that learning and attention are complementary processes (occurring at different time scales) that change the population variability. This illustrates how changing the resolution of an analysis can yield new insights into neural processing.

These data analyses (of correlations, Fano factor, PCA) are based on the co-variability of neurons, and therefore inherently require a sufficient number of time points to be evaluated. Notably, the number of necessary time points grows with increasing recording size. For this reason, they cannot be naturally evaluated instantaneously, *i.e.*, per individual time point. In another study, Reato et al. (2023) approximated neural variability from the temporal standard deviation of the population firing rate directly, requiring 2 seconds of electrophysiology recording. They computed one value per trial, showing that the influence of desynchronisation on perceptual

performance is modulated by reward feedback of the previous trial.

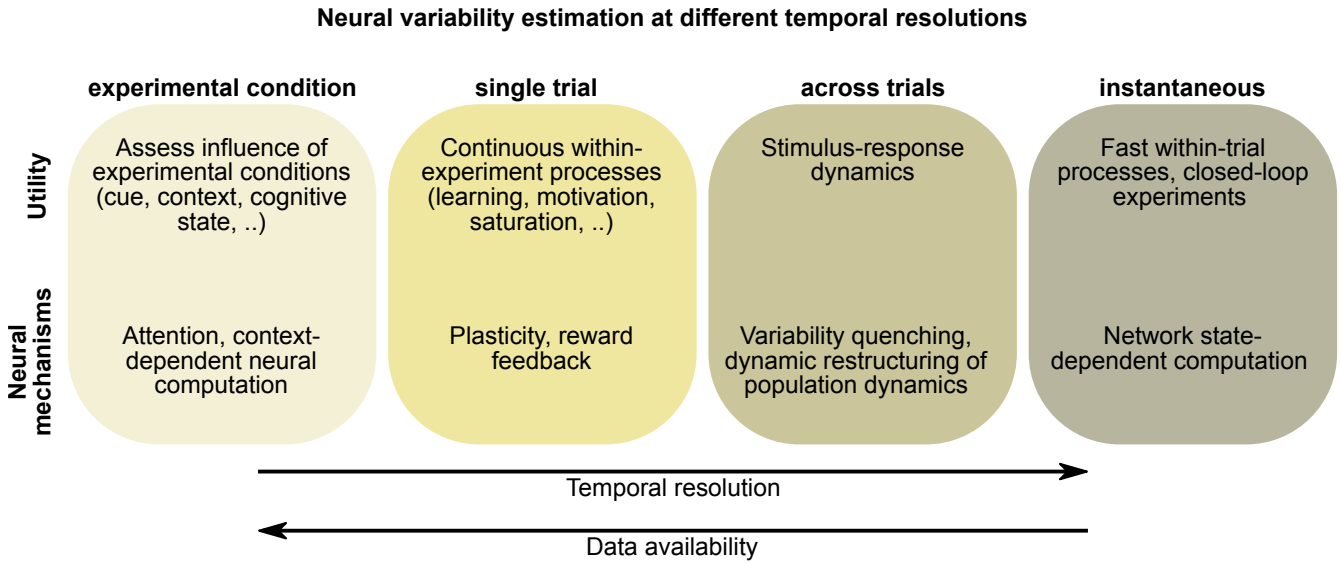


Figure 1.6: Schematic of the three time scale categories for neural variability estimation. Categories are loosely defined depending on whether data is temporally aggregated per experimental condition (first from left to right), single trial (second), across trials (third) or analysed instantaneously (fourth). Neural mechanisms are categorised based on which resolution category aligns best with their intrinsic time scale: attention (Cohen and Maunsell, 2009), context-dependent neural computation (Mante et al., 2013), plasticity (Ni et al., 2018), reward feedback (Reato et al., 2023), variability quenching (Churchland et al., 2010; Mazzucato et al., 2016), dynamic restructuring of population dynamics (Farrell et al., 2022; Javadzadeh and Hofer, 2022) and network state-dependent computation (Waschke et al. (2019) and Rowland*, Van der Plas*, Loidolt* et al. (2023)).

1.4.3 Dynamically quantifying neural variability across trials

Experimental condition and single-trial resolution techniques require all time points of (at least) one trial to calculate variability (Figure 1.5A, B). One strategy to obtain *dynamic* measurements of neural variability instead, is to apply variability metrics across trials, instead of across time points (Figure 1.5C) (Churchland et al., 2010; Farrell et al., 2022; Javadzadeh and Hofer, 2022). This allows one to study the neural variability at each time point separately, but assumes that the dynamics of interest are those that are repeated across trials. This naturally limits this analysis to stimulus-response dynamics, and precludes the analysis of closed-loop experiments (Waschke et al., 2019), spontaneous dynamics (Van der Plas*, Tubiana* et al., 2023) and responses to unique

stimuli (Stringer et al., 2019).

Therefore, across-trial neural variability estimation has predominantly been used to study stimulus-response dynamics. One phenomenon that has stood out across species and experiments is that neural variability decreases after stimulus presentation (Churchland et al., 2010; Mazzucato et al., 2016). This has been quantified by a number of metrics, including temporal standard deviation, Fano Factor and factor analysis (a PCA-based technique) (Churchland et al., 2010), and eigenvalue participation ratio (Mazzucato et al., 2016). This implies that neural dynamics *across trials* become less variable – or, more stable – after stimulus presentation. In other words, neural dynamics become stereotyped during a stimulus-response compared to (spontaneous) pre-stimulus dynamics.

However, networks can also increase their variability in response to a stimulus if this benefits task performance, as was demonstrated by an *in silico* study of Recurrent Neural Networks (RNNs) (Farrell et al., 2022). Specifically, the authors found that RNNs increased their effective dimensionality directly after stimulus presentation to facilitate a linear separable classification in high-dimensional space, and subsequently compressed the dimensionality to maintain a stable solution. To quantify this, the authors also computed the eigenvalue participation ratio across trials.

Lastly, Javadzadeh and Hofer (2022) used across-trial PCA to establish how the effect of one brain region on another changed dynamically. Specifically, by optogenetically silencing a brain region in various 65 ms intervals, they were able to quantify the dynamic effect of this perturbation downstream. They found that inter-area information flow can change rapidly, evidenced by dynamic restructuring of PCA dimensions.

In summary, several studies have quantified the variability among neurons across trials, and thereby shown that variability can change dynamically to facilitate neural computation. However, these across-trial analyses are inherently limited to the evaluation of neural dynamics in

response to repeated stimuli, and necessarily merge all trials to obtain a covariance estimate. This therefore precludes the analysis of experiments without a repeated-trial structure, and it remains unclear whether the neural mechanisms discovered by across-trial analysis also occur within the single trials.

1.4.4 Dynamically quantifying neural variability instantaneously

There are two strategies of obtaining instantaneous metrics of neural variability in single trials; by using short time windows (Figure 1.5D, potentially combined with a running-window approach), or by using metrics that work on 1D data (Figure 1.5E). As an example of the first, Waschke et al. (2019) developed an entropy-based metric that computes the neural variability of electroencephalography (EEG) data continuously using a running window. Here, entropy was calculated using 200 ms of data (collected at 1 kHz) and compared to entropy values of the past 30 seconds to assess whether entropy was relatively high or low. This could be computed online in real-time, allowing a closed-loop experiment where stimuli where either delivered at high- or low-variability states, which showed that neural variability and arousal (as measured by pupil size) differentially affect perceptual performance.

However, this metric is not readily applicable to calcium imaging, as their (spatially masked) EEG signal had only 1 spatial dimension and very high temporal resolution (1 kHz), whereas (our) calcium recordings typically have 100-1000 spatial dimensions (cells) but much lower temporal resolution (30 Hz, including auto-correlated calcium decay dynamics). Hence, for our recordings, a metric of neural variability is required that uses fewer, or just a single, time points (Figure 1.5E), and can be evaluated on the 1D array of neural responses, for each individual trial and time point. For this, we propose population variance, as further motivated and explained in Chapter 3.

1.5 Predictive learning in recurrent neural networks

1.5.1 Predictive learning

Animals learn to create internal representations of their environment, even if these are not directly associated with attaining rewards (Tolman, 1948). How might an organism develop

an internal model of their environment in the absence of reward? A candidate mechanism is predictive learning (Bar, 2009; Friston et al., 2016; Alexander and Brown, 2018; Nagai, 2019; Whittington et al., 2020; Recanatesi et al., 2021) which may arise even through passive exposure (Nagai, 2019). In predictive learning, animals continuously attempt to predict their future state, given their current state and actions. The utility of such models that predict future states has been extensively studied in reinforcement learning. In contrast to 'model-free' agents (that only reinforce actions that have previously led to rewards), 'model-based' agents can rapidly adapt to changes in the environment or reward structure, instead of starting from scratch (Daw et al., 2011; Russek et al., 2017; Wayne et al., 2018). Further, predictive learning in recurrent neural networks (RNNs) can extract the latent structure of the environment during a period of unrewarded exploration, actualising the proposal that predictive learning acts as a helper function in the brain by encoding the environment (Whittington et al., 2020; Recanatesi et al., 2021).

Neuroscience experiments have revealed that passive learning of the environment may indeed encode spatial structure in the brain, as observed in the hippocampus (O'Keefe and Dostrovsky, 1971), but also non-spatial sequences of perceptual states, which can be detected even in primary sensory cortex (Libby and Buschman, 2021). Specifically, Libby and Buschman (2021) showed that neural activity in primary auditory cortex predicted upcoming chords during chord sequences that mice repeatedly experienced (1,500 times per day, for four days), without any form of reward/behaviour structure in this task. Stimuli were split into 'expected' sequences (77% of relevant trials) and 'unexpected' sequences (remainder). Mice learned to associate the chords in expected sequences, and as a result neural activity would predict upcoming activity. (Additionally, Libby and Buschman (2021) found that over the course of learning, stimulus representations rotate orthogonally in neural state space to prevent interference between incoming sensory stimuli and memories of previous stimuli).

This passive learning task was thus able to reveal the existence of passive predictive learning in sensory cortex, but inherently cannot address what the utility of actively maintaining predictions

of upcoming sensory inputs is. In Chapter 4 we address this question, by focusing on a working memory task, which specifically requires maintaining information across time, and ask if and how actively maintaining sensory predictions across a time delay can improve learning.

1.5.2 Studying neural mechanisms with recurrent neural networks

To study the role of predictive learning in cortical networks of neurons, we turn to artificial Recurrent Neural Networks (RNNs). Artificial neural networks are a (simplified) model of biological neuronal networks, which allow researchers to record, perturb and alter neural dynamics unmatched by *in vivo* experiments (Barak, 2017). Whereas conventional (feedforward) deep learning models propagate stimulus information through a network of connected hidden layers (LeCun et al., 2015), RNNs combine information of both stimuli and the network state through recurrent connections within one hidden layer. This property makes them particularly useful for studying mechanisms and tasks that involve working memory (Barak, 2017; Vyas et al., 2020). Therefore, RNNs have now become an indispensable tool to study working memory mechanisms, for example on gating context information (Mante et al., 2013), Bayesian computation (Sohn et al., 2019), movement planning (Hennequin et al., 2014), low-dimensional neural dynamics (Mastrogiuseppe and Ostojic, 2018; Pollock and Jazayeri, 2020; Dubreuil et al., 2022), neural codes (Chaisangmongkon et al., 2017; Orhan and Ma, 2019; Masse et al., 2019) and multi-task learning (Yang et al., 2019; Duncker et al., 2020; Driscoll et al., 2022).

In Chapter 4, RNNs allow us to record the activity of all neurons, stimulate the neurons directly, observe the changes in the network over the course of learning, and compare networks with or without predictive learning.

2
Inhibition Stabilisation

A challenge in quantifying signal propagation is that correlated activity in two regions may not reflect causal influence of one on the other, but rather may reflect common input. Taking causal control of neural activity by directly stimulating neurons circumvents this issue, as activity locked temporally to stimulation of a connected region likely arises from propagation. This can be achieved by driving spikes directly through stimulation of individual neurons (Houweling and Brecht, 2008; Chettih and Harvey, 2019) or groups of neurons (Marshel et al., 2019; Dalgleish et al., 2020). However, so far, detection of optogenetic stimulation with single-cell precision has only been paired with single-region neural recordings, limiting signal propagation studies to within-region dynamics.

Here, we harness the single-cell precision of all-optical interrogation (Packer et al., 2015) to investigate how the collective dynamics of neural activity determine activity propagation and guide behaviour. We imaged two hierarchically organised, densely interconnected and functionally well-characterised (Kwon et al., 2016; Yamashita and Petersen, 2016; Chen et al., 2016) regions, the primary and secondary somatosensory cortex (S1 and S2) while performing

Corresponding publication: Rowland*, J. M., van der Plas*, T. L., Loidolt*, M., Lees, R. M., Keeling, J., Dehning, J., Akam, T., Priesemann, V., and Packer, A. M. (2023). Propagation of activity through the cortical hierarchy and perception are determined by neural variability. *Nature Neuroscience*, 6:1584–1594. *: co-first author.

two-photon optogenetic photostimulation of S1 neurons, and trained mice to report stimulation. By recording neural activity occurring before and after photostimulation in both brain regions simultaneously, we were able to assess how behaviourally salient stimulation is propagated through anatomically distinct brain regions and interacts with their ongoing activity.

2.1 Results

2.1.1 Causally driving behaviour with all-optical interrogation across cortical areas

We developed a preparation that allowed us to record activity in two brain regions (primary and secondary somatosensory cortex, or S1 and S2 respectively) simultaneously, while holographically photostimulating S1. To achieve this, we expressed the genetically encoded calcium indicator GCaMP6s (Chen et al., 2013) and the somatically targeted, red-shifted opsin C1V1-Kv2.1 (Yizhar et al., 2011; Chettih and Harvey, 2019) in layer 2/3 across both S1 and S2. A field of view was selected which spanned the cortical representations of multiple whiskers (across the A, B and C rows, see Diamond et al. (2008)) in both S1 and S2. Mice were head-fixed and we utilised a two-photon microscope, based on previous designs (Packer et al., 2015), adapted to perform two-photon calcium imaging of neural activity across a large field of view (1.35 mm diameter, allowing the recording of thousands of neurons), while performing two-photon photostimulation of S1 neurons. We first verified that targeted neurons, defined as cells within 15 μm of the centre of a photostimulation beam, elicited significant excitatory responses while nearby S1 non-target neurons elicited significant inhibitory responses on average (Figure 2.1a, 2.2a,b, Wilcoxon signed-rank test, $p < 0.05$, Bonferroni corrected). Targeted cells were activated similarly, with the notable exception of Hit-only inhibition at $\sim$ 2 seconds post-stimulus in 3 recording sessions (of the same animal), indicating that the photostimulation-induced activation is largely independent of behavioural outcome (2.2a).

Mice were trained via operant conditioning to associate direct photostimulation of neurons in S1 with water reward. After initial training, mice reported photostimulation by licking a water spout (Figure 2.1b). Crucially, to prevent a guessing strategy and/or random licking, mice had

to withhold licking to initiate a trial after a variable inter-trial interval. Successful report of stimulation on 'go' trials was scored as a 'hit' and the mouse was rewarded; trials in which the animal failed to report the stimulation through licking were scored as a 'miss'. We also randomly interleaved 'catch' trials in which no stimulus was delivered to analyse the detection performance relative to chance. Catch trials in which the animal licked were 'false positive' (FP) trials while trials when the animal appropriately refrained from licking were 'correct rejection' (CR) trials (Figure 2.1b). Finally, we also analysed 'reward only' trials (Methods, section 2.2), in which reward was delivered in the absence of photostimulation (prior to the task, for each recording session). In each trial, the first lick was defined to be the response time, plotted in Figure 2.2c, d. Between 5 and 150 photoresponsive S1 neurons were targeted on a single go trial. The number and identity of cells targeted were varied randomly on each trial to generate a psychometric curve (Figure 2.1c), which shows that stimulating a greater number of neurons increased licking behaviour accuracy. The 50% point of the psychometric curve fit across all sessions is 22 ± 1 neurons ($N = 11$ sessions recorded from 5 mice, $\pm$ 95% confidence interval), in accordance with previously published findings (Dalgleish et al., 2020). Thus, we have developed a preparation in which activity that causally drives behaviour can be recorded, with single-cell resolution, both locally and, after propagation downstream.

2.1.2 Two-photon photostimulation of S1 drives balanced activity that propagates to S2

To understand how neural activity is propagated between brain regions to guide behaviour, we compared neural responses during hit and miss trials, both in the directly stimulated region (S1) and the downstream region (S2). Qualitatively, hit trials elicit a diverse array of relatively strong excitatory and inhibitory responses both in S1 and S2, whereas both excitatory and inhibitory responses were less strong on miss trials (Figures 2.3a, 2.4, 2.5). Excited and inhibited cells are also observed in reward only trials, however, the neuronal responses are less pronounced compared to hit trials, indicating that the response to reported stimulation is not just explained by neural activity driven by reward (Figures 2.3a, 2.4, 2.5)). We later quantify this using a decoding analysis in section 2.1.3.

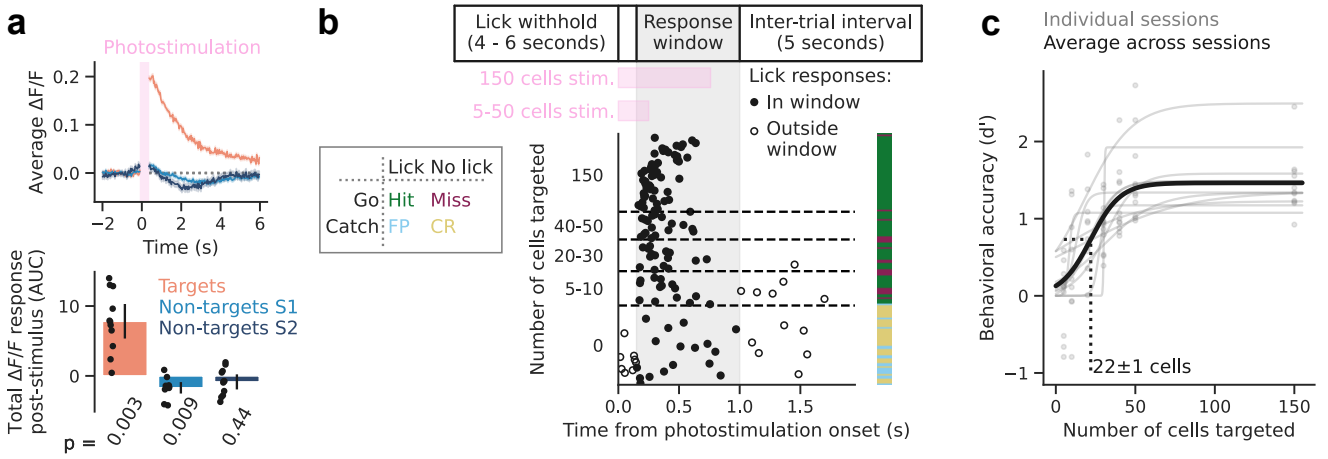


Figure 2.1: Recording neural activity in S1 and S2 during behavioural report of targeted two-photon photostimulation of S1.
a) Top: Example activity responses to photostimulation of a single recording session. Orange shows the response to photostimulation of cells directly targeted with light, averaged across cells and across trials. Light and dark blue show the response of cells not directly targeted in S1 and S2 respectively. Only trials in which photostimulation was delivered were included. Data are blanked while the photostimulation laser was on (pink bar) as this causes a large artifact unrelated to neural activity. Bottom: the total $\Delta F/F$ activity post-stimulus is shown, as defined by the Area Under the Curve (AUC) of the traces of the top panel, for all 11 recording sessions (mean $\pm$ 95% confidence interval across sessions). We tested whether each condition was significantly different from 0 (two-sided Wilcoxon signed-rank test, Bonferroni corrected). **b)** Top: Timing of a single behavioural trial. Bottom left: Behavioural response matrix. Bottom right: Example lick raster from a single session sorted by number of cells targeted and by time within each bin. Each row of the plot shows the first lick within an individual trial. The colour bar shows the outcome of the trial as defined in the behavioural response matrix. **c)** Psychometric curves showing behavioural performance (d') as a function of the number of cells targeted by photostimulation. Each grey point is the d' computed for a given number of cells targeted for an individual session and each grey line is a logistic function fit for an individual session. The thick black line shows the fit for all data points across all sessions ($N = 11$ sessions; $N = 5$ mice). The grey dashed line shows that the 50% point from the fit across all sessions occurs at 22 ± 1 cells targeted.

We then determined the number of (either excitatory or inhibitory) responsive cells for each trial, by comparing the the distribution of $\Delta F/F$ values 500 ms pre-stimulus to 500 ms post-stimulus for each cell. A cell determined to be significantly responsive (on an individual trial) if it passed a significance threshold of p = 0.05, using a two-sided Wilcoxon signed-rank test, following false discovery rate (FDR) correction. We observe that hit trials elicit a significantly greater fraction of responsive cells than any other trial type in S1 (one-sided Wilcoxon signed-rank test, $p < 0.05$ for hit vs all other trial types, Bonferroni corrected; Figure 2.6a). In S2, the fraction of responsive cells was more similar across trial types, suggesting that stimulus information is not encoded

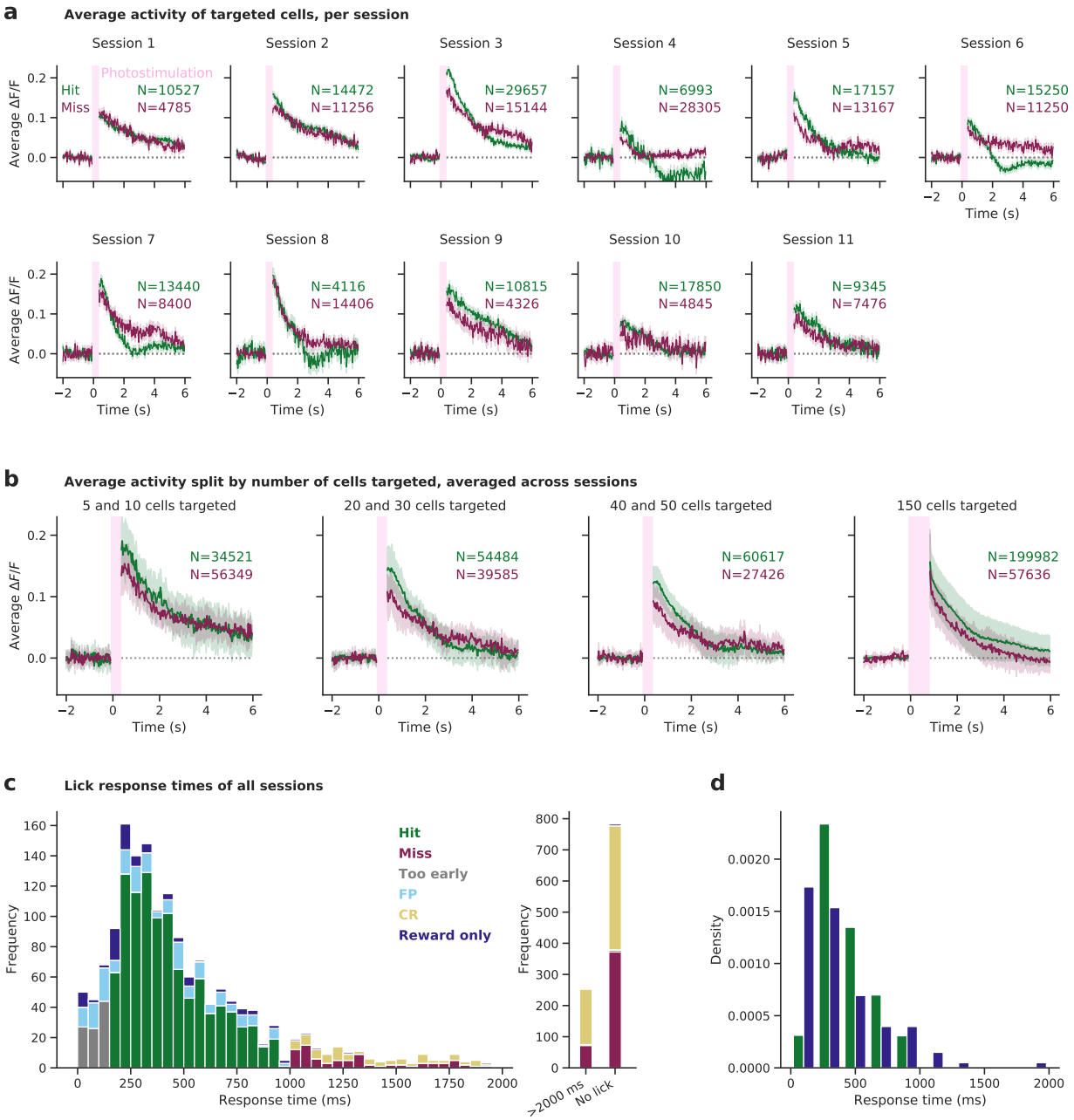


Figure 2.2: Activation of targeted cells by photostimulation.
a) Average $\Delta F/F$ activity of the cells targeted by photostimulation (mean values $\pm$ SEM) per recording session. All trials with 5 to 50 cells targeted were used; trials where 150 cells were targeted were excluded because almost all of these trials resulted in a Hit outcome, biasing the averages. **b)** Same data as panel a, but averaged across sessions and split by number of cells targeted (mean values $\pm$ SEM). **c)** Stacked histogram of the animals' response times (defined by the time of the first lick) for each trial type. **d)** Density histograms of response times of hit and reward only trials. Same data as in panel a, but normalised per trial type and binned per 200ms. Medians (248ms for reward only trials and 383 ms for hit trials) are significantly different (p=0.01, (two-sided) Mood's median test).

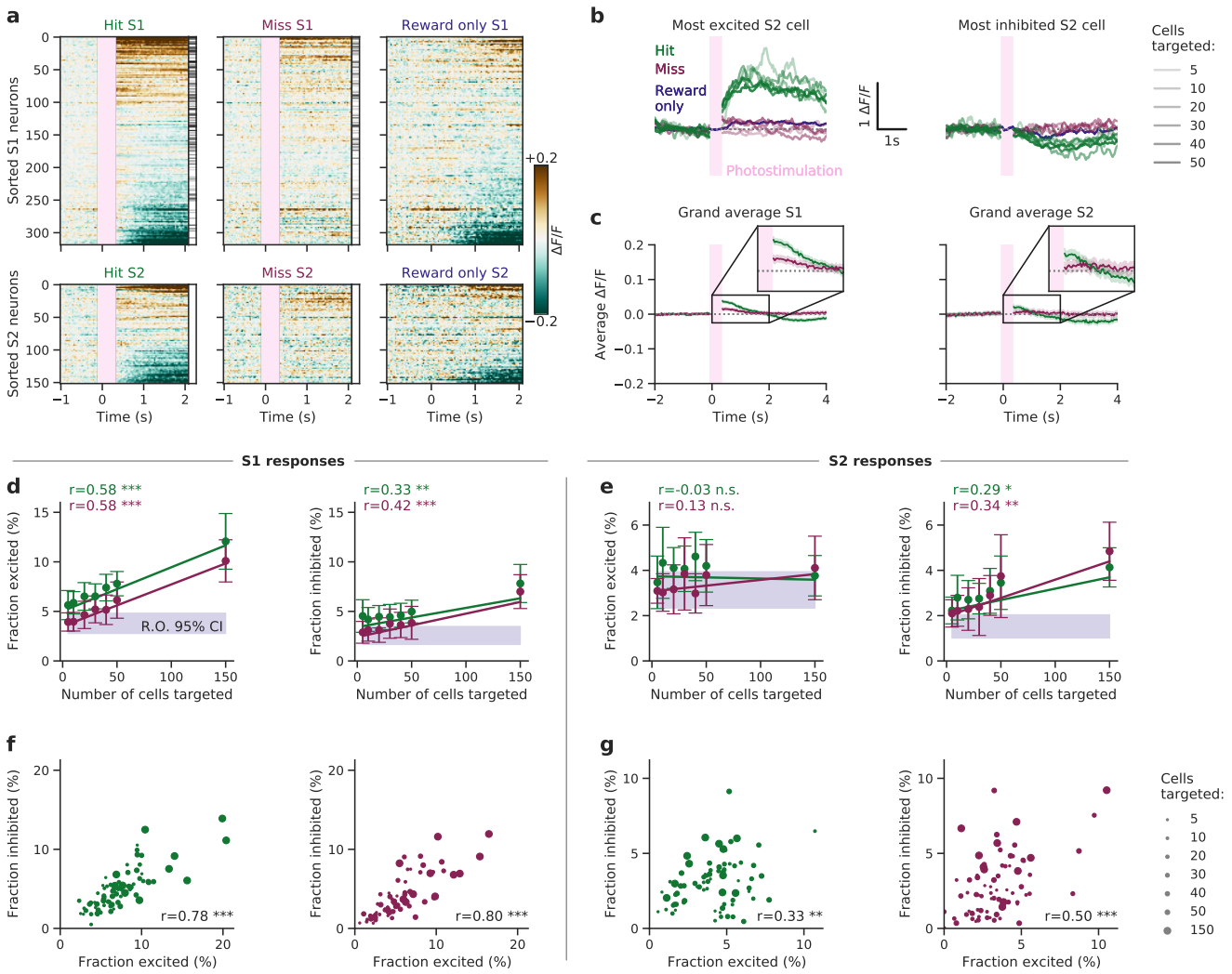


Figure 2.3: (Legend on next page)

by this relatively simple metric (Figure 2.6a, one-sided Wilcoxon signed-rank test, Bonferroni corrected; $p < 0.05$ for hit vs 2 out of 4 trial types). However, on a single cell level, we find some individual neurons in S2 that are excited or inhibited in response to hit trials, but not miss or reward only, hinting that reported stimuli are propagated to single neurons in S2 (Figure 2.3b).

Interestingly, however, averaging responses across cells, trials, and sessions reveals that hit trials elicit only a modest increase in population activity from baseline in both S1 and S2 (peak post-stimulus $\Delta F/F = 0.038 \pm 0.004$ and 0.022 ± 0.006, respectively) followed by a prolonged period of inhibition, while on miss trials the population average response did not deviate from

Figure 2.3: Activity driven by targeted two-photon photostimulation is propagated from S1 to S2. **a)** Trial-averaged neural responses of different trial types to photostimulation (pink vertical bar) and/or reward from a single session, sorted by the sum of their responses (Methods, section 2.2). The intensity of the grey-scale bar on the right hand side of the hit and miss plots is proportional to the number of times each cell was targeted, for hit and miss trials separately. Trials in which 150 cells were targeted were removed for display because their stimulation period is longer. Data are bound between -0.2 and +0.2 $\Delta F/F$. **b)** Left: $\Delta F/F$ activity traces of the most excited S2 cell (of the session of panel a), averaged across stimulus conditions. This S2 cell shows a large response on hit trials (green) but no response on miss trials (red) or reward only trials (blue). Right: Equivalent plot for the most inhibited S2 cell of panel a. **c)** The average population response to hit and miss trials across all sessions is shown (shaded areas show 95% confidence interval across trials of all sessions). Traces are averaged across cells, trials and sessions for a given trial type. Trials in which 150 cells were targeted were removed for display. **d)** The fraction of excited (left) and inhibited (right) cells in S1 as a function of the number of cells targeted on both hit and miss trials. Data are presented as mean values $\pm$ SEM. The shaded purple bar shows the 95% confidence interval across sessions of the fraction of excited or inhibited cells in S1 on reward only (R.O.) trials. The linear fit was determined using Weighted Least Squares, where the weights were the inverse variance of the trials that constituted a data point, and subsequently bound between their $25^{\text{th}}$and 75^{th}percentiles to prevent extreme weight values. P values were computed using a two-sided t-test, where significance is indicated by: *** ($p < 0.001$), ** ($p < 0.01$), * ($p < 0.05$) or n.s. (not significant). **e)** Equivalent panel for S2 responses. **f)** Same data as in panel d, but plotted per session and hit (left) and miss (right) trial type. The size of the circle indicates the number of cells photostimulated. The Pearson correlation coefficient is denoted by r, with significance indicated as before (two-sided t-test). **g)** Same but for S2.

baseline in S2 (Figure 2.3c). The population responses of all other trial types are shown in Figure 2.6b,c.

Taken together, the population responses of S1 and S2 suggest that excitatory stimulation is balanced by an inhibitory response, also supported by the relatively weak grand average. Therefore, we analysed the excited and inhibited cells separately. As before, we considered the fraction of responsive cells, split by the sign of the responsive. In other words, significant *positive* post-stimulus responses were said to be significantly *excited* cells, and vice versa for inhibited cells. We could then analyse the fraction of excited and inhibited cells, as a function of stimulus strength (*i.e.*, number of cells targeted). As expected, the inhibited responses increased with the strength of optogenetic stimulation (Figure 2.3d,e right; correlation values reported in the figure). In S1, but not in S2, the excitatory response also increased with stimulation strength (Figure 2.3d,e left). This may reflect the strong contribution of the targeted cells in S1.

To quantify this effect directly, we analysed the excitation-inhibition ratio across all sessions. We find a strong correlation between the fraction of excited and inhibited cells for both hit and miss trials in S1, and a slightly weaker correlation in S2 (Figure 2.3f,g; correlation values reported in the figure). This indicates that the excitation-inhibition ratio was maintained regardless of photostimulation strength or whether the stimulation was perceived. However, downstream from the stimulation, in S2, the correlation between excitation and inhibition was weaker than in S1 (excitation-inhibition correlation of S1 hit versus S2 hit, $p = 1.2 \cdot 10^{-5}$; S1 miss versus S2 miss, $p = 0.001$; Two-tailed Steiger test; Figure 2.3f,g). Taken together, we show that as activity is propagated from its site of initiation in S1 to S2, the excitation-inhibition interplay persists but is less tight.

2.1.3 S1 and S2 populations encode representations of perceived photostimulation

Individual cells displayed a diverse array of responses, and balanced population responses masked the magnitude and time course of propagated activity (Figure 2.3). To gain insight into what task-related information is contained in these complex neural responses, we assessed propagation of stimulus-relevant activity from S1 to S2 and its relationship to behaviour. To this end, we trained classifiers to dynamically decode the trial type of individual trials using the activity of all cells in S1 or S2 separately (Figure 2.7a). We used logistic regression classifiers that were trained to distinguish two trial types, for each time point separately. Classifiers were trained on 75% of trials and evaluated on the remaining 25%, and 4 classifiers were trained from scratch (for each time point) using cross-validation such that all trials were evaluated exactly once. All results and figures only use the concatenated test data (withheld from training). For more details, see Methods, section 2.2.

First, we trained classifiers that distinguished hit trials from correct rejection trials in order to detect the neural signature of both the stimulus and the activity resulting in its perception and report. Trained independently on each time frame, the classifiers performed significantly above chance for more than three seconds after stimulation, both in S1 and S2 (two-sided Wilcoxon signed-rank test, $p < 0.05$, Bonferroni corrected for number of tested time points; Figures 2.7b,c;

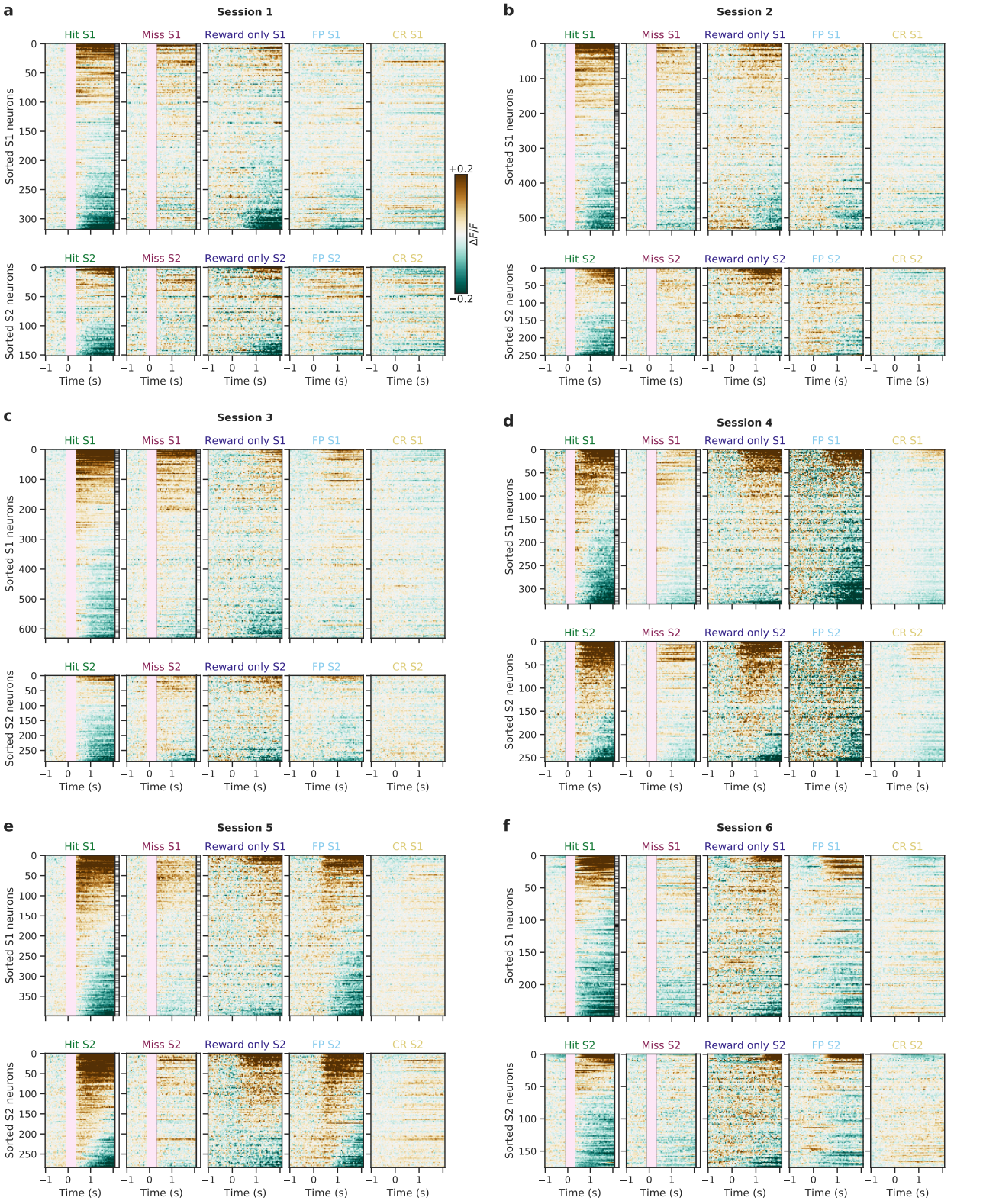


Figure 2.4: Trial-averaged raster plots of trial types for first 6 sessions.

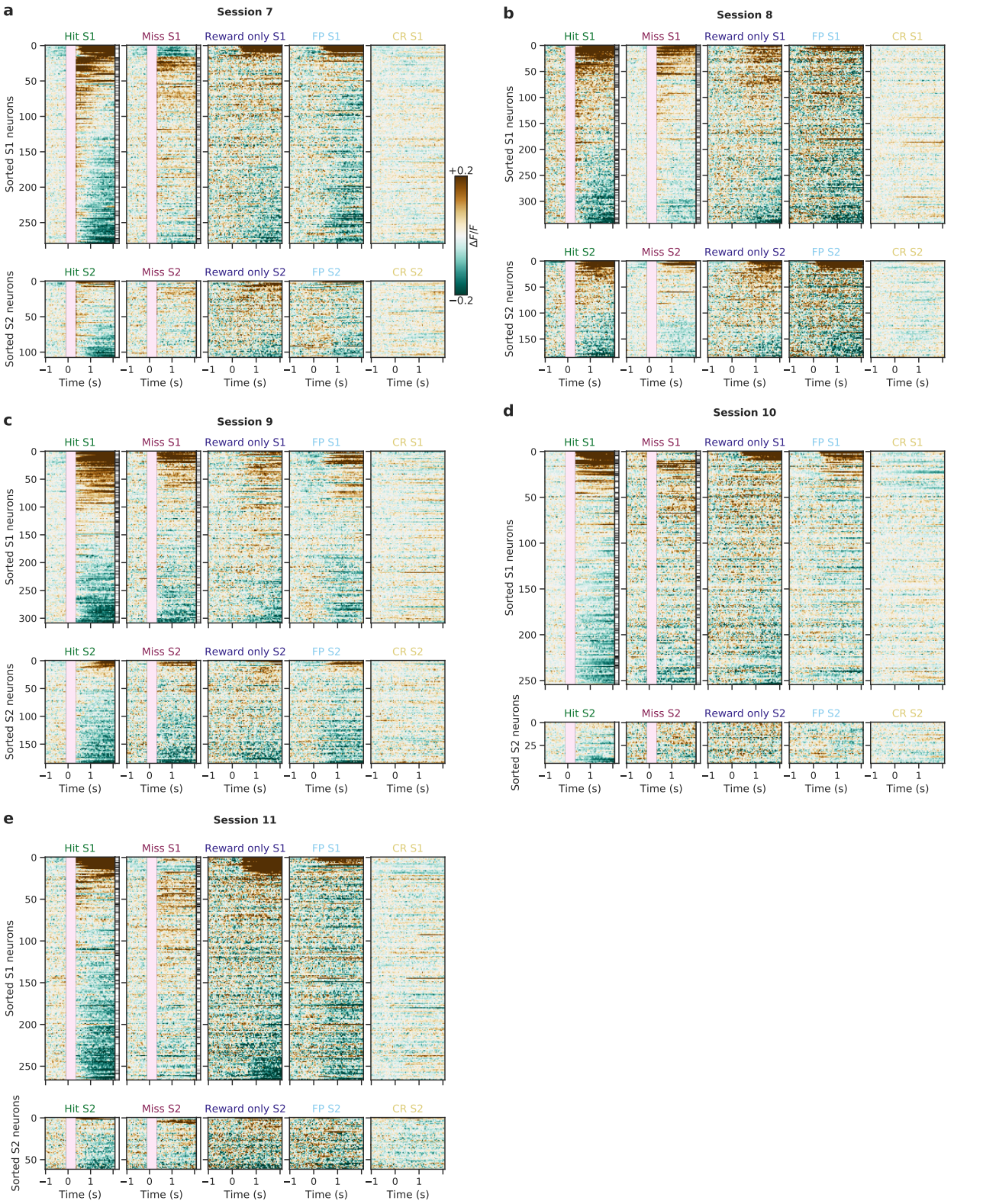


Figure 2.5: Trial-averaged raster plots of trial types for next 5 sessions.

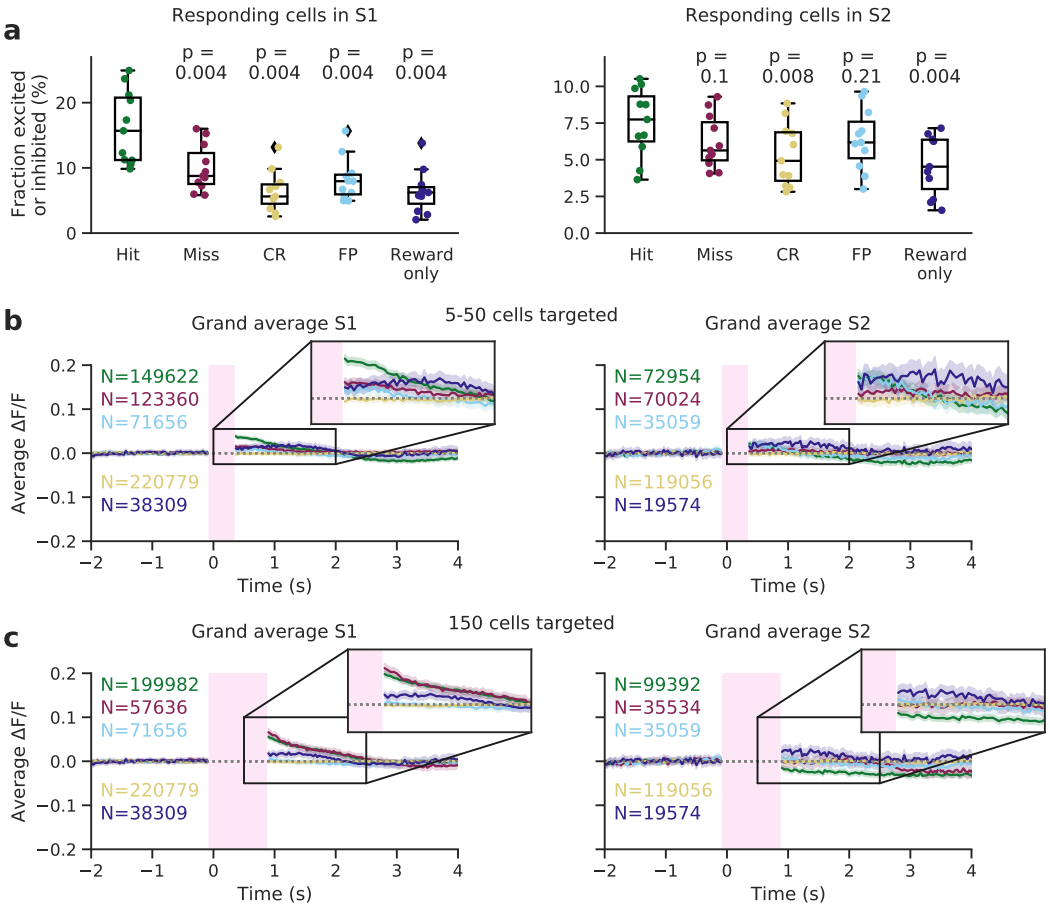


Figure 2.6: Fraction of cells responding and grand average traces of all trial types in S1 and S2. **a)** The fraction of responsive cells (both excited and inhibited) on each trial type across all numbers of cells targeted. One-sided Wilcoxon signed-rank tests were used to test whether the fraction of responsive cells of hit trials was significantly greater than on other trial types (Bonferroni corrected for 8 tests). $N = 11$ independent recording experiments. Boxplots are defined by the median (centre), interquartile range (IQR) box, 1.5 IQR whiskers plus any outliers. **b)** Average population responses of all trial types across all sessions. Traces are averaged across cells, trials and sessions for a given trial type. Trials in which 150 cells were targeted were removed for display (and shown in c). The time course of reward only trials is different from the time course of hit trials, hinting that the neural activity on hit trials constitutes more than motor preparation, movement, and reward related activity. This is further quantified in Figures 2.3, 2.7. **c)** As above but showing exclusively trials in which 150 cells were targeted.

2.8a,b; 2.9a,b). This shows that neural activity underpinning perceived stimulation persisted both locally in S1 and downstream in S2 for several seconds, indicating that activity that arose from perceived stimulation in S1 was propagated to S2.

However, the classifiers could just have decoded the signatures of reward or the motor command

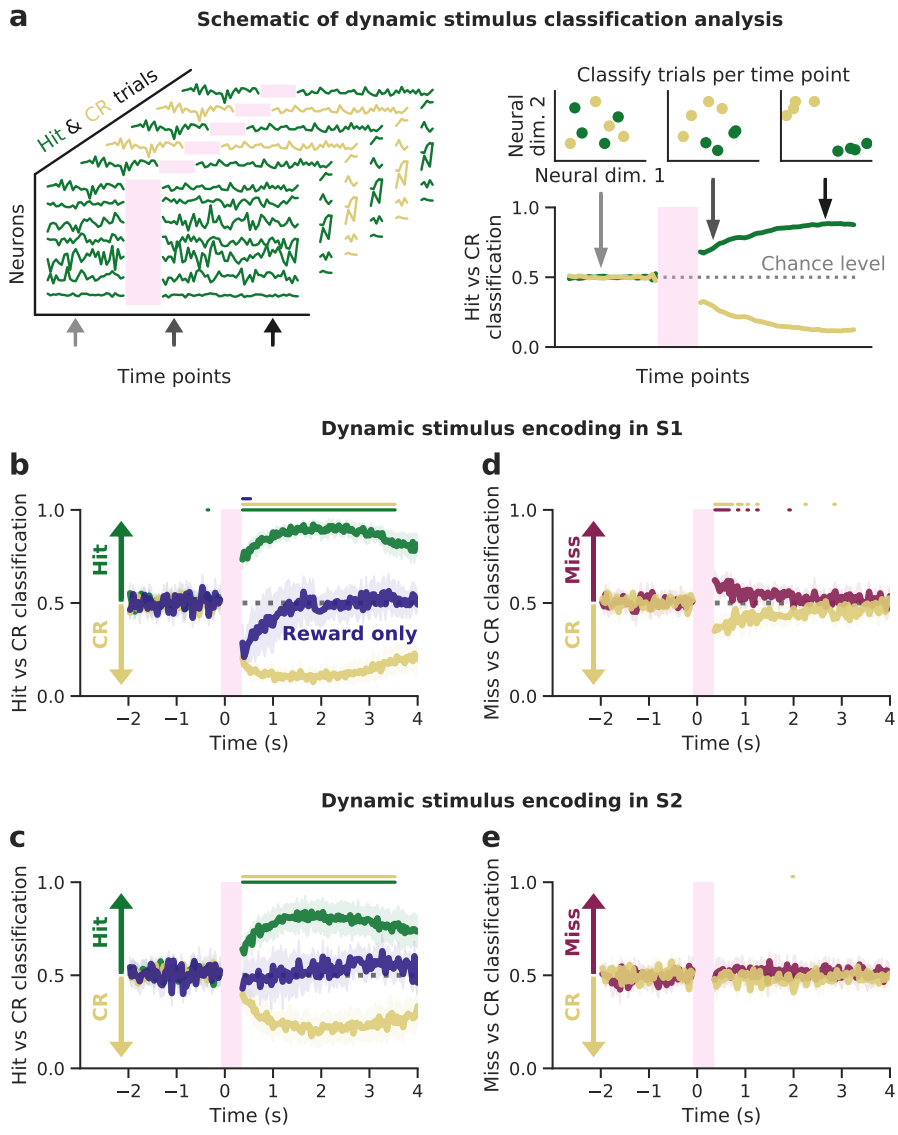


Figure 2.7: Perceived photostimuli elicit persistent activity in both S1 and S2 populations.
a) Illustration of dynamic trial type decoding in the neural population. Classifiers were trained on each time frame individually, with activity of all cells in S1 or S2, and tested on held-out data. **b)** Classifiers were trained, for each time point, on S1 activity to classify hit trials from correct rejection trials and then tested on held-out hit trials (green), correct rejection trials (yellow) and reward only trials (blue). Classifications are presented as mean values $\pm$ SEM across $N = 11$ sessions. Coloured bars above the traces show time points at which classifier performance was significantly different from chance (two-sided Wilcoxon signed-rank test, $p < 0.05$, Bonferroni corrected, see Methods, section 2.2). **c)** Same, but for S2. **d)** S1 classifiers trained to classify miss trials from correct rejection trials, and then tested on miss trials (red) and correct rejection trials (yellow). **e)** Same, but for S2.

required for licking, which are present on hit trials but not correct rejection trials, instead of decoding stimulus perception. To disambiguate this, we first tested the classifiers (that had been trained to distinguish hit versus correct reject trials) on 'reward only' trials (*i.e.*, trials in which reward was delivered in the absence of photostimulation). Performance on reward only trials was either below or at chance level in S1, and remained at chance throughout the trial in S2, indicating that neural signatures of motor commands and licking were not the driving features

of the hit/correct rejection classifiers (Figure 2.7b,c; blue lines; Figure 2.8a, b, $p > 0.05$, two-sided Wilcoxon signed-rank test, Bonferroni corrected).

Secondly, we investigated the possibility that reward-related responses are different on hit and reward only trials, in principle possible because the water reward on hit trials was expected, but not expected on reward only trials. We addressed this by training a model to classify reward only trials from correct rejection trials, thus this classifier should detect only the neural signal from rewards, and related motor activity. Testing this model on hit trials resulted in identical performance to the reward only condition both in S1 and S2 (Figure 2.8e, f), evidenced by no significantly different decoding accuracies at any time point in S1 or S2 ($p > 0.05$, two-sided Wilcoxon signed-rank test, Bonferroni corrected). Hence, classifiers trained specifically to detect rewards could not distinguish hit and reward only trials, meaning that the neural response to reward is comparable in both trial types and the classification of hit trials from correct rejection trials is indeed decoding the neural signature of perceived stimulation, both in S1 and S2. Finally, we repeated the hit vs correct rejection decoding analysis using the same subsampling procedure as reward only vs correct rejection decoders (see Methods, section 2.2), and confirmed that results were equivalent to before (Figure 2.8g, h).

Next we asked whether neural activity from unperceived photostimulation was reliably propagated from S1 to S2 by training models to classify miss and correct rejection trials (Figure 2.7d, e; 2.8c, d; 2.9c, d). We found that in S1, these classifiers decoded miss trials slightly above chance immediately following stimulation ($p < 0.05$, Bonferroni corrected), however performance rapidly decayed back to chance level after 1 second. This implies that non-perceived stimuli do not generate persistent stimulus representations. In S2, the classifiers were not able to decode miss trials, indicating that stimulus information that is not perceived does not propagate to S2.

Taken together, these results show that only perceived stimulus representations are reliably propagated out of the local brain region in which they originate to a downstream brain area. Further, the representations of stimuli that are propagated and drive perception persist both

locally and downstream for several seconds following the injection of activity.

We further verified that the dynamic decoding results of Figure 2.7 were reproduced when larger time windows of 2 seconds were used (Figure 2.8i-l). Three windows were used (-2.0s up to and including -0.1s, 0.4s u/i 2.3s, 2.4s u/i 4.3s), shown at the bottom of each panel. For each window, neural activity was averaged across time, per neuron per trial. Classifiers were then trained and evaluated as before. This did not lead to different results: stimulus information was only propagated to S2 on hit trials. Additionally, we verified that the dynamic rise in S2 Hit/correct rejection decoding accuracy during the response window (0 ms - 1000 ms) of Figure 2.7c was not solely due to an increase of the number of trials that responded to the stimulation with time (Figure 2.6c,d). To do so, we considered the first decoder post-stimulus (at t = 367 ms), and compared the decoded prediction of individual hit trials (of withheld test data only, eventually evaluating all trials using the previously described cross-validation procedure) to their response times (Figure 2.8m). We found that no session exhibited a significant correlation (Figure 2.8n), with non-significant correlations spreading across both positive and negative values, showing that there was no consistent correlation between response time and decoding accuracy across sessions.

2.2 Methods

2.2.1 Experimental procedures

As the imaging experiments were conducted by others (see Author Contributions), I refer to our publication for full details on surgical methods, widefield imaging, two-photon calcium imaging, two-photon optogenetic stimulation, behavioural training, calcium imaging pre-processing and pupil size pre-processing (Rowland*, Van der Plas*, Loidolt* et al., 2023). However, in the below, I will summarise very briefly some of these aspects for the purpose of readability of this thesis. Unavoidably, this summary glosses over many details for the purpose of brevity, which can all be found in Rowland*, Van der Plas*, Loidolt* et al. (2023).

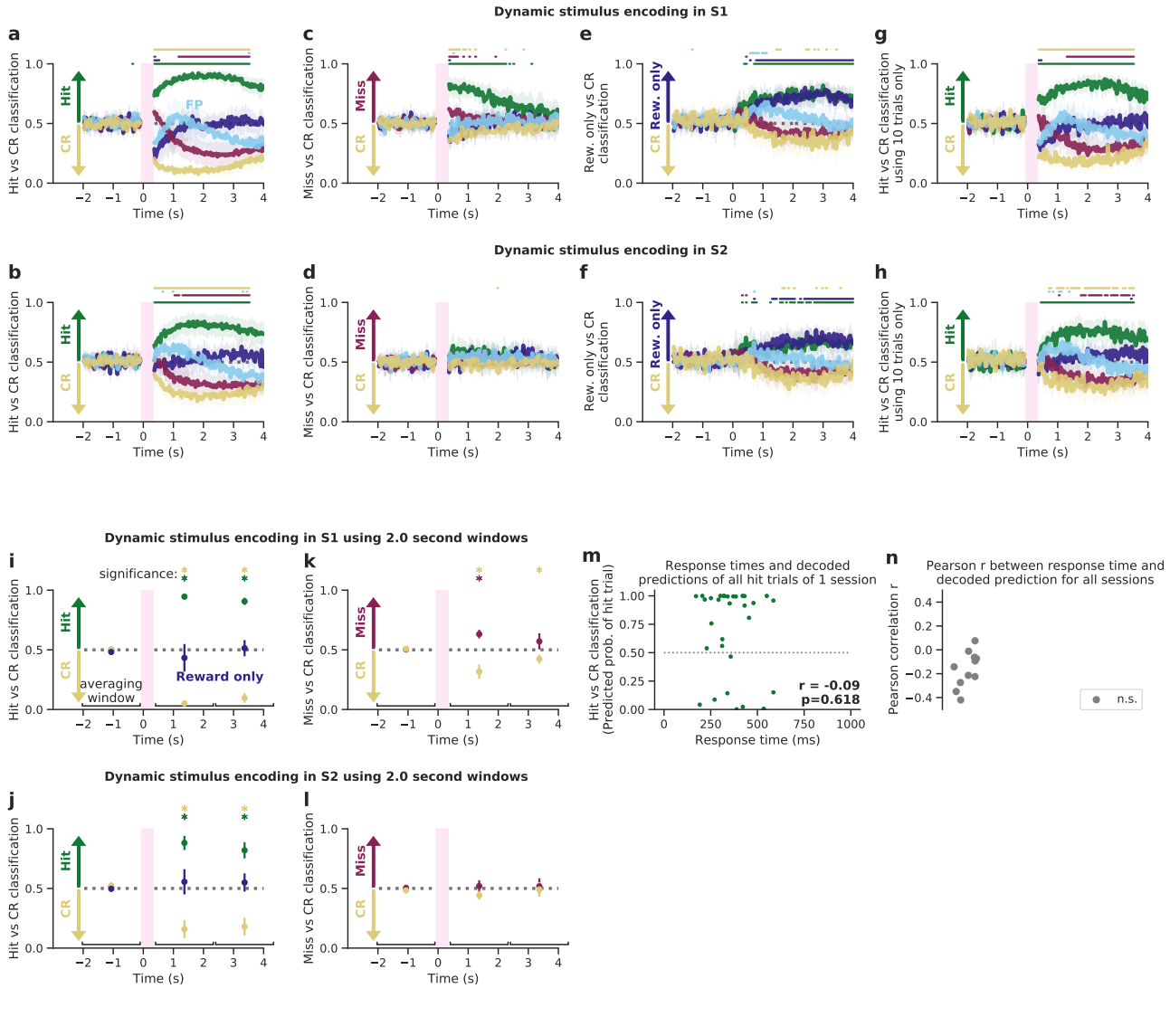


Figure 2.8: (Legend on next page)

All experimental procedures involving animals were conducted in accordance with the UK animals in Scientific Procedures Act (1986).

Behaviour and trial structure. Mice were trained to report optogenetic stimulation of S1 by licking a water spout, and correct responses (within 1 second of photostimulation) were rewarded with water. Prior to the imaging experiments, mice were trained to perform this task by starting with a strong (1-photon) stimulus, and gradually reducing stimulus intensity until

Figure 2.8: Dynamic decoders of all trial types.
a) Models were trained, for each time point, on S1 activity to classify hit trials from correct rejection trials and then tested on held-out hit trials (green), correct rejection trials (yellow), reward only trials (dark blue), false positive trials (light blue) and miss trials (red). Classifications are presented as mean values $\pm$ SEM across $N = 11$ sessions in panels a-l. Coloured bars above the traces show time points at which classifier performance was significantly different from chance (two-sided Wilcoxon signed-rank tests, p < 0.05, Bonferroni corrected, see Methods, section 2.2). **b)** As above, but trained on S2 activity. **c)** Models were trained on S1 activity to classify miss trials from correct rejection trials and then tested on all trial types. **d)** As above, but trained on S2 activity. **e)** Models were trained on S1 activity to classify reward only trials from correct rejection trials and then tested on all trial types. **f)** As above, but trained on S2 activity. **g, h)** As panels a and b, but the number of trials was restricted to 10 for each type, matching the total number of reward only trials recorded. **i-l)** Equivalent to Figure 2.7, but now using 2-second windows to train the classifiers. Asterisks indicate significant differences with chance level accuracy 0.5 (two-sided Wilcoxon signed-rank test, Bonferroni-corrected p value < 0.05). Classifications are presented as mean values $\pm$ SEM across $N = 11$ sessions. **m)** Comparison between lick response time and predicted outcome for all hit trials in S2. We considered the first decoder post-stimulus (at t=367 ms), and compared for each trial the predicted outcome evaluated on withheld test data to the response time on that trial. This panel shows one example session (same session as Figure 2.3a), and its Pearson correlation value r and associated (two-sided) p value. **n)** Pearson correlation values r between response time and decoded S2 predictions of hit trials of all 11 sessions are shown. A Bonferroni multiple-comparison correction of $N = 11$ (sessions) was used.

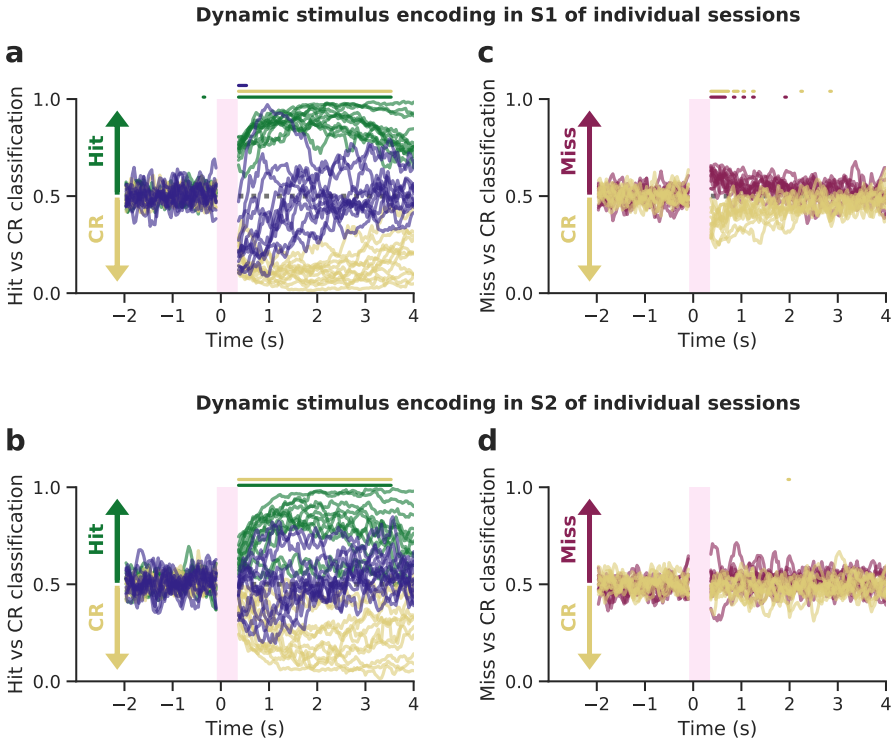


Figure 2.9: Dynamic decoders of individual sessions.
As Figure 2.7, but displaying classifier performance for each individual session. Traces were smoothed using a running mean of 5 time points for visual clarity.

they could perform this task with stimulation of tens of neurons.

During the imaging experiments, no stimulation was delivered ('catch' trials) on 1/3 of trials. On another third, 150 cells were stimulated, and on the remaining third a number between 5 and 50 cells was stimulated. Each trial was separated by a five-second inter-trial interval, and started with a variable 4-6 lick withhold period to trigger the trial onset (photostimulation or catch) (Figure 2.1b). This both prevented mice from adopting a strategy of licking very often, and from basing their response on the timing of the task. If mice responded within 1 s of photostimulation, they were rewarded with water and the trial was labelled a 'hit' trial. If they didn't response (on time), they did not receive water and the trial was labeled as 'miss' trial. No reward was given on catch trials, which could either be 'false positives' (on responding) and 'correct rejections' (on not responding). Behavioural accuracy was quantified by d' (Figure 2.1c), which is defined by:

$$d' = z(\text{hit rate}) - z(\text{false positive rate}) \tag{2.1}$$

where z is the z-transform.

Prior to each photostimulation experiment, spontaneous activity was recorded during which 10 non-cued rewards were delivered in 10 second intervals. These were labeled 'reward only' trials, and were subsequently analysed.

Recording and perturbation of neural activity. Neural activity was recorded in S1 and S2 using two-photon calcium imaging, while simultaneously, neurons in S1 could be photostimulated using two-photon optogenetics (Packer et al., 2015). To achieve this, we expressed the genetically encoded calcium indicator GCaMP6s (Chen et al., 2013) and the somatically targeted, red-shifted opsin C1V1-Kv2.1 (Yizhar et al., 2011; Chettih and Harvey, 2019) in layer 2/3 across both S1 and S2. Note that in our experiments the calcium indicator was expressed in both excitatory and inhibitory neurons (via Synapsin), and the opsin was only expressed in excitatory cells (because CamKIIα is mainly expressed in excitatory (Pyramidal) neurons (Nicole and Pacary, 2020), see Rowland*, Van der Plas*, Loidolt* et al. (2023) for more information). In other words, we optogenetically stimulated excitatory cells, while we recorded calcium activity from both excitatory and inhibitory cells.

Up to 50 individual neurons could be stimulated simultaneously using a spatial light modulator, which split the laser beam. Before each recording experiment, photoresponsive cells in S1 were identified by online stimulus-triggered-average analyses of two-photon photostimulation. These then formed the pool of potential photostimulation targets during the recording experiment. Cells in each simultaneously targeted subset were no more than 350 μm apart. Cells were stimulated for 250 ms. To stimulate 150 cells, 3 groups of 50 cells were stimulated consecutively, leading to a total photostimulation period of 750 ms (Figure 2.1b).

Calcium imaging pre-processing analysis The resulting calcium imaging videos were pre-processed using Suite2p (Pachitariu et al., 2017) for cell segmentation and obtaining a fluorescence trace per cell. To remove contamination of surrounding neuropil, fluorescence was corrected by:

$$F(t) = F_{\text{soma}}(t) - 0.7 \cdot F_{\text{neuropil}}(t) \tag{2.2}$$

where the 0.7 constant was taken from Chen et al. (2013). Next, fluorescence was normalised (against baseline variations across cells) by:

$$\Delta F/F(t) = \frac{(F(t) - \langle F(t) \rangle_t)}{\langle F(t) \rangle_t} \tag{2.3}$$

denoted $\Delta F/F$ in short. Cells that exhibited artificially high $\Delta F/F$ (if $\max_t \Delta F/F > 10$) were discarded from further analysis.

For the purpose of analysis, trials were defined as 2 seconds preceding to 8 seconds following the onset of photostimulation (which was set to be $t = 0$, *e.g.*, Figures 2.7, 2.4, 2.5). Frames during the photostimulation period (250 ms for 5-50 cells, 750 ms for 150 cells stimulated) were excluded from analysis and visualisation because of laser-induced artificial imaging signals. This is indicated by the pink bar in all relevant figures. (Note that due to the slow decay of the GCaMP6s calcium sensor, calcium events are typically longer than the 250 ms window).

2.2.2 Behavioural data analysis

As imaging was stopped intermittently, neural activity was not recorded for every trial performed by the animal; trials that were not imaged were excluded from all analysis. Hit trials in which the animal licked with an exceptionally short-latency (< 150 ms) are likely to have been driven by random licking rather than perception of the stimulus and were thus marked as 'too-soon' and not included in further analysis (Dalgleish et al., 2020). Further, after three consecutive misses mice were 'auto-rewarded' in order to keep them engaged with the task. These trials were excluded from further analysis too. Finally, on rare occasions (33 out of 1043 total hit trials), lick responses just before the end of the response window (928 - 985 ms interquartile range from photostimulus onset) were registered just outside the response window and therefore not rewarded. These were also excluded from all further analyses.

For trial-wise analysis of neural activity, each trial's time series was aligned to the onset of the (sham) photostimulation (time = 0 s) or to the onset of reward for the reward only condition.

Psychometric curves were fit to behavioural data by computing the value of d' separately for trials in which a given number of cells was targeted (Figure 2.1c). This was achieved by comparing the hit rate for a given number of cells targeted to the false positive rate across all catch trials. Data was fit using a logistic function adjusted lower bound at $d' = 0$.

2.2.3 Normalisation and sorting of post-stimulus neural activity

Post-stimulus neural activity was baselined relative to pre-stimulus activity in order to assess the relative change in activity after the photostimulation period, and to compare this change across cells and trials. On each trial, for each individual cell, the average $\Delta F/F$ activity in the 2 seconds preceding the photostimulation was subtracted from the post-stimulus activity trace. This normalisation procedure was applied to all analyses and visualisations of post-stimulus neural activity, except for Figures 2.3d-f, 2.6a where the difference between pre-stimulus and post-stimulus activity was assessed. Neurons were sorted for visual clarity only (in Figures 2.3,

2.4, 2.5), using the sum of the post-stimulus $\Delta F/F$ activity on hit, miss and reward only trials. This yields a sorting from strongly inhibited to strongly excited cells.

2.2.4 Definition of targets, photostimulation responsiveness, and excited/inhibited neurons

Neurons were defined as 'targeted' on an individual trial if any part of their cell body was located within 15 μm of the centre of the photostimulus beamlet. Neurons were categorised as responsive or unresponsive to stimulus or reward in a trial-wise manner. For each trial, the distribution of $\Delta F/F$ values 500 ms pre-stimulus were compared to 500 ms post-stimulus for each cell. A cell was deemed as responsive to a stimulus on an individual trial if it passed a significance threshold of p = 0.05, using a two-sided Wilcoxon signed-rank test, following false discovery rate (FDR) correction. The threshold of the FDR correction (0.015) was set empirically as the value which yielded $\sim$ 5% of cells responding on correct rejection trials, where there was no stimulation or licking response. Following significance testing, cells were split into positively and negatively responding cells based on whether the mean $\Delta F/F$ value post-stimulus was greater or less than the mean $\Delta F/F$ value pre-stimulus respectively. The fraction of neurons excited or inhibited on a given trial, in a given brain region, was defined as the number of positively or negatively responding cells respectively, divided by the total number of neurons recorded in that brain region on that session. The linear fit was determined using Weighted Least Squares, where the weights were the inverse variance of the trials that constituted a data point, and subsequently bound between their 25$^{\text{th}}$and 75$^{\text{th}}$percentiles to prevent extreme weight values (Strutz, 2011).

2.2.5 Dynamic decoding of stimulus representation

Logistic regression classifiers. The dynamic decoding classifiers of Figure 2.7 used a logistic regression model with an L2 penalty term to the weights with a regularisation strength of 0.001 (optimised by a parameter sweep with values 10^{-7} to 10^{3} with increments of factor 10). The Scikit-learn implementation of logistic regression was used (Pedregosa et al., 2011). Classification accuracy was computed per time point on a session-wise basis and then averaged across sessions, with shaded areas showing the 95% confidence interval of performance across sessions.

Each model was trained to classify the probability that a trial belonged to one of two different trial types. A 3:1 train:test split was employed and model performance was assessed on held-out test trials only. Trials were subsampled if necessary to prevent biases, see the next paragraph for more details. 4-fold cross validation was used on each session, with a new model trained for each fold for each time point, and classification accuracy is reported as the average of the test data across folds, meaning all (potentially subsampled) trials were in the the held-out test set exactly once. A new model was trained from scratch for each imaging frame within a trial, hence the training data consisted of a vector containing a single scalar $\Delta F/F$ value for each cell of all trials on a given frame.

Subsampling. Trials of the majority trial type class were randomly subsampled to adjust for sessions in which there was an unequal number of different trial types and ensure that models were not biased to a major or minor class. This meant that the reward only / correct rejection classifiers (Figure 2.8e, f) could only use 10 trials: the number of reward only trials. To ensure that this low number of trials did not underlie the difference in performance with the hit / correct rejection classifier, we also trained a hit / correct rejection classifier using 10 trials only. For this analysis, we sampled 10 hit trials (for each time point) according to the response time distribution of reward only trials, to also control for the difference in response times between the two trial types (Figure 2.2d). We found that hit predictions were significantly greater than reward only predictions on all post-stimulus time points in S1 and 56/110 post-stimulus time points in S2 (Figure 2.8g, h).

Statistical tests. Significance of decoding predictions was assessed by two-sided Wilcoxon signed-rank tests, where the predicted classification was compared to chance level (= 0.5). Predictions were binned per 2 time points, such that each test consisted of 22 (2 time points x 11 sessions) paired samples (prediction and the constant chance level). Predictions were said to be significant if their p value was below 0.05, Bonferroni corrected for the number of tests performed per trial type (83, both the pre-stimulus and post-stimulus activity time points,

excluding the photostimulation artifact). The binning of 2 time points was used because single time point statistics had too little data to be significant given the strong Bonferroni correction. Hence we chose the smallest bin size (2) through which it was mathematically possible to yield significant p values. Significant time points are indicated in Figures 2.7, 2.8a-h, 2.9 by thick lines, coloured by the trial type they refer to, at the top of each panel. To quantify the difference between predictions of two trial types, the same tests were performed comparing the two trial type predictions directly.

2.2.6 Data and code availability

All source data used for the data analysis and visualisation are publicly available on a GIN data repository. This includes the two-photon calcium imaging and optogenetics recordings of all experiments, as well as the pupil size data.

All data analysis and visualisation presented was performed in Python 3.7, which we have made publicly available on Github. In particular, all figures were made fully in Python, and the code is available on Github.

3
Neural Variability

Perception of optogenetic stimuluation depends on the number of cells stimulated (Marshel et al., 2019; Dalgleish et al., 2020), as we have replicated in Chapter 2. At the same time, we know that neural variability can influence perceptual performance in various ways (Waschke et al., 2021), as we reviewed in section 1.4. Therefore, in this Chapter, we ask if neural variability influences the likelihood of perception, and thus propagation of neural activity, independent of stimulus strength using the photostimulation experiments of Chapter 2. To do so, we propose an instantaneous metric of neural variability tailored to calcium recordings (high spatial resolution, low temporal resolution): population variance.

Next, we also analyse a separate set of experiments, where S1 neurons were stimulated without behaviour/reward (Figure 1.1, middle). We ask if propagation of activity to S2 still occurs in this passive stimulation setting, and if neural variability influences this.

Corresponding publication of sections 3.1.1 and 3.1.2: Rowland*, J. M., van der Plas*, T. L., Loidolt*, M., Lees, R. M., Keeling, J., Dehning, J., Akam, T., Priesemann, V., and Packer, A. M. (2023). Propagation of activity through the cortical hierarchy and perception are determined by neural variability. *Nature Neuroscience*, 6:1584–1594. *: co-first author.

3.1 Results

3.1.1 Pre-stimulus S1 activity predicts trial outcome

While the results of Chapter 2 clearly showed that stimulation of S1 can be reliably propagated to S2 and reported behaviourally, it is still unclear what intrinsic network conditions facilitate its detection. To identify the conditions that facilitate both the propagation to S2 and the behavioural report, we analysed population activity in S1 immediately prior to stimulation (Figure 3.1a) and asked whether the pre-stimulus population activity could predict whether the stimulus would be perceived (*i.e.*, whether the upcoming trial type would be a hit or a miss). To characterise population activity, we analysed the distribution of instantaneous activity across neurons and computed the population mean, population variance and temporal variance as scalar measures. The mean population activity (averaged per session) was not predictive of whether a trial would be a hit or miss ($p = 0.28$, Wilcoxon signed-rank test, Figure 3.1b left). In contrast, the population variance across S1 neurons strongly predicted whether the upcoming stimulus would be a hit or miss trial. Specifically, the population variance in S1 was larger prior to miss trials than prior to hit trials ($p=0.001$, Wilcoxon signed-rank test, Figure 3.1b right). Finally, temporal variance (*i.e.*, the variance across time, instead of across neurons) did not significantly differ between hit and miss conditions (Figure 3.2a, b, note that p values are shown without multiple comparison correction).

We further observed that population variance was correlated with task variables such as trial number and reward history (Figure 3.2), and that population variance in S1 was correlated with that of S2 (Figure 3.2e). In other words, as recording sessions progress, reward history declines, and population variance is correlated to this trend. This suggests that population variance in both, S1 and S2, may be a global mechanism underpinning how arousal and/or motivational state are instantiated in neural circuitry. The relationship between population variance and behavioural performance was also evident on a trial-by-trial level, whereby increased population variance pre-stimulus in S1 is negatively correlated with the probability that the upcoming stimulus elicits a hit (Figure 3.1c; one-sided regression Bonferroni corrected, test $p < 0.001$ on 10/11 sessions). Further, even though the population variance influenced the probability of

the stimulus being perceived (and thus propagated downstream) (Figure 3.1c), the subsequent strength of stimulus encoding was not influenced by population variance (Figure 3.5a-d). This implies that the neural representation of perceived stimuli was independent of the preceding population variance, and therefore generalised.

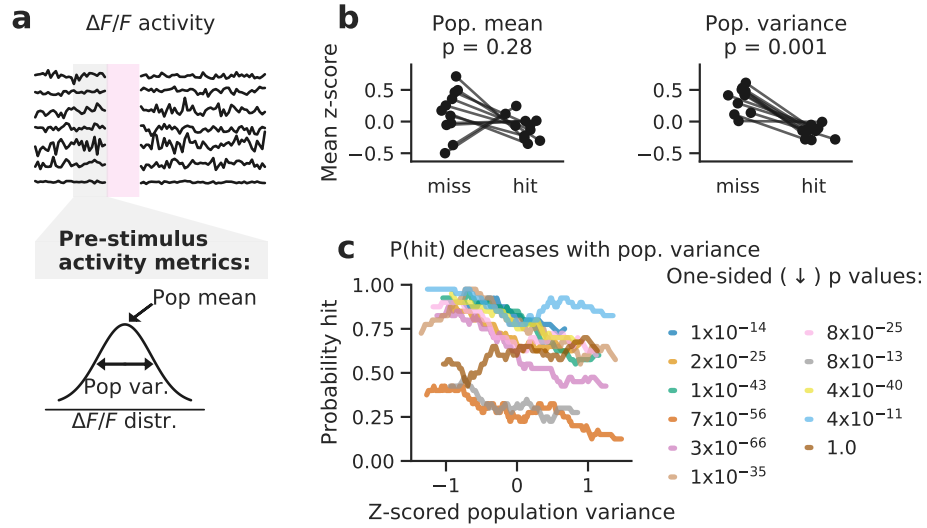


Figure 3.1: Pre-stimulus population in S1 predicts the upcoming trial outcome.
a) Illustration of neural activity throughout a trial. Only the activity in the 0.5 seconds before the stimulus on a given trial is included in subsequent panels. **b)** Comparison of population metrics of pre-stimulus S1 activity prior to hit trials and prior to miss trials. Left: No evidence that mean population activity pre-stimulus predicts the upcoming trial outcome. Right: Population variance is significantly higher before miss trials than before hit trials. Pre-stimulus population metrics in S1 were computed trial-wise and z-scored across trials within a session before being split into hit and miss trials and averaged across a session. P values tested for a difference in session-wise population metrics between hit and miss trials (two-sided Wilcoxon signed-rank test). **c)** The probability of detecting the photostimulation decreased linearly with increasing pre-stimulus population variance in S1 for 10/11 sessions (one-sided t-test, Bonferroni corrected). Trials within a session were binned by their z-scored population variance and this was correlated to the probability of a hit trial within that bin.

3.1.2 Stimulus strength and neural population variance determine probability of perception

Our previous analyses show that we have identified two separate variables that predict the probability a stimulus will be perceived: the strength of the stimulus (number of cells targeted, Figure 2.1c), and the variance of the population activity (Figure 3.1b,c). Next we asked if these two predictors work in tandem, *i.e.*, whether stimulus perception depends on both the strength

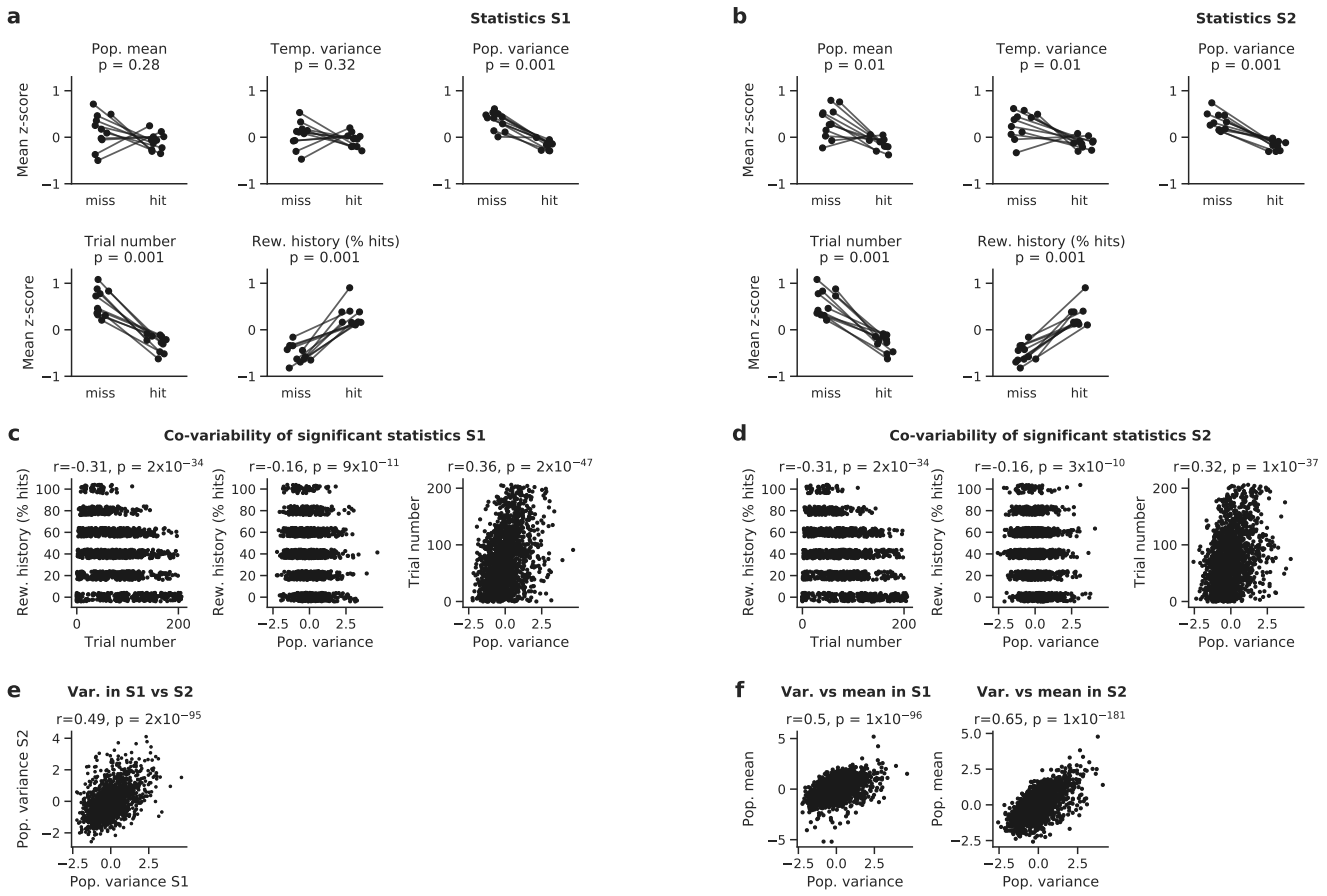


Figure 3.2: Further population metrics of pre-stimulus activity.
a, b) Comparison of population metrics of pre-stimulus in S1 (a) and S2 (b) activity prior to hit trials and prior to miss trials. The first three panels show the two metrics of Figure 3.1b as well as temporal variance. Next, two other significant task variables are shown; the trial number (an integer between 0 and the total number of trials in a session, indicating the number of trials previously undertaken by the animal in a given session) and the reward history (defined as % hit trials in the 5 preceding photostimulated trials). All variables were z-scored for clarity, and p values were calculated using two-sided Wilcoxon signed-rank tests. (No multiple correction was applied.) **c, d)** The co-variability of the significant statistics of panels a and b was assessed using Pearson correlation (two-sided t-test). **e)** Population variance is significantly correlated between S1 and S2 on a single-trial level (two-sided t-test). **f)** Population variance is significantly correlated to the population mean on a trial-by-trial basis in both S1 and S2 (two-sided t-test). Population mean and variance were z-scored in all panels to allow comparison between different recording sessions. Trial number and reward history were z-scored in panels a-b only.

of the stimulus and the ongoing state of the population.

We find that the probability that a stimulus elicits a hit can be expressed on a two-dimensional axis (Figure 3.3a), where hit probability is highest when population variance is minimised and the number of cells targeted is maximised. More targeted cells were required to reliably drive hits, as

pre-stimulus population variance increased. This phenomenon can be conceptualised by a signal-to-noise ratio (SNR) framework, in which activity injected into cells through photostimulation forms the signal and the magnitude of background noise is measured by the population variance. The higher the SNR, the more likely the animal is to detect the signal above ongoing noise and respond, as quantified by collapsing the population variance and the number of cells targeted onto a single SNR axis (Figure 3.3b). We performed a logistic regression analysis to validate that both stimulus strength (*i.e.*, the number of targeted cells) and the noise level (*i.e.*, pre-stimulus population variance) significantly contributed to predicting trial outcome (Figure 3.3). Logistic regression was performed on the data of all hit and miss trials of all sessions (as in Figure 3.3a,b). Both regressors significantly contributed to explaining trial outcome (both p values < 10^{-13}, two-sided t-test on the logistic regression coefficients), with explained variance R^2 of 8.4% (as calculated by McFadden's pseudo-R^2 for logistic regression). This was greater than the R^2 obtained by regressing trial outcome to each of the two variables individually (R^2 = 5.4% for number of targeted cells and R^2 = 2.7% for population variance). Hence, trial outcome was best predicted using both regressors. In summary, our results show that the SNR of a stimulus determines how likely it is to be perceived and propagated downstream to drive behaviour (Figure 3.3c).

Further, as pupil size changes have been associated with spontaneous activity fluctuations (McGinley et al., 2015a), we measured the change in pupil size for 3 sessions during the all-optical experiments (sessions 7, 8 and 9, corresponding to Figure 2.5), see the Supplementary Information from Rowland*, Van der Plas*, Loidolt* et al. (2023) for methodological details (Figure 3.4a). We found that both on average across sessions, and for each session individually, the pre-stimulus pupil size was not significantly different between hit and miss trials (two-sided t tests, Figure 3.4b, d-f). (Although session 7 (panel d) almost reached the significance threshold (p=0.053, no multiple comparison correction)). We then assessed the correlation between pre-stimulus population variance and pre-stimulus pupil size. We found a slightly significant, weak correlation on average (Figure 3.4c, two-sided Pearson correlation r=0.12, p=0.02). Inspection of the individual sessions revealed that this was most likely caused by one session (session 7, panel

g) which was strongly correlated (r=0.31, $p < 10^{-4}$, two-sided Pearson correlation test), while the other two were not significantly correlated (Figure 3.4g-i). Therefore, we conclude that we did not find conclusive evidence that pre-stimulus pupil size consistently influenced trial outcome (Figure 3.4).

Finally, we assessed whether *after perception*, the stimulus encoding strength depended on the strength of the stimulation (*i.e.*, number of cells targeted) and/or the S1 population variance. To this end, we considered the results of the decoders of Figure 2.7 again, but split the visualisation of the results into 3 bands, corresponding to low/medium/high population variance (Figure 3.5a-d) or number of cells targeted (Figure 3.5e-h). First, we observed that the number of cells targeted led to a significantly different decoding performance directly post-stimulus in S1 only (Figure 3.5e-h; On hit vs correct rejection classifiers, for the 18 time points (from post-stimulus up to 1 s): hit 40/50 was significantly different from hit 5/10 in S1 for all time points; miss 40/50 was significantly different from miss 5/10 in S1 for 16 time points; there were no significantly different time points for the two equivalent comparisons in S2; Wilcoxon signed-rank test, Bonferroni corrected). In S1, stronger stimulation led to more positive decoding performance of both hit and miss trials (Figure 3.5e-h). This likely indicates that the photostimulation-induced perturbation of cells in S1 was picked up by the classifiers, but that the ensuing network response (in S2 and later in S1) was equivalent for hit and miss trials irrespective of the strength of stimulation. Second, we found that S1 population variance did not influence the stimulus encoding strength in either S1 or S2, except for a slight difference in S1 for miss trials on the hit vs correct rejection classifier (Figure 3.5a-d; On hit vs correct rejection classifiers, for the 18 time points up to 1 s: miss high-population-variance was significantly different from miss low-population-variance in S1 on 8 time points; there were no significantly different time points for the equivalent tests in S2 and for hit trials in both S1 and S2; Wilcoxon signed-rank test, Bonferroni corrected). From this, we conclude that although both (pre-stimulus) population variance and stimulus strength influence the likelihood that a stimulus will be perceived (Figure 3.3), the ensuing network response to perception (on hit trials) or no perception (miss trials) is generalised and independent of these conditions.

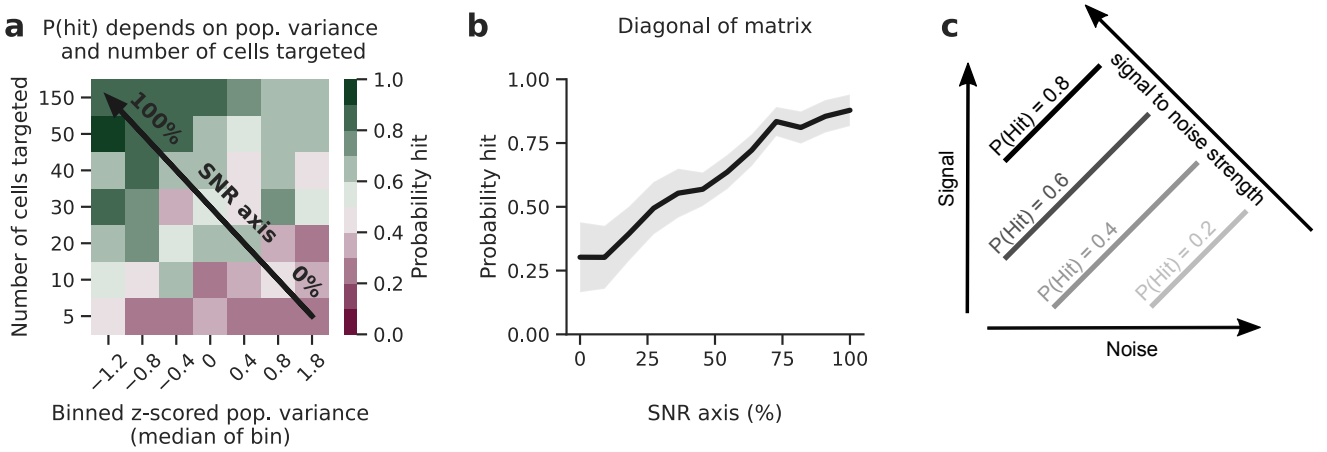


Figure 3.3: Stimulus strength and neural population variance underpin perception.
a) The probability of a hit trial as a function of the interaction between pre-stimulus population variance in S1 and the number of cells targeted by photostimulation. Trials were binned by their z-scored population variance and by the number of cells targeted; the probability of a hit within each bin is plotted on a 2-dimensional axis, pooled across all sessions. Increasing the number of cells targeted (*i.e.*, signal strength) and decreasing the pre-stimulus variance (*i.e.*, 'noise') generally yielded a greater probability of a hit trial. This is referred to as the Signal to Noise Ratio (SNR), as indicated by the diagonal black arrow. **b)** Maximising the SNR of the stimulus resulted in the maximal probability of a hit trial. Data as in a, but projected onto the SNR axis – as indicated in panel a – by averaging across all bins that project orthogonally onto each point on this axis. Data are presented as mean values ± SEM. **c)** Schematic outlining the intuition for the SNR axis. Increasing the number of cells targeted on a given trial maximises the signal of that stimulus. Noise is proportional to the population variance as there is more excitation and suppression from baseline in a population with high variance. The probability of hit is maximal when SNR is maximal, as the stimulus is more likely to be detected above ongoing activity.

Figure 3.4: Pre-stimulus pupil size does not influence trial outcome.
a) Example pupil size detection. This panel was created by Matthias Loidolt (Rowland*, Van der Plas*, Loidolt* et al., 2023). **b)** Trial- and session-averaged pupil size dynamics for hit and miss trials, where shaded areas indicate the 95% confidence interval of the mean (across sessions and trials). Significance was assessed with a two-sided t-test. **c)** Population variance was very weakly correlated to pupil size across all sessions (both averaged across 500 ms pre-stimulus per trial) (two-sided Pearson correlation r=0.12, p=0.02). **d-f)** Trial-averaged pupil size dynamics per session, where shaded areas indicate the 95% confidence interval of the mean (across trials). Significance was assessed with a two-sided t-test. **g-i)** Population variance vs pupil size per session. Two-sided Pearson correlation was significant for only one of the sessions (session 7).

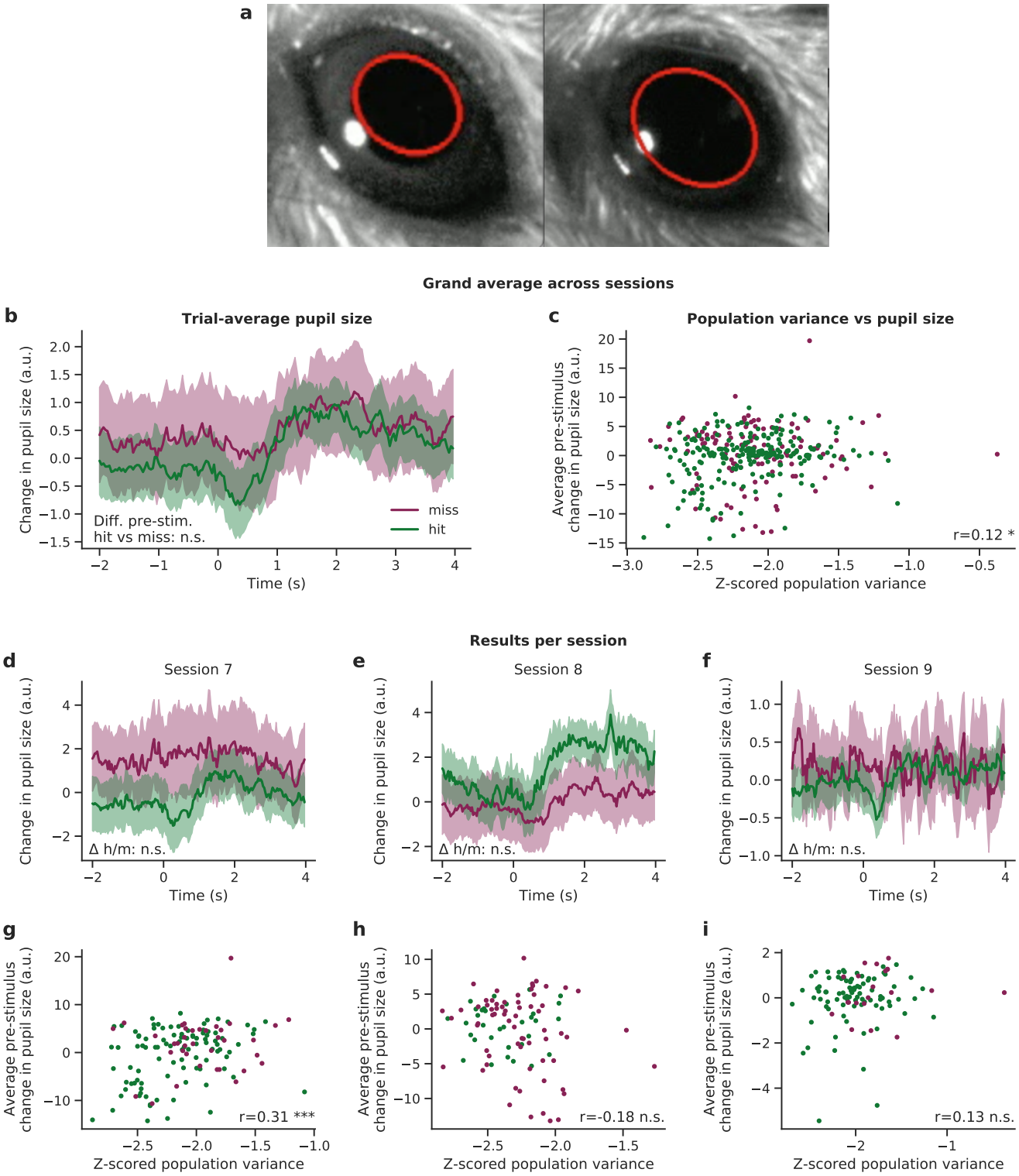


Figure 3.4: (Legend on previous page.)

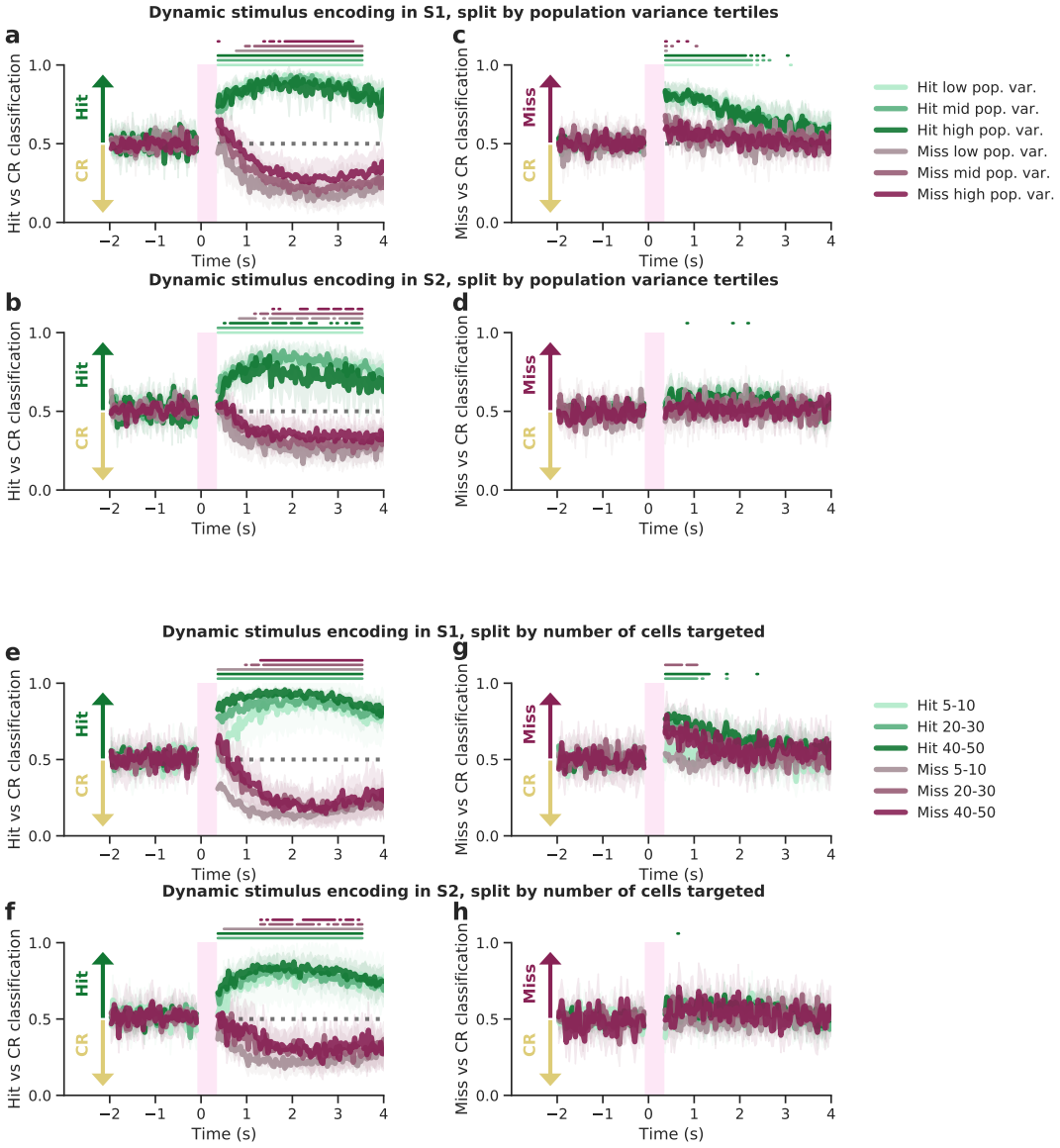


Figure 3.5: Dynamic decoding results split by population variance.
a-d) The decoders of Figure 2.7 are shown, but additionally split by population variance. In other words, this figure shows the same data, but with an additional condition to separate trials by. Each recording session was split into 3 tertiles of equal size based on population variance (*i.e.*, 3 bins were used), which were then averaged across sessions. Classifications are presented as mean values ± SEM across $N = 11$ sessions. **e-h)** The decoders of Figure 2.7 are shown, but additionally split by number of cells targeted. Three 'number of cells targeted' conditions were used (instead of six) to increase data size, in particular for the rare scenarios (such as, *e.g.*, 'Miss 40-50'). Two recorded sessions did not include data for one trial type/cells targeted combination, and were therefore omitted (for that combination only). Hence, classifications are presented as mean values ± SEM across $N = 9$ sessions.

3.1.3 Passive processing of optogenetic stimulation

In the previous sections, we found that perception during a stimulus-detection task is linked to the propagation of activity, which in turn is influenced by the population variance. To understand if these mechanisms are linked to behaviour/reward, or if these are intrinsic to the (S1-S2) cortical circuit, we also analysed a second set of experiments where S1 neurons were passively stimulated – *i.e.*, with no behavioural relevance. In this data set, 30 S1 cells were stimulated on each trial, while the calcium activity in S1 and S2 was recorded. Different S1 stimulation patterns were tested: 1) a random, unspecific set of cells, 2) a sensory, stimulus-selective set of cells (defined by cells that responded to physical whisker stimulation), 3) S2-projecting cells in S1, 4) non-S2-projecting cells in S1. Furthermore, whisker-stimulation trials and sham trials (*i.e.*, no stimulation) were analysed for comparison.

We observed that whisker-stimulation was the only stimulation paradigm that reliably elicited a positive average neural response in both S1 and S2 (Figure 3.6a, b, time points of statistically significant responses are indicated in the figure). However, we did find significant periods of inhibition for the optogenetically stimulated trials (3/4 trial types in S1, 2/4 trial types in S2), which could be a photostimulation-induced inhibitory network response, reminiscent of the previous analysis of neural responses during the stimulus-detection task (Figures 2.3c and 2.6b, c). In these previous stimulus-detection experiments, photostimulation induced a reliable, balanced network response, where although the average activity was close to zero, the stimulation could be decoded from the network activity because the same cells were excited or inhibited across trials (Figure 2.7). We thus asked if this was the case for the passive stimulation experiments too, by evaluating a population-activity decoder on the post-stimulus activity, comparing stimulated trials to sham trials (for each trial type separately). However, we found that only whisker-stimulated trials could be reliably decoded in S2, while all trial types could be decoded in S1 (where the S1 data includes the optogenetically targeted cells) (Figure 3.7). This means that there is no consistent S2 population response to this passive photostimulation, indicating that different cells are activated on different trials.

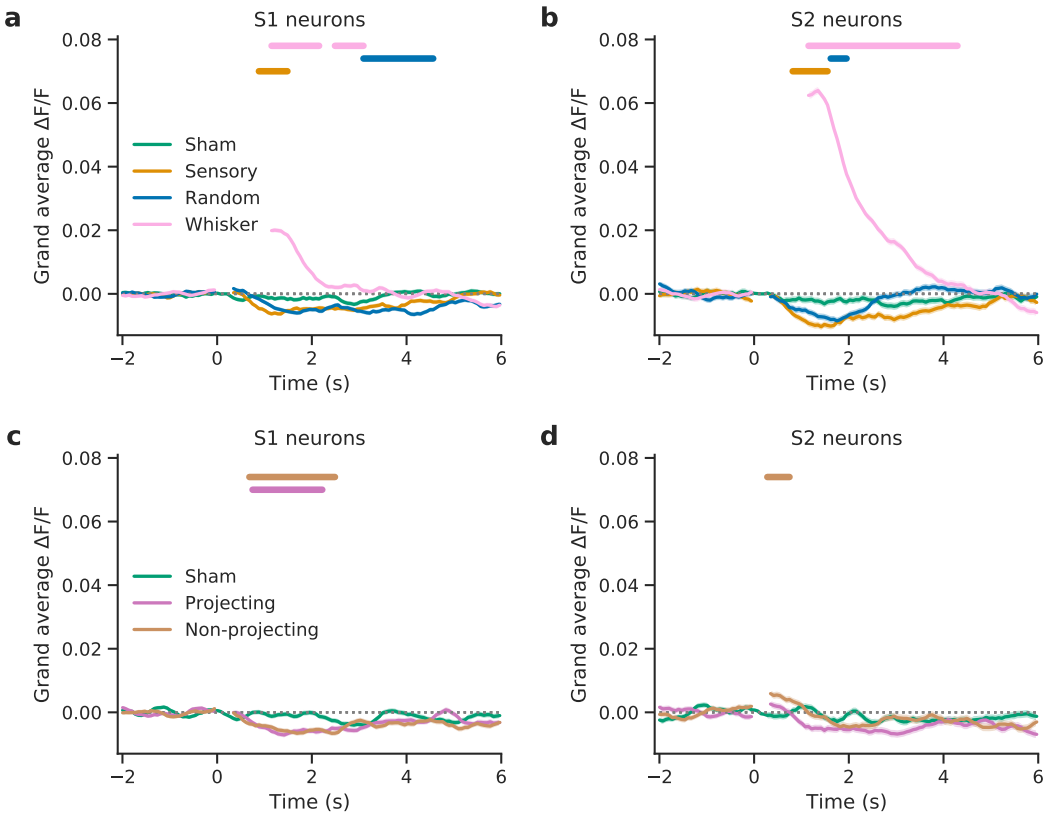


Figure 3.6: Optogenetic stimulation causes a slight inhibitory responses.
Neural activity was averaged across cells, trials and sessions, such that the mean ± SEM is shown per trial type and brain region. For each trial type other than sham, significant deviations from sham are indicated by the horizontal stripes at the top of each panel. Presence of these stripes indicates a cluster of consecutive time points that is significantly different from sham trials (t-test, multiple comparison correction done by suprathreshold cluster size test). Means are temporally smoothed for visual clarity only (*i.e.*, not for assessing significance), with a running window of size 5.

We next assessed whether photostimulation still elicits a reliable S2 response, even if these are not the same cells across trials. We therefore determined the number of responding cells per trial, using the same method as in Figure 2.3d-g, and described in section 2.2.4: for each trial for each cell, 500 ms of pre-stimulus activity was compared to the 500 ms of post-stimulus activity, and tested for statistically significant differences. False discovery rate (FDR) analysis was used to correct for multiple comparisons, and thresholds were chosen such that sham trials had approximately 5% responding cells on average (0.02 for 'projection' experiments and 0.1 for 'sensory' experiments, Figure 3.8). We found that in S1, where the optogenetic stimulation occurred, all trial types had significantly more responding cells than sham trials, as expected

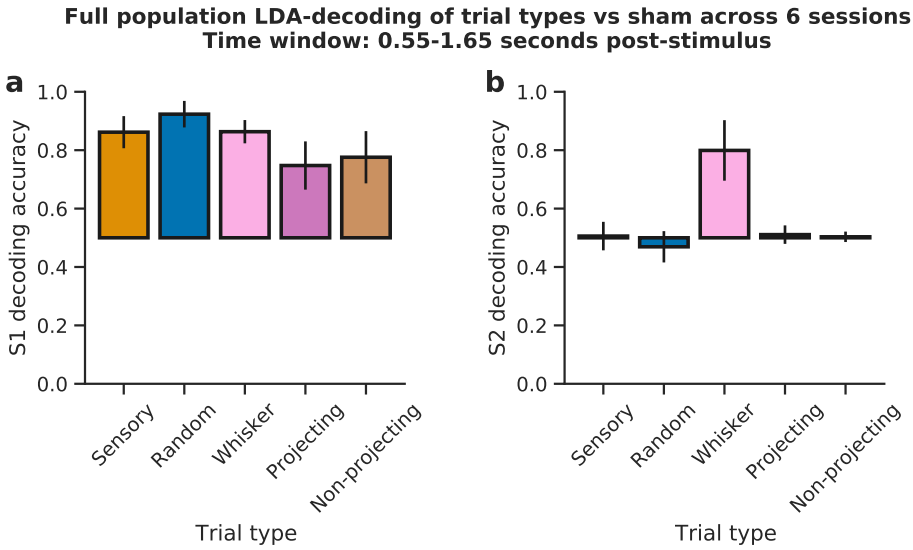


Figure 3.7: All trial types can be decoded in S1, but only whisker-trials in S2.
a) Decoding accuracy (= fraction of correctly decoded trials) per trial type of S1 activity post-stimulus. Decoders were trained and evaluated for each trial type (versus sham) individually, where 0.5 is equivalent to a naive classifier. Error bars denote SEM across $N = 6$ sessions. **b)** Equivalent analysis for S2.

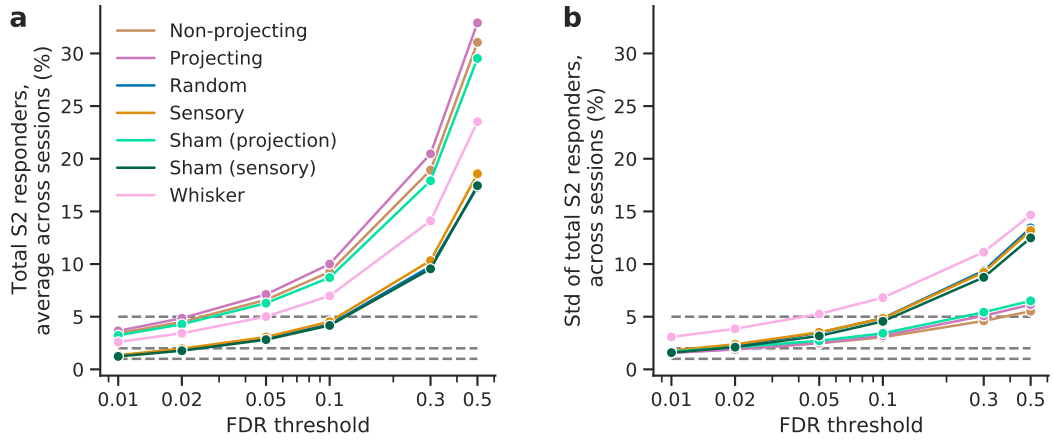


Figure 3.8: Number of responding S2 cells as a function of FDR threshold.
a) Average number of responding S2 cells (across sessions), per FDR (False discovery rate) threshold value. The 'projection' experiments had more responding cells for all FDR values. Grey lines indicate 1%, 2% and 5%. **b)** Standard deviation (std) of data points shown in panel a).

(Figure 3.9a, b). However, only whisker trials and S2-projecting trials contained more responding cells than sham trials in S2 (Figure 3.9c, d).

The results from Figure 3.9 (specifically, the statistically significant differences from sham trials) were equivalent for a range of other FDR-threshold values. In detail, for most FDR-threshold values, we found that all trial types were significantly different from sham in S1, but only whisker and S2-projecting cells were significantly different from sham in S2. For the 'sensory' experiments, this was true for FDR-threshold values of 0.01, 0.02, 0.05 and 0.1 (but not for 0.3

and 0.5, where in S2, sensory trials were significantly different from sham trials with $0.01 < p < 0.05$). For the 'projection' experiments, this was true for FDR-threshold values of 0.01, 0.02 and 0.05 (but not for 0.1, 0.3 and 0.5, where non-S2-projecting trials were also significantly different from sham trials with $0.001 < p < 0.05$). Furthermore, when using a t-test instead of Wilcoxon signed-rank test to define responding cells, significant differences from sham were again equal, except for non-S2-projecting trials which were significantly different from sham with $0.01 < p < 0.05$.

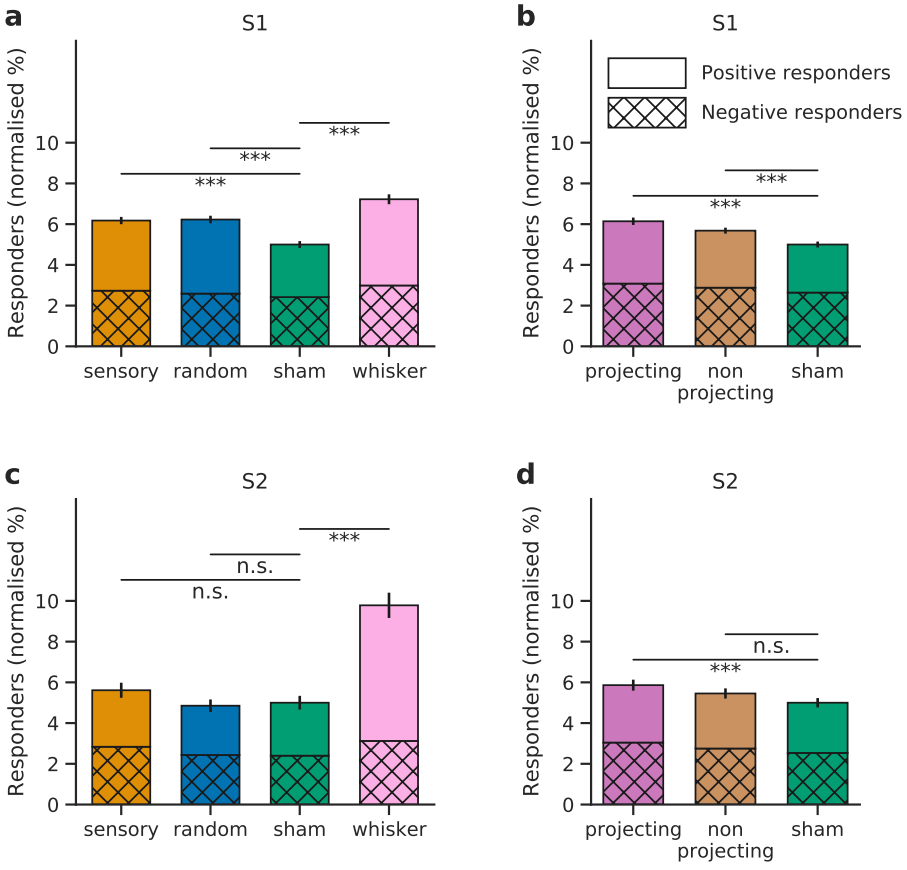


Figure 3.9: Whisker-trials and S2-projecting trials have more responding cells in S2.
The total number of responding cells is shown, averaged across all trials of all sessions, split by positive/negative responders. For this analysis, the number of responding cells was normalised by the trial-averaged number of responding cells on sham trials (such that sham trials had 5% responding cells), per session, before averaging across sessions. Error bars denote the SEM across trials and sessions ($N = 600$ per trial type). Mann Whitney U tests were performed to assess whether the average number of responding cells was significantly different across trial types (Bonferroni corrected for 10 tests). ***: p value < 0.001, n.s.: not significant. Data are split by S1 (**a-b**) and S2 (**c-d**).

Previously, we observed that population variance influenced the likelihood that an optogenetic S1 stimulus was propagated to S2 during a stimulus-detection task (Figures 3.1, 3.3). However, it is unknown whether this mechanism is maintained when the stimulus is not associated with reward. To test this, we asked whether population variance would affect the number of responding cells in the passive stimulation experiments. If the same mechanism would be present, we would expect that lower S1 population variance would yield more S2 responding cells. However, we observed that S1 population variance was positively correlated to the number of S2 responding cells on all trial types, including on sham trials, meaning that lower variability led to fewer responding cells (Figures 3.10, 3.11). Hence we found no evidence of a similar SNR mechanism for passive stimulation. Instead, the positive correlation on sham trials suggests that more variability simply leads to more significant activity changes from pre- to post-stimulus (*i.e.*, how we defined 'responding cells'). In support of this, we observed that population variance-responding cell correlations were even stronger for S1 (sensory: $r = 0.29$, random: $r = 0.23$, whisker: $r = 0.23$, sham: $r = 0.21$, S2-projecting: $r = 0.36$, non-S2-projecting $r = 0.37$, sham: $r = 0.28$, all $p < 0.001$, Bonferroni corrected for 7 tests) than S2 (correlation values in Figures 3.10, 3.11).

A possible explanation for the much-reduced – if present at all – activity propagation to S2 in the passive stimulation experiments, compared to the stimulus-detection task, could be the difference in stimulation paradigm. In the passive experiments, the same set of 30 cells was optogenetically stimulated for all 100 trials of any trial type, while in the active experiments the stimulation varied both in number of cells and cell identities. Hence, the passive experiments could undergo adaptation in the neural response to optogenetic stimulation, for example because of opsin desensitisation or adaptation in the neural response. Figure 3.12 shows the percentage of responding targeted S1 cells as function of trial number, confirming that the number of responding cells generally decreased as the experiment progressed. We therefore assessed whether analysing the first 10 trials (out of 100) only would change the results of Figure 3.9. However, when we quantified the number of S2 responding cells during the first 10 trials of each trial type only, we observed no qualitative difference: again, only whisker trials and S2-projecting



Figure 3.10: Number of responding S2 cells is positively correlated to population variance. Scatter plots of population variance (measured from S1 pre-stimulus activity, log-transformed and z-transformed) versus total number of S2 responding cells (normalised by average number of S2 responding cells on sham trials). Each panel shows all trials of all sessions, for sensory (**a**), random (**b**), whisker (**c**) and sham (**d**) trial types. The Pearson correlation value *r* and p value (Bonferroni corrected for 7 tests) is indicated at the top. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

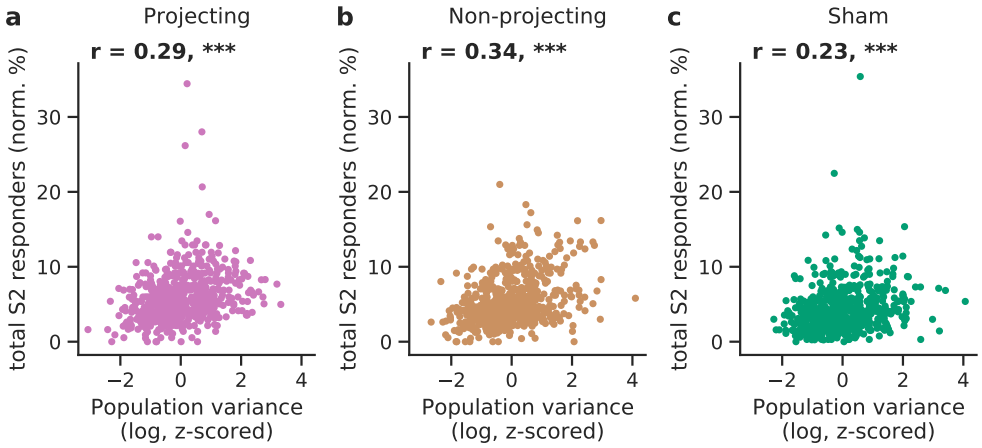


Figure 3.11: Number of responding S2 cells is positively correlated to population variance, continued. Same as Figure 3.10, but for the 'projection' experiments: S2-projecting (**a**), non-S2-projecting (**b**) and sham (**c**) trial types.

trials were significantly different from sham trials in S2 (Figure 3.13).

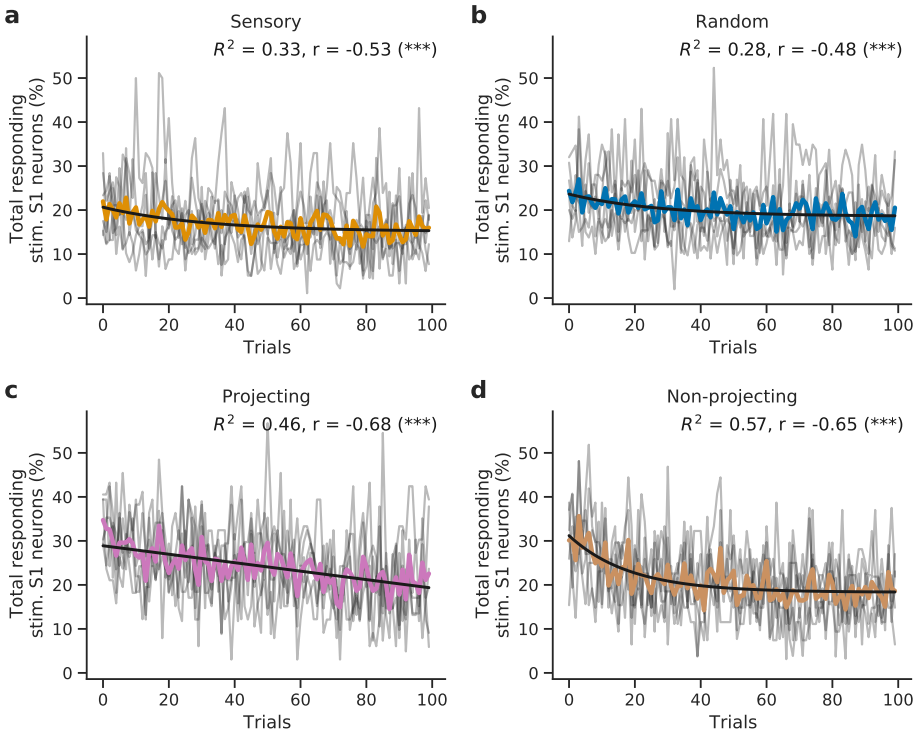


Figure 3.12: The number of responding target cells decreases with experiment duration.
The number of responding optogenetically targeted cells in S1 is shown per trial (in the order that the trials occurred), split by trial type: sensory (**a**), random (**b**), S2-projecting (**c**) and non-S2-projecting (**d**). Note that targeted cells include both the 30 directly targeted cells and all cells within 15 μm of target cells. Grey lines show individual sessions ($N = 6$ per panel/trial type), thick coloured lines the mean across sessions. The means were fitted with an exponential decay function and linear function, and the fit with the highest R^2 is shown in black (exponential for panels a, b and d; linear for c, where exponential fit had an equal $R^2 = 0.46$). R^2 values, Pearson correlations r and one-sided p values (testing for negative correlations, Bonferroni corrected for 4 tests) are denoted at the top of each panel. ***: $p < 0.001$.

3.2 Methods

3.2.1 Pre-stimulus population metrics

All pre-stimulus population metrics were computed across a 500 ms period immediately prior to photostimulation. All metrics were calculated on a trial-wise basis. The natural logarithm was taken of metrics that were fit better by a log-normal distribution as opposed to a normal distribution as assessed by Kullback-Leibler divergence (for clarity, this includes population variance).

Mean population activity (Figures 3.1b – left, 3.2a, b – left) was computed by first averaging

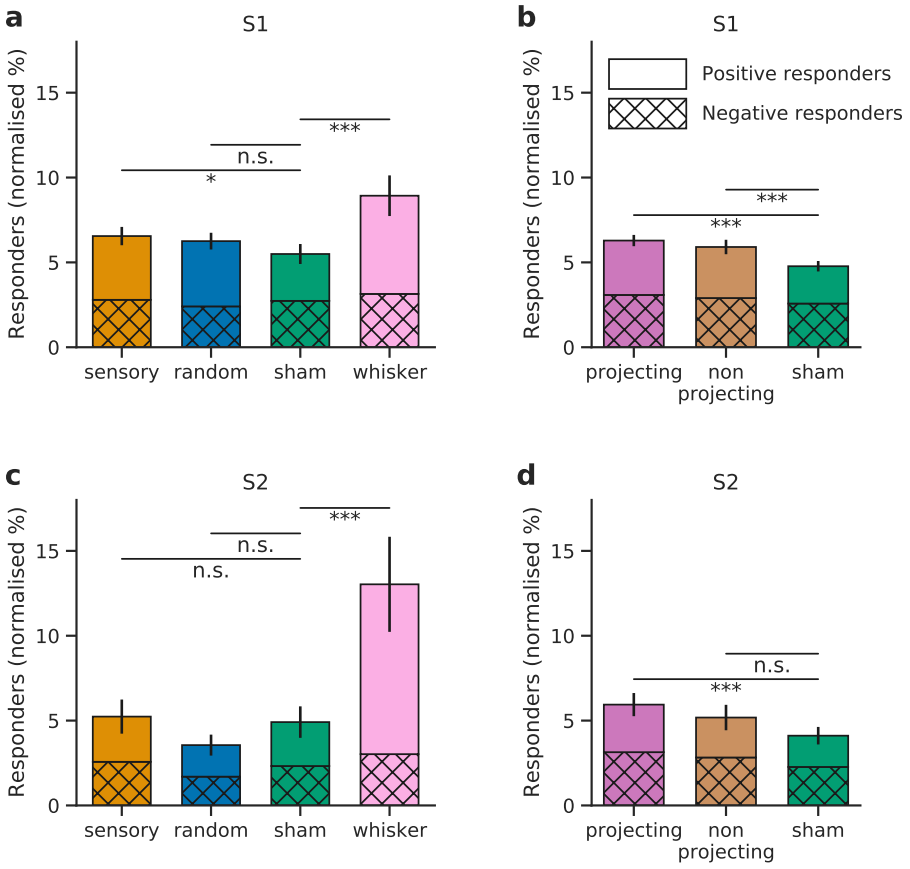


Figure 3.13: The first 10 trials are representative of all 100 trials.
As in Figure 3.9, but only the first 10 trials (out of 100) of each trial type are analysed. Otherwise all analysis and parameters were exactly replicated.

$\Delta F/F$ activity across all pre-stimulus frames for each neuron to yield a vector containing a scalar value for each neuron defining its pre-stimulus activity. Next firing rates were averaged across all neurons to give a single scalar value for each trial, defining the average population activity.

Population variance (Figures 3.1b – right, c, 3.2a, b – right, 3.3, 3.5a-d) was computed in a similar fashion, first by averaging across all pre-stimulus frames (in the 500 ms window prior to stimulation) for each neuron. However rather than taking the mean of the activity vector as above, the variance of the vector was used to generate a single scalar value for each trial.

Temporal variance (Figure 3.2a, b – middle) was calculated by first computing the variance across time (in the pre-stimulus window), for each neuron, and was subsequently averaged across neurons.

3.2.2 Experimental details of section 3.1.3

Here we will describe in brief the experiments of section 3.1.3. This thesis primarily concerns the data analysis of these experiments, rather than the experimental procedure itself (performed by Robert Lees), so the descriptions will be concise. All animal experimentation was carried out with approval from the UK Home Office and University of Oxford Animal Welfare and Ethical Review Board.

As before, calcium activity of both S1 and S2 was imaged using two-photon imaging (using viral or transgenic GCaMP6s (Chen et al., 2013)), while S1 excitatory cells were stimulated using two-photon optogenetics (using C1V1-Kv2.1 (Yizhar et al., 2011; Chettih and Harvey, 2019)). Imaging was done at 15 Hz (as opposed to 30 Hz in the previous experiments), and was pre-processed using Suite2p (Pachitariu et al., 2017), and $\Delta F/F$ was calculated as before. Two sets of experiments were performed: 'sensory' experiments ($N = 6$) that contained sensory (stimulation of whisker-responsive S1 cells), random (stimulation of unspecific S1 cells), whisker (physical whisker stimulation) and sham (no stimulation) trials, and 'projection' experiments ($N = 6$) that contained S2-projecting (stimulation of S1 cells that project directly to S2), non-S2-projecting (stimulation of S1 cells that do not project directly to S2) and sham (no stimulation) trials.

All recording sessions consisted of 100 trials per trial type, carried out over 25 minutes, with 15 second inter-trial intervals. The time points during which optogenetic or whisker stimulation occurred were blanked out for analysis. On all optogenetic stimulation trials, 30 cells were stimulated, but all cells within 15 μm of the photostimulation beam were defined to be 'targets'. On whisker trials, a single whisker (B2 or B3) was threaded in a capillary and deflected at 10 Hz. On sham trials, the same experimental procedure was used as for stimulated trials, except for the laser power, which was set to zero. For the 'projection' experiments, whether S1 cells projected to S2 was determined using retrograde labelling of S2.

3.2.3 Suprathreshold cluster size test

We performed statistical tests to assess whether average $\Delta F/F$ responses on stimulated trials were significantly different from sham trials (Figure3.6). At each time point, a t-test was performed (across cells and trials) to assess significance ($p < 0.05$). Hence, multiple comparison correction was required, as 15 tests per second are performed per trial type (*i.e.*, just under 90 tests for 6 seconds of post-stimulus activity excluding the stimulation period). Previously, we used Bonferroni correction (Figures 2.7, 2.8), which is a very stringent, or conservative, correction because it assumes that all tests are independent, which they are not because of the auto-correlation of calcium activity (Chen et al., 2013). As the deviations in Figure3.6 are of much smaller magnitude than before, we corrected for multiple comparison by permutation tests instead.

Specifically, we implemented the suprathreshold cluster size test (Nichols and Holmes, 2002). This tests whether the size of clusters – defined as consecutively significant time points – is significantly larger than expected by chance. First, we computed at which time points stimulated trial types significantly differed from sham trials (t-test, $p < 0.05$). Then, to calculate the null distribution of cluster sizes expected by chance, this time series (of significant/non-significant binary data points) was shuffled and the largest cluster identified. This was repeated 5,000 times, such that we obtained a distribution of maximum cluster sizes, expected by chance for the number of significant time points in the data. Then, the threshold was set at $p < 0.05$, or the 95^{th}percentile of maximum cluster sizes. Any cluster found in the data larger or equal to this threshold was said to be significant (plotted by horizontal stripes in Figure3.6). This analysis was performed separately for each trial type.

3.2.4 Population decoder

We used a Linear Discriminant Analysis (LDA) decoder for the analysis of Figure3.7, using the scikit-learn implementation (Pedregosa et al., 2011). The same post-stimulus windows as for the responsive cell analysis were used: 0.55-1.65 seconds post-stimulus for all trial types except for whisker, which used 1.1-2.2 seconds post-stimulus because more stimulation frames

were blanked out. Neural data was averaged across these time points, such that each data point was a 1D vector of time-averaged neural responses of one trial. We used stratified 5-fold cross-validation to create train/test splits, and the average accuracy (= fraction of correctly classified test trials) across folds was used to evaluate performance.

Only S2 neurons were analysed, and for the purpose of this analysis only, post-stimulus activity was baselined with respect to the pre-stimulus activity (to correct for baseline changes/drifts across trials). Baselining was done by subtracting the time-averaged $\Delta F/F$ per cell per trial of the 2 seconds preceding stimulation from its post-stimulus activity.

We verified that changing the decoder to logistic regression, as well as using a shorter time window of 1.1-1.6 seconds post-stimulus, did not qualitatively change the results of Figure3.7.

3.2.5 Data and code availability

Section 3.1.1 and 3.1.2: All source data used for the data analysis and visualisation are publicly available on a GIN data repository. This includes the two-photon calcium imaging and optogenetics recordings of all experiments, as well as the pupil size data.

All data analysis and visualisation presented was performed in Python 3.7, which we have made publicly available on Github. In particular, all figures were made fully in Python, and the code is available on Github.

Section 3.1.3: All source data used for the data analysis and visualisation are publicly available on a GIN data repository.

All data analysis and visualisation presented was performed in Python 3.7, which we have made publicly available on Github. In particular, all figures were made fully in Python, and the code is available on Github.

4
Predictive Learning

We use artificial neural networks to show that predictive learning can give rise to an internal stimulus representation that enables networks to learn a more difficult working memory task. Specifically, in our experiments, small, sparse RNNs could not learn a delayed-match-to-category (DMC) task without simultaneously engaging in predictive learning. When RNNs were faced with these two tasks, they would first solve the prediction task, which created an internal stimulus representation that kick-started learning of the DMC task. This phenomenon – where two tasks depend on a latent feature that can only be learned from one of the tasks, thereby initiating learning of the other – has been termed 'eavesdropping' (Caruana, 1997), and constitutes a form of multitask learning that is known to boost learning performance in artificial networks (Bell and Renals, 2015; Yu and Jiang, 2016; Liu et al., 2020). Our results suggest that neural systems could engage in eavesdropping by using predictive learning as helper task.

We also show that predictability does not need to persist constantly, in order to produce long-term benefits. It was sufficient to introduce stimulus predictability only transiently, making simulated annealing an effective mechanism to kick-start the learning of complex

Corresponding publication: Van der Plas, T. L., Manohar†, S. G., and Vogels†, T. P. (2022). Predictive learning enables neural networks to learn complex working memory tasks. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 518–531. †: co-senior author.

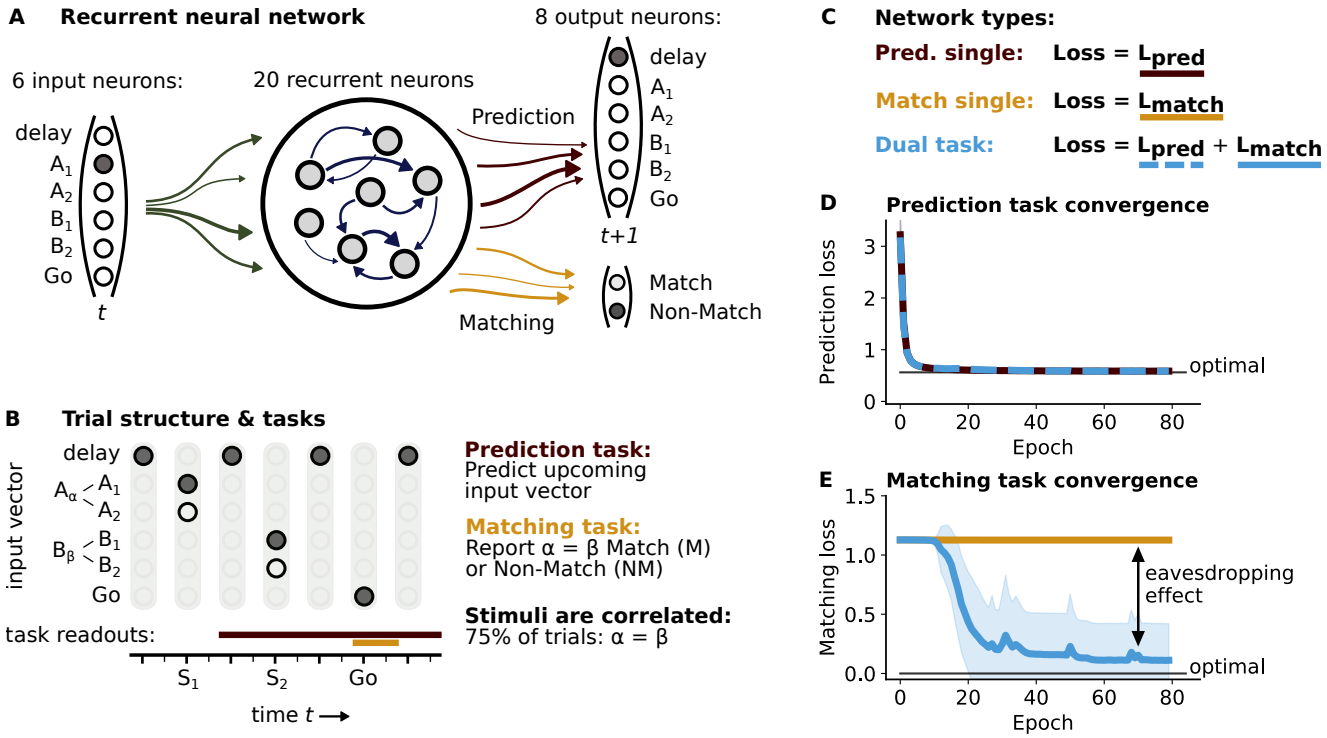


Figure 4.1: RNNs use eavesdropping to bootstrap learning.
A) RNNs projected to two sets of output neurons, corresponding to the prediction task and matching task. **B)** Each trial consisted of two stimuli S_1 and S_2 followed by a GO response window, interleaved by delay periods. The stimuli take the possible values $S_1 = A_\alpha \in \{A_1, A_2\}$ and $S_2 = B_\beta \in \{B_1, B_2\}$, where $P(\alpha = 1) = P(\alpha = 2) = \frac{1}{2}$ and $P(\alpha = \beta) = \frac{3}{4}$ (hence S_1 is predictive of S_2). **C)** Definitions of loss functions. **D)** Convergence of loss functions on prediction task. **E)** Convergence of loss functions on matching task. Panels D and E show the mean $\pm$ std of 20 simulated networks per network type.

concepts. This finding aligns with the dynamic nature of natural environments, where stimulus correlations come and go. Finally, we leverage our results to formulate testable predictions for experimental neuroscience.

4.1 Results

4.1.1 Recurrent neural networks use eavesdropping to bootstrap learning

We trained standard recurrent neural networks (RNNs) to solve two working-memory tasks, either individually or simultaneously (Figure 4.1A). All RNN weights were randomly initialised, and then trained with backpropagation through time (Algorithm 1). See Methods, section 4.2 for a detailed description of the RNN, synthetic data and training procedure. Trials consisted

of a sequence of two stimuli and a GO cue, interleaved by delay periods. The prediction task asked networks to continuously predict the next upcoming input stimuli. The matching task prompted during the GO period whether the two previously presented stimuli matched (Figure 4.1B). Importantly, the stimuli were predictable, or biased, because a match occurred on 75% of trials. This means that the first stimulus S_1 was predictive of the second stimulus S_2 on most trials.

Networks were trained with one of three possible loss functions: prediction single task learning (red, Figure 4.1C), matching single task learning (orange) or dual task learning (blue). The dual task loss function was defined as the sum of the prediction task and matching task loss functions. Notably, we found that networks could only learn the matching task when it was presented jointly with the prediction task (Figure 4.1D, E). Here, networks first solved the prediction task, after which learning of the matching task ensued. Both prediction single task networks and dual task networks learned the prediction task equivalently well, converging to the optimal solution equally quickly. But only dual task networks could learn the matching task, after they solved the prediction task (Figure 4.1D, E). This sequence of task learning was not specified, as the loss function was set to be the sum of the loss functions of both tasks (Figure 4.1C).

As we will detail in the next sections, this happened because the prediction task unveiled a latent feature – the memory of the first stimulus S_1 – which was required to solve the matching task. This phenomenon, in which two tasks share a feature that can (easily) be learned from one, thus catalysing learning in the second, is a multitask learning mechanism called 'eavesdropping' (Caruana, 1997).

4.1.2 Quantification of the eavesdropping effect

To quantify the eavesdropping effect (Figure 4.1E) and explore what conditions facilitated this learning strategy, we first optimised the number of RNN hidden units based on their performance on the prediction task alone (Figure 4.2A). We found that networks with 20 neurons converged on the prediction task with near-maximal speed of convergence. Here, convergence speed was

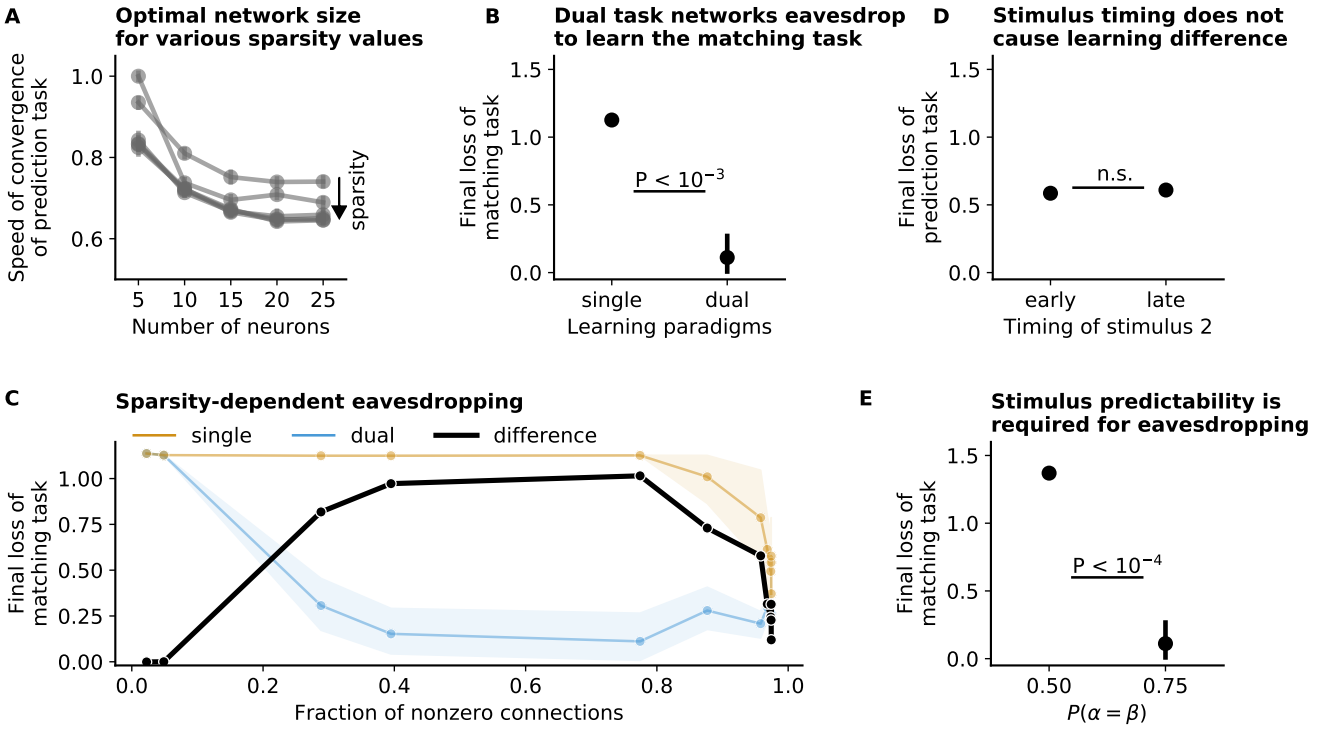


Figure 4.2: Quantification of the eavesdropping effect.
A) Speed of convergence of the prediction task as a function of network size and sparsity. **B)** The eavesdropping effect shown in Figure 4.1E was quantified by measuring the difference in final loss. **C)** The magnitude of the eavesdropping effect depended on the number of nonzero connections (defined by $|W| > 0.01$, result is robust to different thresholds). **D)** The prediction task was solved equally well when S_2 was delivered later, during the GO window. **E)** Eavesdropping only occurred when the stimuli were predictable ($P(\alpha = \beta) = 0.75$), as in prior panels). Panels B, D and E used a two-sided Wilcoxon signed-rank test to assess significance (P values shown). All panels show the mean $\pm$ 95% confidence interval of 10 simulated networks per data point for panel A and 20 simulated networks per data point of all other panels (error bars or shaded area).

defined as the integral of the loss function over epochs, see Figure 4.1D), with an optimum reached at $N = 20$ neurons for the majority of sparsity values. (Based on the final loss metric, the optima would be reached at $N = 10$ to $N = 15$ for all sparsity values.) We quantified the eavesdropping effect by calculating the difference in performance (*i.e.*, their final loss value after training, averaged over 5 final epochs) between matching single task networks (Figure 4.1C, E, in orange) and dual task networks (Figure 4.1C, E, in blue), averaged over 20 simulations of each, confirming that only dual task networks could learn the matching task ($p < 10^{-3}$, two-sided Wilcoxon signed-rank test, Figure 4.2B).

Next, we quantified how the eavesdropping effect depended on the difficulty of the tasks, by varying the strength of sparsity regularisation (Figure 4.2C). We found that eavesdropping was most effective at a range of medium sparsity values from 30% to 80% non-zero connections. The number of nonzero connections (defined by $|W| > 0.01$) was varied by parametrically changing the sparsity regularisation parameter λ of Equation 4.7 from 0 to 0.1 and measuring the fraction of nonzero connections. When sparsity regularisation was too strong (*i.e.*, less than 25% of all possible connections were allowed), both learning paradigms failed to learn. When sparsity regularisation was very weak (i.e., more than 80% of all connections were allowed) both paradigms partially converged. Hence, the effectiveness of eavesdropping depended on the difficulty of the tasks, and eavesdropping allowed sparse networks to learn the matching task which is typically beneficial to prevent overfitting. Correspondingly, the best absolute performance on the matching task was only achieved with sparse networks that used eavesdropping (Figure 4.2C). We performed this analysis across a range of network sizes N (i.e., number of hidden units) from 5 to 100, which resulted in an equivalent range of sparsity values where networks could only learn the matching task with eavesdropping (Figure 4.3). Specifically, very small networks ($N = 5$) could not solve the task, while larger networks generally require eavesdropping for optimal performance, especially when sparsity is constrained. For the remainder of this study, we used the optimal fraction of nonzero connections (77%) from Figure 4.2C, which corresponds to $\lambda = 10^{-3}$ in Equation 4.7.

Networks had to create an internal representation, *i.e.*, a memory, of S_1 to predict the upcoming stimulus S_2. For the matching task, networks also had to memorise S_1 (to compare the two stimuli), but for a longer period of time (to compare with S_2). Therefore, the difference in performance (Figure 4.2B) could in theory be explained by vanishing gradients. However, we found that this is not the case, as the prediction task was equally well solved when the delay period between S_1 and S_2 was increased to mirror the matching task (not significant, two-sided Wilcoxon signed-rank test, Figure 4.2D).

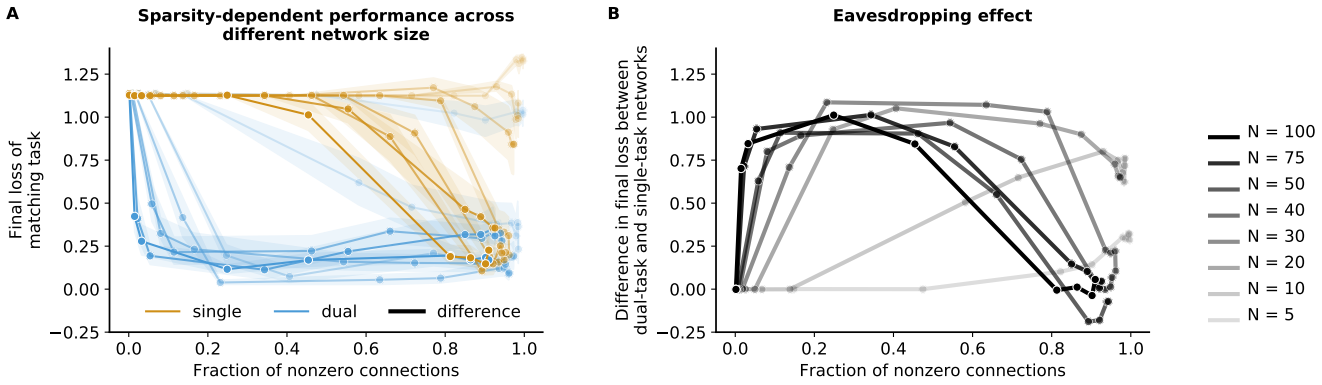


Figure 4.3: Eavesdropping effect dependency on network size.
The eavesdropping effect as in Figure 4.2C was quantified for a range of network sizes N. The single task and dual task match losses are shown in panel **A**, and their differences in panel **B**. Network size is indicated by transparency (see legend on the right). The average $\pm$ 95% confidence interval of 50 simulations per data point is shown.

Finally, as expected, we found that stimulus predictability $P(\alpha = \beta) > 0.5$ was crucial for eavesdropping (on the prediction task) to occur: dual task networks failed to learn the matching task when they were trained with uncorrelated stimuli ($p < 10^{-4}$, two-sided Wilcoxon signed-rank test, Figure 4.2E), because S_1 is not predictive of S_2 if stimuli are not predictable ($P(\alpha = \beta) = 0.5$).

We have shown that RNNs solved the matching task by eavesdropping on the solution of the prediction task. We have purposefully used a small and simple network, and have optimised the parameters that give rise to eavesdropping (Figure 4.2). This now enables us to pinpoint why learning the prediction task catalysed learning of the matching task.

4.1.3 Matching task eavesdrops on the stimulus memory of the prediction task

To understand the effect of eavesdropping, we considered the neural representations of both the S_1 stimulus and the Match/Non-Match trial type at various stages during learning. Figure 4.4A,B shows these for an example network at 4 different epochs during training. Here, we observe that dual task networks gradually enhanced their neural representation of the S_1 stimulus over the course of training (Figure 4.4A). First, the prediction task induced a memory of S_1 until

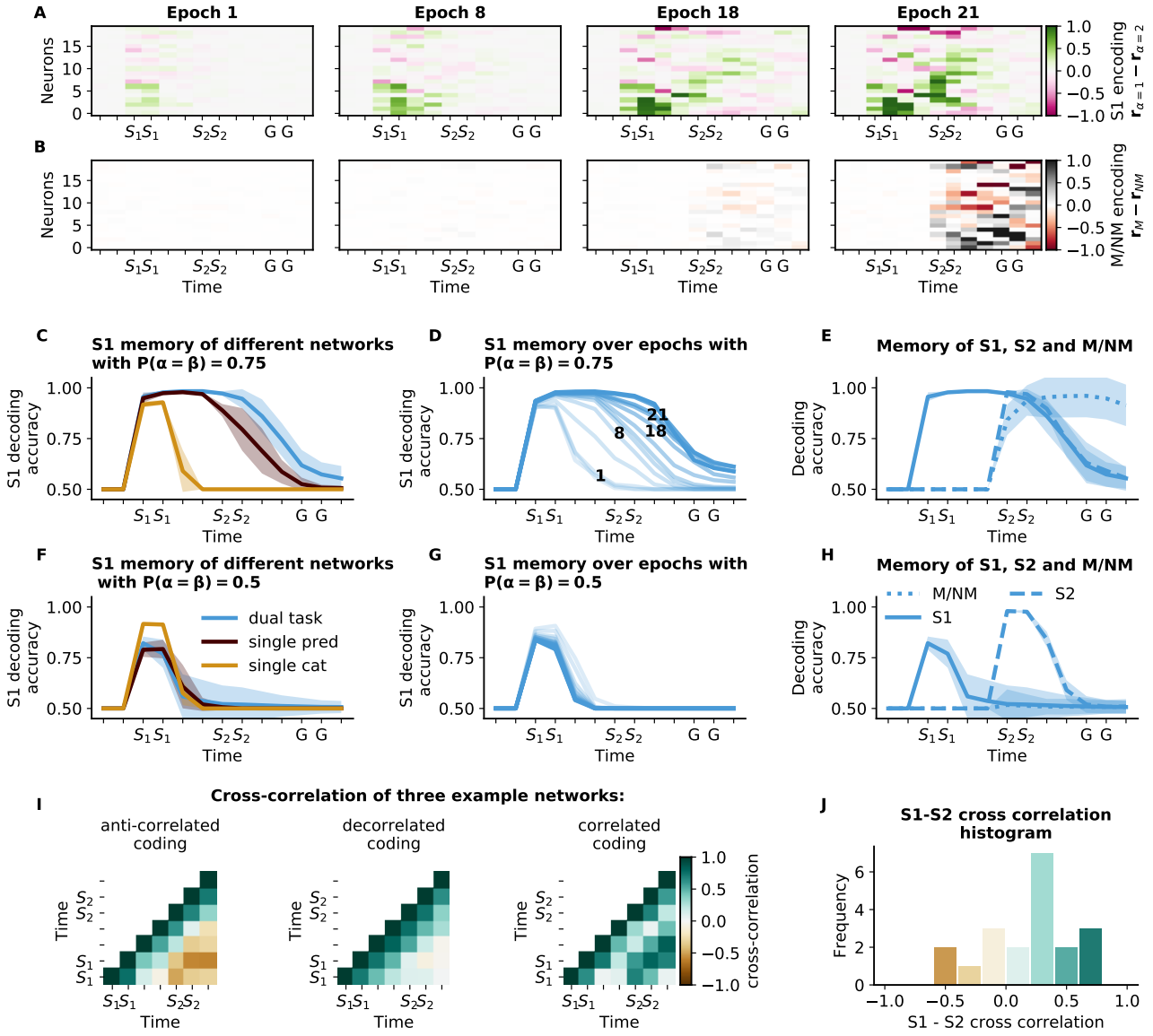


Figure 4.4: Matching task eavesdrops on the stimulus memory of the prediction task.
A) The neural representation of S_1 is shown. Each raster plots shows the differentiated activity between $\alpha = 1$ and $\alpha = 2$ trials of all neurons (clipped between -1 and 1 for clarity). **B)** Similarly, the neural representation of the Match/Non-Match (M/NM) trial type is shown at the same 4 epochs. **C)** The memory strength of S_1 was determined per time point by decoding the stimulus identity from the network activity. Panels C-H show the average $\pm$ std across 20 simulations. **D)** Dual task networks gradually learned to enhance the memory of S_1 (different epochs are shown from light to dark blue). The 4 epochs of panels A and B are indicated in the figure. **E)** Dual task networks memorise S_1 until the S_2 stimulus is presented, after which they represent the Match/Non-Match identity of the trial. **F, G, H)** Equivalent figures for networks that were trained with uncorrelated stimuli. **I)** The cross-correlation of the differentiated activity (as in panel A) is shown for 3 example networks that show anti-correlated, decorrelated and correlated coding respectively (quantified by the S_1-S_2 cross-correlation). **J)** The distribution of S_1-S_2 cross-correlation values of the 20 network simulations (as determined by the average of the 4 cross-temporal data points with S_1 vs S_2 from panel I).

the S_2 presentation (epoch 8, Figure 4.4A), which is subsequently strengthened to facilitate the matching computation (Figure 4.4B). To quantify this learning dynamic, and study this effect across all networks, we then defined the memory strength of stimulus S_1 at time t as the decoding accuracy of S_1 of the differentiated neural activity at time t. We then evaluated this quantity across trials using logistic regression (see Methods, section 4.2). Figure 4.4C shows how the memory of S_1 dynamically evolved for all three network types after training. Matching single task networks (orange) failed to converge, as they did not create a memory of S_1 beyond the time when S_1 was presented to the networks. However, prediction-only networks (red) created a memory representation of S_1 and maintained this at perfect accuracy until the S_2 presentation, as expected (because the optimal solution would use S_1 information to predict S_2). Dual task networks (blue) also maintained a perfect memory of S_1 after convergence that slightly extends the duration of prediction-only networks, so that S_1 and S_2 could be compared to perform the matching task.

The prediction task and matching task thus share the same latent feature: S_1 memory. The prediction task was solved by creating a memory of S_1, on which dual-task networks could subsequently eavesdrop to solve the matching task (Figure 4.4D). Conversely, prediction single task networks that are trained on uncorrelated data ($P(\alpha = \beta) = 0.5$) do not require a memory of S_1 to achieve optimal performance, because S_1 is no longer predictive of S_2. Indeed we found that no S_1 memory was created in this case (Figure 4.4F, red), which subsequently failed to bootstrap the matching task for the dual task learning networks (Figure 4.4G, H). Hence, only networks trained on predictable stimuli created a memory of S_1 when solving the prediction task, and this facilitated learning of the matching task.

We next investigated how the neural representation of S_1 after learning (as shown in Figure 4.4A) dynamically evolved during a trial, which is known to be variable network property (Orhan and Ma, 2019). We quantified this by computing the cross-correlation (across time) of the differentiated neural activity (as in Figure 4.4A). Intriguingly, we found that dual task networks could embed several different solutions to achieve optimal performance, depending on the

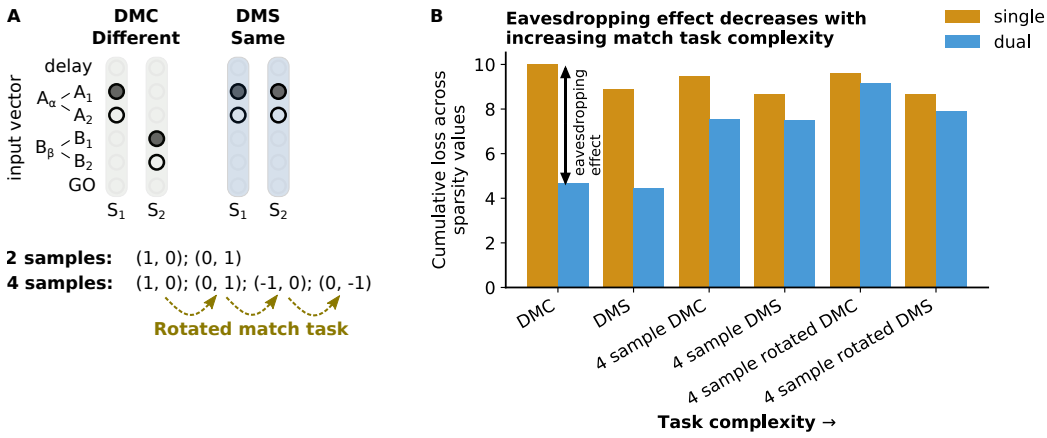


Figure 4.5: Increasing matching task complexity reduces eavesdropping effect.
A) We investigated the benefit of eavesdropping for 6 different tasks (including the task of previous figures). **B)** Eavesdropping effect quantified for the 6 different matching tasks, by summing the sparsity-dependent loss of the matching task of Figure 4.6 (*i.e.*, the orange and blue lines, experiment was repeated for all tasks, with 20 networks per sparsity value).

stability of the neural representation of S_1. When the subpopulation of neurons that encoded S_1 changed over the course of a trial, this gave rise to a cross-temporal anti-correlation, while an unchanged representation induced a cross-temporally correlated code (Figure 4.4I). Finally, a mix of these neuron types decorrelated the memory code. In other words, the cross-temporal correlation of the S_1 representation depended on how variable the identity of its neurons was over time. We quantified this variability across different network simulations, which differed only in initial conditions, illustrating the multitude of solutions that networks could use (Figure 4.4J). This highlights that eavesdropping was not specific to a particular neural representation of the S_1 stimulus, as each of these solutions maintained the S_1 memory equally well.

4.1.4 Increasing matching task complexity reduces eavesdropping effect

Eavesdropping allowed RNNs to learn the matching task by utilising memories from solving the prediction task. Consequently, eavesdropping failed when stimuli were not predictable (Figure 4.4F-H), *i.e.*, when S_1 was not predictive of S_2 such that the prediction task was no longer sufficiently informative of the matching task. Next we asked if we could similarly identify the limits of when eavesdropping is an efficient learning strategy by increasing the difficulty of the matching task instead.

We changed the matching task by considering 6 different combinations of three possible conditions: input domain, number of samples and matching task type (illustrated in Figure 4.5A). Each task is a combination of the following 3 conditions (inspired by Yang et al. (2019); Masse et al. (2019); Duncker et al. (2020)):

- **Input domain**: stimuli S_1 and S_2 either used the same input domain (A_α and A_β, i.e., delayed-match-to-sample (DMS) task) or different domains (A_α and B_β, i.e., delayed-match-to-category (DMC) task).
- **Number of samples**: each stimulus could either be chosen from 2 samples ($A, B \in \{(1,0),(0,1)\}$) or 4 samples ($A, B \in \{(\cos\theta, \sin\theta); \theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}\}$). Because these stimuli can also take negative values, a different output non-linearity (tanh instead of softmax) and loss function (mean least squares instead of cross-entropy) was used for the prediction task. (We verified that the 2-sample prediction task also converged when optimised with this output non-linearity and loss function.)
- **Type of matching task**: For the 4-sample condition, we considered a regular matching task ($A_\alpha = B_\beta$ in the case of different input domains), and a rotated matching task ($A_\alpha = B_{\beta+\frac{\pi}{2}}$).

Hence, we analysed performance on the following six tasks: 1) 2-sample DMC, 2) 2-sample DMS, 3) 4-sample DMC, 4) 4-sample DMS, 5) 4-sample rotated DMC, 6) 4-sample rotated DMS.

We then quantified the eavesdropping effect across sparsity values (as before in Figure 4.2C), detailed in Figure 4.6. To obtain a single metric, we summed over the sparsity-dependent loss of the matching task of Figure 4.6 (*i.e.*, the orange and blue lines, experiment was repeated for all tasks, with 20 networks per sparsity value). In short, we found that as the matching task complexity increased, eavesdropping gradually became less effective because the dual-task networks could no longer solve the matching tasks. This decrease in eavesdropping efficiency was not caused by a failure of the networks to solve the prediction task, as networks still solved the prediction task at all task complexities and sparsity values (except for the two strongest sparsity values that led to $\sim 0\%$ nonzero weights), see Figure 4.7. Task performance was similar

for both input domains (match-to-sample or match-to-category tasks), but decreased when more stimuli or a rotated-match task were introduced (Figure 4.5B). Hence, as the difficulty of the matching task increases, eavesdropping on the solution of the prediction task becomes less useful for learning the matching task.

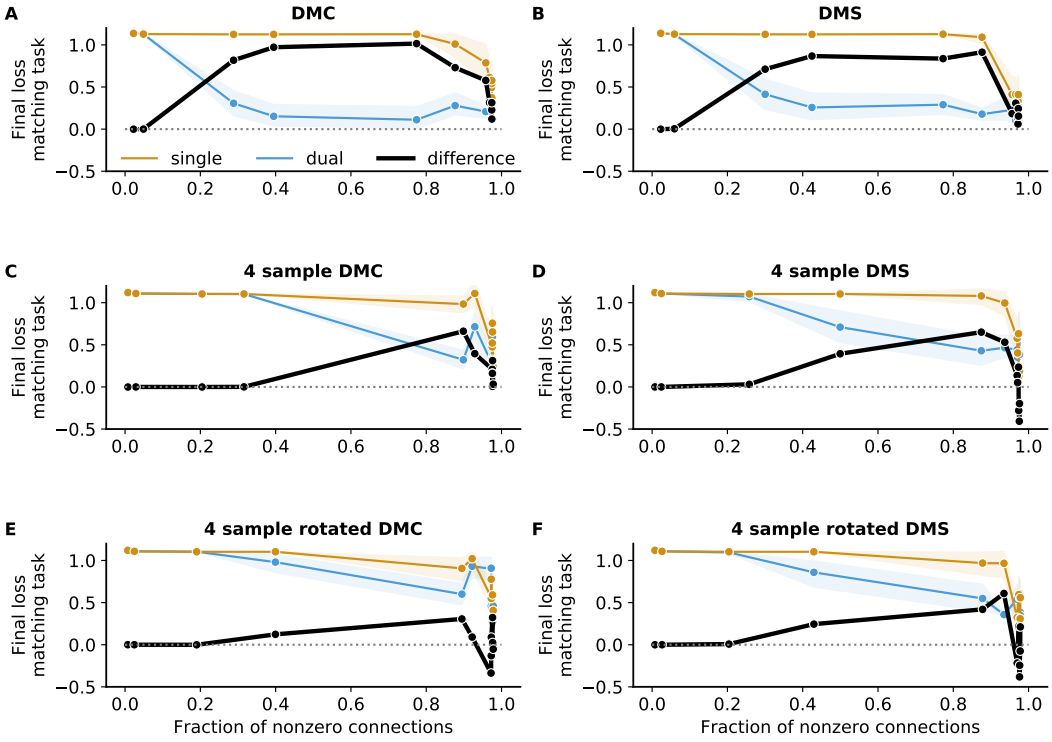


Figure 4.6: Detailed breakdown of Figure 4.5B.
Each panel is as in Figure 4.2C, conducted for each of the 6 tasks as defined in Figure 4.5. NB: For each task the same range of sparsity regularisation parameter λ values were evaluated, which could result in slight variations in number of nonzero connections (hence the variation of the x-axis positions of the data points across panels). The optimal loss is 0 for all tasks.

4.1.5 Simulated annealing enables RNNs to learn the matching task with uncorrelated stimuli

In the previous sections we have shown that eavesdropping only worked with predictable stimuli, because only then a memory of S_1 was created. To prove that the stimulus S_2 predictability – and the thereby induced S_1 memory – only helped to *initiate* the learning process of the matching task, but was not necessary for performing the task after learning converges, we considered the following: we dynamically varied the stimulus S_2 predictability $P(\alpha = \beta)$ from a plateau of 0.5

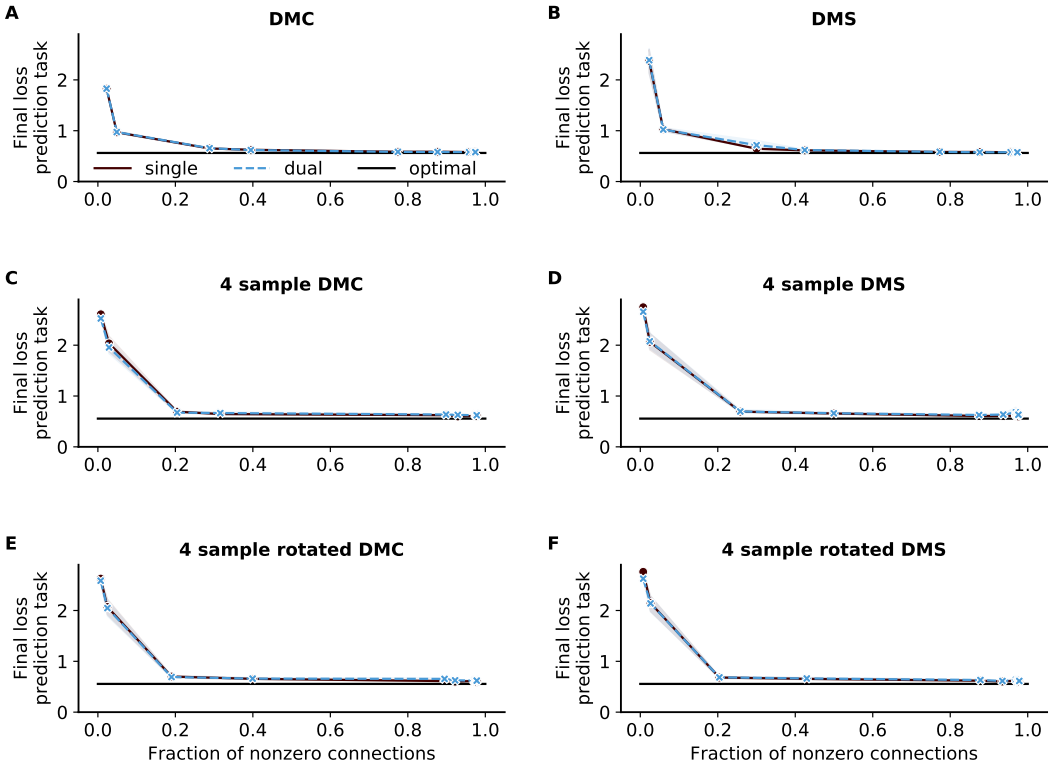


Figure 4.7: The prediction task is solved for all tasks.
Each panel shows the prediction task loss of both the dual task networks and single (prediction) task networks, for each of the 6 tasks as defined in Figure 4.5. Only networks with severe sparsity regularisation, such that almost all connections were 0, failed to learn the task.

to a plateau of 0.75 and back again to 0.5. We name this procedure *simulated annealing*, after the commonly-used technique in Monte Carlo sampling (MacKay, 2003, Ch. 30).

We found that dual task networks that engaged in this setting first learned to solve the prediction task using uncorrelated stimuli, and then gradually learned to memorise S_1 for predicting S_2 as stimulus predictability increased. This S_1 memory could then bootstrap learning of the matching task, and networks maintained a stable solution after the stimuli decorrelated again (Figure 4.8A). Hence, the matching task could be performed with uncorrelated stimuli, but only by simulated annealing of the stimulus S_2 predictability, as networks with constant $P(\alpha = \beta) = 0.5$ did not converge (Figure 4.8B). In summary, a transient period where stimuli are predictable provides the bootstrapping required to learn the demanding matching task, by inducing stimulus memories in the network activity.

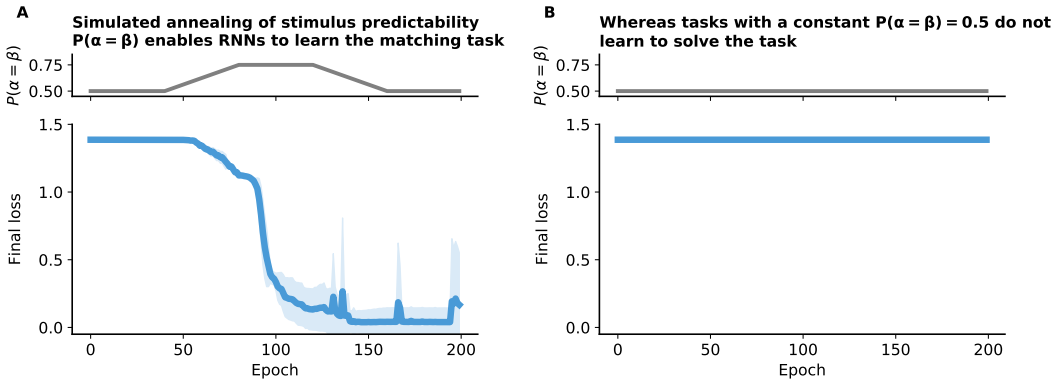


Figure 4.8: Simulated annealing enables RNNs to learn the matching task with uncorrelated stimuli. **A)** Dual task networks were able to learn the matching task with uncorrelated stimuli, by temporarily increasing and decreasing the stimulus S_2 predictability $P(\alpha = \beta)$ from 0.5 to 0.75 and back to 0.5. This simulated annealing of the stimulus predictability causes the prediction task to learn the memory of S_1, which then bootstraps learning of the matching task. Networks are able to maintain their stable solution of the matching task – which does not depend on stimulus S_2 predictability – when the stimuli are uncorrelated again. The end result is a network that has learned the matching task with uncorrelated stimuli. **B)** In contrast, dual task networks are not able to learn the matching task with uncorrelated stimuli without simulated annealing (also see Figure 4.2E). Both panels show the mean $\pm$ std of 10 simulated networks per network type.

4.2 Methods

4.2.1 Terminology

Vectors are denoted by bold notation ($\mathbf{x}$) and matrices by capital notation (X). Time is denoted by subscript t, trials are denoted by subscript k.

4.2.2 Tasks

We considered two tasks in Figures 4.1, 4.2, 4.4, 4.8: first, the prediction task required networks to continuously predict the next upcoming input stimulus, second, the matching task required networks to report whether two stimuli matched at the end of each trial (Figure 4.1A).

In Figure 4.5 we considered the task described above, and 5 additional variants, by changing 1) the input domain, 2) the number of possible stimulus samples and 3) a direct match or rotated match. These are further described in the caption of Figure 4.5.

4.2.3 Synthetic data

Each trial consists of a sequence of 2 stimuli S_1 and S_2 and a GO cue, interleaved by delay periods (Figure 4.1B). Stimuli are one-hot vectors, with $S_1 = A_\alpha \in \{A_1, A_2\}$ and $S_2 = B_\beta \in \{B_1, B_2\}$ (see Figure 4.1B).

Each trial is a sequence of input vectors:

$$\text{trial sequence} = (\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5, \mathbf{z}_6, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_9, \mathbf{z}_{10}, \mathbf{z}_{11}, \mathbf{z}_{12}, \mathbf{z}_{13}) \tag{4.1}$$

where each element $\mathbf{z}_t$ is a one-hot vector:

$$\mathbf{z}_t = \left(1_{\{0\}}, 1_{\{A_1\}}, 1_{\{A_2\}}, 1_{\{B_1\}}, 1_{\{B_2\}}, 1_{\{\text{GO}\}}\right)^T \tag{4.2}$$

Hence, a trial sequence is (in short-hand notation):

$$\text{trial sequence} = \left(0, 0, A_\alpha, A_\alpha, 0, 0, B_\beta, B_\beta, 0, 0, \text{GO}, \text{GO}, 0, 0\right) \tag{4.3}$$

Networks are either trained on *predictable* sequences ($P(\alpha = \beta) = \frac{3}{4}$), or *unpredictable* (uncorrelated) sequences ($P(\alpha = \beta) = \frac{1}{2}$). The duration of all stimuli, GO and the delay periods is 2 time points. This ensures that networks that are trained on predictable sequences will encode both stimuli for optimal performance: A_α information is used to predict the first B_β with 75% accuracy, and then the first B_β is used to predict the second B_β with 100% accuracy.

The network input $\mathbf{x}$ is defined by the one-hot vectors of Equation 4.3 (Figure 4.1B) with added white noise (to prevent networks from overfitting) with standard deviation $\sigma_x = 0.15$ (we found numerically that the results were robust to variations in σ_x). For each network, we randomly generated 1000 trials, which are stratified-split into an 80/20 train/test ratio of trials. All results presented were evaluated on test data only.

4.2.4 Recurrent neural network training

We used standard recurrent neural networks (RNNs), consisting of 1 hidden layer of $N = 20$ neurons that all receive input from 6 input neurons and project to 8 output neurons (Figure 4.1A). The (rate) activity of neurons $\mathbf{r}_t$ is defined by :

$$\mathbf{r}_t = \tanh\left(U \cdot \mathbf{x}_t + W \cdot \mathbf{r}_{t-1} + \mathbf{b}_W\right) \tag{4.4}$$

where $\mathbf{x}_t$ is the input at time t, $\mathbf{b}$ is the learned bias, U is the input weight matrix and W the recurrent weight matrix. At the start of each trial, $\mathbf{r}$ is initialised with $\mathbf{r}_{-1} \sim N(0, 0.1 \cdot I)$.

The 6-element prediction task output vector $\hat{\mathbf{y}}_t^{\text{pred}}$ that estimates the next stimulus is given by a softmax function:

$$\hat{\mathbf{y}}_t^{\text{pred}} = \text{softmax}(V \cdot \mathbf{r}_t + \mathbf{b}_V) \tag{4.5}$$

where V is the prediction output weight matrix. Hence, $\sum_i \hat{y}_{t,i}^{\text{pred}} = 1$, and the network predicts the probability of each possible outcome, rather than one deterministic value. Similarly, the 2 output neurons of the matching task $\hat{\mathbf{y}}_t^{\text{match}}$ must estimate $\mathbf{y}_t^{\text{match}} = (1, 0)^T$ on Match (M, $\alpha = \beta$) trials and $(0, 1)^T$ on Non-Match (NM, $\alpha \neq \beta$) trials, and are defined by

$$\hat{\mathbf{y}}_t^{\text{match}} = \text{softmax}(\text{ReLU}(T \cdot \mathbf{r}_t + \mathbf{b}_T)) \tag{4.6}$$

where T is the task output weight matrix. (Two output units instead of one were used to ensure networks actively had to report both possible outcomes, and an additional ReLU (Rectified linear unit) non-linearity was introduced to prevent the trivial solution $T_{\text{M}} = -T_{\text{NM}}$.) The loss function $\mathcal{L}_k$ of trial k, used to train the model parameters $\theta = \{U, W, V, T, \mathbf{b}\}$, is the sum of the cross-entropy loss of either one or both tasks – depending on the network type – and a sparsity L1 regularisation term:

$$\mathcal{L}_k = \mathcal{L}_k^{\text{pred}} \cdot \mathbb{1}_{\text{pred}} + \mathcal{L}_k^{\text{match}} \cdot \mathbb{1}_{\text{match}} + \lambda \cdot \mathcal{L}_k^{\text{L1}} \tag{4.7}$$

with

$$\mathcal{L}_k^{\text{pred}} = \sum_{t'=4}^{13} \sum_i \left(-y_{k,t',i}^{\text{pred}} \log \hat{y}_{k,t',i}^{\text{pred}}\right) \tag{4.8}$$

$$\mathcal{L}_k^{\text{match}} = \sum_{t'=10}^{11} \sum_j \left(-y_{k,t',j}^{\text{match}} \log \hat{y}_{k,t',j}^{\text{match}} \right) \tag{4.9}$$

$$\mathcal{L}_k^{\text{L1}} = \lambda \cdot ||\theta||_1 \tag{4.10}$$

where λ is the sparsity regularisation parameter (optimised in Figure 4.2), and t' defines the trial times that are evaluated (also see Figure 4.1B). During the training phase, the loss $\mathcal{L}_k$ is accumulated during each trial k. At the end of each trial, $\mathcal{L}_k$ is backpropagated through time to update $\theta = \{U, W, V, T, \mathbf{b}\}$ using stochastic gradient descent (SGD) with a learning rate set to 0.002. Hence, the pseudo-algorithm is given by Algorithm 1:

Algorithm 1: RNN Backpropagation Through Time algorithm

```
generate {x, y}_{k=(1,..,K)} ;
initialise rnn = RNN() with random θ ;
for epoch in E do
    for k in K do
        initiate r_{-1} ~ N(0, 0.1 · I) ;
        for t in T do
            ŷ_{k,t}, r_t = rnn.forward(x_{k,t}, r_{t-1}) using equations 4.4, 4.5 and 4.6 ;
        end
        compute loss l_k(y_k, ŷ_k) using equation 4.7 ;
        update parameters θ using SGD and l_k ;
    end
end
```

where we use $E = 40$ epochs and $K = 800$ trials.

4.2.5 Quantifying stimulus representation strength

The strength of stimulus representations was quantified in Figure 4.4C-H using the stimulus decoding accuracy of logistic regression. We trained decoders to decode the stimulus identity (*e.g.*, S_1) at each time point t: $P(\alpha = 1|\mathbf{r}_t) = 1 - P(\alpha = 2|\mathbf{r}_t)$. Decoders were trained across trials using logistic regression with L1 regularisation (scikit-learn implementation with a SAGA solver, `max_iter` $= 250$ and `C` $= 0.1$ (Pedregosa et al., 2011)). To ensure that the stimulus predictability does not bias the decoders, we trained decoders with a new set of 800/200 train/test trials with uncorrelated stimuli $P(\alpha = \beta) = \frac{1}{2}$. Decoding accuracy is subsequently defined as the mean probability of correctly decoding the stimulus identity.

4.2.6 Software & code availability

All data was analysed and visualised with Python 3.7, using PyTorch as automatic differentiation package, and various other standard libraries for analysis and visualisation (Numpy, Scipy, Scikit-Learn, Pandas, Matplotlib, Seaborn, IPython). All experiments and analysis were done on a 16-thread CPU of a desktop computer. All code and trained network simulations are available on Github. In particular, all figures were made fully in Python, and the code is available on Github.

5 Discussion

5.1 Optogenetic stimulation

5.1.1 Measuring the propagation of neural activity

Critical to the brain's ability to process stimuli is that activity is robustly propagated between functionally and anatomically distinct regions. However, given the technological complexity required to study propagation of neural activity between brain regions *in vivo*, much remains unknown about how the dynamics of neural activity are structured at the single-cell level during propagation (but see Van Vugt et al. (2018); Semedo et al. (2019); Javadzadeh and Hofer (2022); Ebrahimi et al. (2022)). We focused on S1 and S2, two strongly interconnected regions (Kwon et al., 2016; Chen et al., 2016; Yamashita and Petersen, 2016; Yang et al., 2016), where perceptual detection is thought to arise through these recurrent connections (Kwon et al., 2016; Condylis et al., 2020). By using an all-optical interrogation technique across multiple brain areas for the first time, we demonstrate that behaviourally salient photostimulation of S1 as a causal intervention elicits robust activation of S2 (Figures 2.3, 2.7).

As expected, stronger stimulation elicits a more reliable behavioural response (Figure 2.1). Interestingly, we also find that pre-stimulus activity in S1 influences the detectability of the upcoming stimulus (Figures 3.1, 3.3). This is consistent with a signal-to-noise (SNR) framework

in which the signal strength is the number of photostimulated cells and the noise is the pre-stimulus neural variability in S1. Thus our results provide new insights into cortical function as we show for the first time that the SNR of a stimulus, relative to network state, is critical for the propagation of the stimulus and its ability to drive action.

Furthermore, we also performed passive stimulation experiments, where optogenetic stimulation of a similar number of S1 neurons was not associated with reward or behaviour. Here, we did not observe reliable propagation to S2, and pre-stimulus neural variability was not found to modulate the neural response (Figures 3.6-3.11). Hence, our results suggest that the propagation of cortical activation depends on the behavioural context.

5.1.2 Technical limitations

In Chapters 2 and 3, we focused on communication between two neighbouring areas at the single-neuron level. We used state-of-the art recording and manipulation techniques (two-photon imaging and photostimulation), which requires a trade-off between spatiotemporal resolution and field-of-view size. For example, the 'slow' dynamics of GCaMP6s (the best-performing calcium sensor available at the time the experiments were performed) precluded us from reading out millisecond-level spike timings, but allowed us to observe the activity of hundreds of neurons across a field of view spanning more than a square millimetre of cortex. Crucially, these slow dynamics also enabled us to read out the neural activity occurring immediately after trial onset despite the photostimulation artifact (but note that, *e.g.*, recent advances in sensor biotechnology could facilitate a higher temporal resolution (Abdelfattah et al., 2019; Zhang et al., 2023)).

Differences in experimental procedures complicate the comparison of the active and passive stimulation experiments. In particular, the same cells were stimulated per trial type during the passive experiments, which led to adaptation in their response (Figure 3.12), whereas the cell identities varied per trial for the active experiments. Although re-analysis of the first ten trials per trial type resulted in similar outcomes (Figure 3.13), it remains an open question what the effect of maintaining versus changing the target cells is.

5.2 Inhibition stabilisation

5.2.1 Inhibition stabilisation in a stimulus-detection task

One feature of cortical dynamics that may allow activity to reliably propagate is inhibition stabilisation in which inhibition tracks excitation to prevent unstable dynamics (section 1.3). Indeed, we observed that excitation and inhibition co-varied with stimulus strength in S1 and S2, and that both regions demonstrated correlated excitation and inhibition (Figure 2.3). For these active stimulation experiments, our results show that the population response of somatosensory cortex to excitatory photostimulation of small groups of neurons was stabilised by inhibition both in the excited region and in a downstream region. Further, we observed that both the correlation between excitation and inhibition, as well as the fraction of inhibited cells, were lower in S2 than S1.

These findings align with theoretical work from Huang et al. (2022), which firstly predicts that stronger inhibition increases stimulus decodability and (thereby) the propagation of information, and, separately, that decodability decreases as the signal travels downstream. Our observed differences between S1 and S2 corroborate both of these predictions: both inhibition strength and decodability were higher in S1. Hence, extrapolating these theoretical predictions, we would expect that if excitation/inhibition (E/I) correlation would be stronger in S2, decodability would be higher too. This begs the question, why is E/I correlation not stronger in S2? Importantly, the (only) point of reference here is S1, which has a stronger E/I correlation than S2. We believe this strong E/I correlation may be caused by the fact that the stimulation itself – a strong excitatory response – occurs in S1, forcing the local network to strongly inhibit in order to maintain a 'healthy' state. This would confirm theoretical predictions by Marshel et al. (2019), who demonstrated that strong inhibition following network excitation (which leads to strong E/I correlation) is required in order for a large network to be highly excitable (*i.e.*, respond to a small stimulus), see also Figure 1.2. Hence, instead of viewing S2 as weakly E/I correlated (compared to S1), we believe that S1 may be strongly E/I correlated (compared to S2), due to the large optogenetic stimulation that has to be stabilised.

5.2.2 Generalised or random neural stimulus-response

In the active experiments, we found that the population representation of the stimulus was generalised, as the neural response to perceived stimulation was comparable across trials in both regions. We demonstrated this by showing that classifiers robustly decoded hit trials using cross-validation across trials, even though we varied the identity of targeted neurons on each trial (Figure 2.7). This variation of the stimulus requires sufficient consistency of the representation of the stimulus, implying that perceived neural activity remains generalised throughout its journey through the cortex. Conversely, on miss trials, the presence of the stimulus was only decodable for a brief period in S1 immediately following stimulation, and not at all in S2, showing that non-perceived stimulus information was not generalised across both brain regions (Figure 2.7).

Interestingly, passive stimulation did not elicit a generalised neural response either (Figure 3.7). Just like the miss trials of the active stimulation experiments, stimulated trials could not be distinguished from sham trials despite the inhibitory response to stimulation (leading to zero-average activity, Figure 3.6).

In summary, (optogenetic) stimulation of excitatory cells leads to an inhibitory reaction of the network, matching the strength of the stimulus such that the dynamics are stabilised. Without behavioural context, this response is 'random', or unspecific, in the sense that different neurons are inhibited on different trials. We observed this for both passive stimulation and 'missed' active stimulation. However, when the stimulus is associated with reward, as observed on hit trials of active stimulation, the network response remains balanced but generalises, such that the stimulus identity can reliably be decoded from the population activity. Furthermore, only then does neural variability modulate the likelihood that the activity propagates downstream.

5.3 Neural variability

5.3.1 Signal-to-noise ratio as a mechanism for stimulus detection

The SNR of a sensory neuron or population of neurons can be used to quantify the salience of the neural representation of a stimulus, whereby a higher SNR means that the sensory stimulus is more robustly represented in neural activity. Indeed, one of the functions of the highly recurrent circuitry in sensory cortex is thought to be the amplification of relevant activity arising from feedforward inputs, in order to enhance SNR (Kwon et al., 2016; Van Vugt et al., 2018). Despite the well-characterised importance of high SNR in the local representation of sensory stimuli, it is unclear how the SNR of a stimulus relates to its likelihood to propagate downstream. Here we generated a signal with direct cortical activation and quantified noise using the population variance metric. We found that the probability a photostimulus was propagated and perceived was determined by the SNR of the stimulus (Figure 3.3).

5.3.2 Measures of Neural Variance

Previous studies have looked at neural variability in the pre-stimulus time period – though major differences in experimental techniques make it notoriously difficult to directly compare results (Umakantha et al., 2021; Waschke et al., 2021). We expand on this in following, and also summarise the relevant methods covered here in Appendix A.

In Chapter 3, we quantified the neural variability in the moments before optogenetic stimulation by calculating the population variance (Figure 3.1). This aligns with previous findings that reduced variability pre-stimulus is optimal for stimulus detection tasks (McGinley et al., 2015a; Neske et al., 2019). In these studies, variability was quantified by the temporal variance in the membrane potential obtained by whole-cell patch-clamping, measured in auditory (McGinley et al., 2015a) or visual (Neske et al., 2019) cortex. Even though these studies thus studied different sensory cortices, stimuli and recording techniques, they did identify the same directional effect (less variability enhances stimulus detection performance), hinting a possible link between

measures of single-cell and population variability.

Other studies have considered the role of cortical (de)synchronisation during perceptual 2AFC (2-alternative-forced-choice) tasks in the auditory and visual domains. In general, desynchronisation benefits task performance, although this effect has been found to be variable (Jacobs et al., 2020; Reato et al., 2023). Reato et al. (2023) derive synchrony from the temporal variability of the population firing rate obtained from electrophysiology, while Jacobs et al. (2020) analysed the power spectrum of widefield calcium imaging. It is therefore not obvious to quantitatively compare these metrics of desynchronisation to population variance, and would require a thorough analysis (similar to the comparison from Umakantha et al. (2021) on correlations vs PCA). (A key difference between these metrics, is that synchrony requires multiple time points for evaluation and population variance does not.) However, these studies do underscore that pre-stimulus variability impacts sensory processing.

In similar vein, literature on human and monkey perceptual decision making has found that the pre-stimulus power of neural oscillations can impact decision making (Van Vugt et al. (2018); Stephani et al. (2021), and reviewed by Samaha et al. (2020)), typically measured by EEG (electroencephalography) or MEG (magnetoencephalography) in humans or multi-electrode arrays. Similarly, recent work by Davis et al. (2020) found that travelling waves of excitability in MT (middle temporal visual area) could facilitate visual stimulus detection in monkeys, as measured by multi-electrode arrays. This again confirms that pre-stimulus activity fluctuations/variability plays a role in stimulus detection, but the differences in recording method, stimulus modality, animal model and metric of variability preclude us from directly comparing our findings.

Finally, and notably, Waschke et al. (2019) studied the effect of both pre-stimulus neural variability and pupil size for humans performing an auditory pitch discrimination task. This experiment was closed-loop, meaning that stimuli were only presented when neural variability was measured to be significantly high or low (in real-time). Here, neural variability was

quantified by the weighted permutation entropy, a metric of temporal variability of a single time-series. Waschke et al. (2019) found that neural variability and pupil size (thought to index alertness/arousal) differentially influence perceptual performance: neural variability could optimise response time and response bias, while pupil size optimised response sensitivity. Our results – quantifying population instead of temporal variability – confirm that the influence of neural variability might be independent of pupil size (Figure 3.4), corroborating the theory that neural variability might be linked to fast, localised fluctuations in perceptual ability (Waschke et al., 2019).

In sum, these studies confirm our finding that pre-stimulus variability influences sensory processing. To the best of our knowledge, our work is the first to show this effect in an optogenetic stimulus-detection task. This is important, because we thereby directly control the stimulus strength in sensory cortex, which has allowed us to study propagation from the region of stimulation to a downstream region.

5.4 Predictive learning

5.4.1 Summary and limitations

We have shown that predictive learning acted as a helper task for sparse RNNs to achieve optimal performance in a delayed-match-to-category (DMC) task. This was facilitated by eavesdropping, a multitask learning mechanism whereby the easier prediction task created a persistent neural representation of the stimulus that catalysed learning of the matching task. Crucially, it was this stimulus representation (or memory) that was necessary for the matching task to initiate learning. Our results show that predictive learning can induce this representation, and was therefore sufficient to initiate learning of the matching task (Figure 4.4).

Eavesdropping benefits were specific to certain parameters and task variants. The sparsity optimisation showed that eavesdropping is effective for a range of parameter settings that ensure the matching task is neither too difficult (strong sparsity regularisation) nor too easy (no sparsity regularisation) (Figure 4.2). Further, as the matching task complexity increased,

networks failed to converge, even with eavesdropping (Figure 4.5). Lastly, previous studies have optimised very similar matching tasks (including a variant of the rotated match task) without eavesdropping, using similar RNNs of 100-500 hidden units (Yang et al., 2019; Masse et al., 2019; Orhan and Ma, 2019). Our results show that although smaller networks naturally fail to learn these memory tasks, eavesdropping is able to facilitate success.

Small, interpretable neural networks are crucial for understanding learning mechanisms (Rudin, 2019), and can provide a good approximation of larger biological networks (Barak, 2017). Previous studies have shown that larger networks with more complex architectures can use predictive learning to boost performance in more difficult tasks, such as 3D navigation (Wayne et al., 2018), but their underlying mechanisms are difficult to dissect. After we confirmed that task performance was equivalent across a range of network sizes (Figure 4.3), we adhered to a minimal working example of predictive learning, which enabled us to uncover eavesdropping as its facilitating mechanism (Figure 4.4). Notably, the stimulus-prediction and DMC tasks that we considered are typical of controlled experiments performed in animals and humans (Freedman and Assad, 2006; Wan et al., 2020; Mohan et al., 2021; Libby and Buschman, 2021; Zhou et al., 2021), which allows us to use this theoretical study to make experimental predictions that can readily be tested *in vivo*. We will do so in section 5.4.3. Hence, crucially, the aim of our study was not to develop a new Machine Learning heuristic, but rather to use artificial networks to further develop our understanding of biological learning.

5.4.2 Predictive learning and eavesdropping in natural learning

Eavesdropping is a form of multitask learning where two (or more) tasks that are optimised simultaneously share a latent feature that can only (easily) be learned in one of the tasks, hence catalysing learning in the other task (Caruana, 1997). Eavesdropping, and multitask learning more generally, is conceptually related to both transfer learning and curriculum learning, which both use 'helper' tasks to improve learning. All these learning strategies have been proposed to occur naturally in animals (Elman, 1993; Ben-David and Schuller, 2003; Bengio et al., 2009;

Weiss et al., 2016; Hassabis et al., 2017). However, it remains unclear what the helper task of eavesdropping could be in neural systems.

Simultaneously, predictive learning has been shown to benefit reinforcement learning (RL) agents (Russek et al., 2017; Wayne et al., 2018). In particular, the successor representation – a form of predictive learning (Dayan, 1993; Russek et al., 2017) – has been able to provide a link between model-free and model-based RL (Russek et al., 2017), which are both known to occur in the brain (Daw et al., 2011).

In neuroscience, brains are thought to continuously engage in predictive learning, often motivated by normative frameworks (Rao and Ballard, 1999; Friston et al., 2016). Our study builds on this, by proposing an additional function of predicting sensory inputs. We have linked eavesdropping and predictive learning, by showing that predictive learning was sufficient to induce eavesdropping in environments where stimuli are correlated in time. Other learning processes that create stimulus representations, such as solving a delayed-response task (Yang et al., 2019; Duncker et al., 2020), could potentially also facilitate eavesdropping. However, in this work we have focused on predictive learning, given its prevalence in neuroscience.

Prediction may be useful as sensory correlations are ubiquitous. Our study suggests that prediction functions as a natural prerequisite for the ability to remember and process sensory experiences over time. Importantly, this does not entail that ongoing predictability is necessary to maintain a memory: In Figure 4.8 we show that transient predictability is already enough to bring the system into a new minimum, which remains stable after stimulus predictability disappears. Thus, brief moments during which a complex task can be deconstructed into smaller parts can reveal a path to the solution, after which these components can be disregarded again (also see Knoblich et al. (1999)).

5.4.3 Experimental predictions

Libby and Buschman (2021) found that mice actively encoded stimulus predictions during a passive learning paradigm (i.e., with no reward structure). Our study suggests that this learning is functionally important: representing stimulus correlations that are not directly relevant for behaviour might still be a rewarding long-term strategy. Specifically, predictive learning can act as a module that facilitates learning of complex future task structures.

Taking these experimental findings together with our results, we propose the following hypothesis. Animals that have to solve a complex working memory task, such as DMC, should benefit from prior passive exposure to correlated sequences of the stimuli (that are later used in the DMC task). With no requirement for reward, they should internally attempt to predict these stimuli, thereby creating neural representations of the stimuli that can later accelerate learning. Prior exposure to *uncorrelated* sequences, in contrast, should not benefit learning, as random sequences cannot be predicted.

5.5 Conclusion

In summary, we first developed a new preparation to inject sparse, task-relevant activity into the cortex while recording its propagation downstream and resulting behaviour. Using this, we elucidated how, at the level of neuronal networks, perception depends on the signal-to-noise ratio of the stimulus relative to background activity. In contrast, passive cortical stimulation of similar magnitude did not generate reliable activity propagation downstream, even when neural variability was low. Together this reveals an organising principle of the cortex that explains how inhibition-stabilised circuits can remain susceptible to behaviourally relevant stimuli. Secondly, our modelling study revealed that previously observed neural populations that engage in predictive learning might be crucial when neural networks are faced with demanding working memory tasks.

Appendices

A

Neural variability estimation methods in neuroscience

Effective dimensionality of neural networks

Neural networks in the brain consist of large numbers of neurons, which work together to engage in cognitive functioning. The neural activity of brain networks is therefore not random, but rather displays coordinated and correlated dynamics (Stringer et al., 2019; Urai et al., 2022). This degree of coordination of activity can be quantified by the *effective dimensionality* of the network. The effective dimensionality describes how many variables are needed to explain the activity of the network, and is therefore generally lower than the total *network dimensionality* (*i.e.*, number of neurons) for connected networks (Gao et al., 2017), see Figure A.1. This property has been widely leveraged by neuroscientists to efficiently represent large-scale neural data using a small number of dimensions, called dimensionality reduction, for ease of both analysis and visualisation (Cunningham and Byron, 2014; Engel and Steinmetz, 2019). However, beyond a tool, the magnitude of the effective dimensionality is a defining property of a neural system, and is closely related to the variability of a neural population (Umakantha et al., 2021).

Linear versus nonlinear dimensionality estimation methods In section 1.4, we reviewed how learning can gradually change the effective dimensionality of a neural population (Ni et al., 2018). A key feature of learning is to create an internal model of the environment

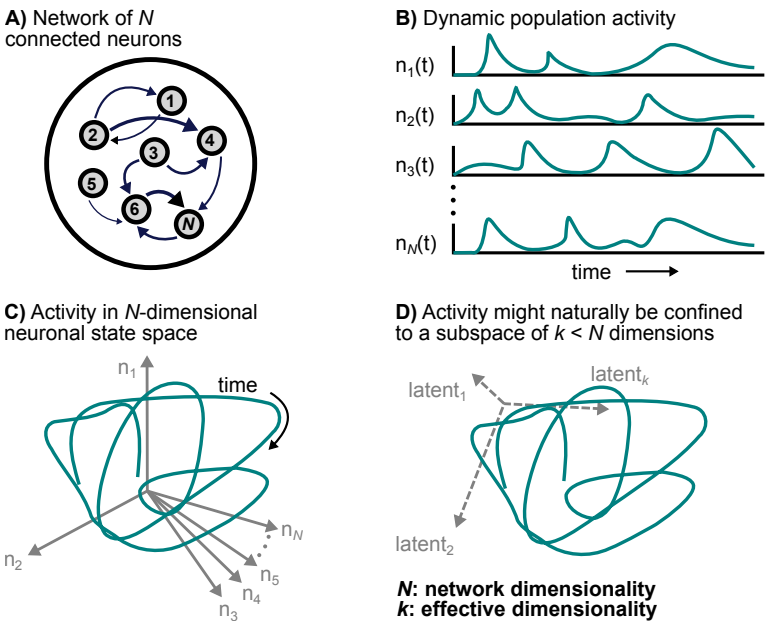


Figure A.1: Schematic explanation of effective dimensionality.
A) Connected network of N neurons **B)** Dynamic activity of all neurons. **C)** Population activity of all neurons in their (high) N-dimensional state space. **D)** Although the state space is N-dimensional, the activity might be confined to a subspace that can be described by $k < N$ dimensions. k is therefore called the effective dimensionality of the network. If $k \ll N$ then the population activity can thus be described by a low-dimensional representation. NB: this example shows a linear dimensionality reduction, for ease of visualisation.

for model-based reinforcement learning (Sutton and Barto, 2018). Natural environments are generally characterised by a low-dimensional structure, but conveyed through high-dimensional observations. Recanatesi et al. (2021) show that neural networks that learn to represent such environments reflect this property. Linear measurements of the effective dimensionality are high-dimensional, as a result of the high-dimensional observations. But the low-dimensional latent structure is expressed locally, quantified by nonlinear measurements of effective dimensionality.

Evidently, linear and nonlinear measurements yield different results. However, they are not contradictory, but rather capture different aspects of how population activity is structured (Recanatesi et al., 2021; Jazayeri and Ostojic, 2021). The nonlinear dimensionality – called *intrinsic dimensionality* – generally is the minimal effective dimensionality (as it encapsulates linear estimates), quantifying the minimal number of variables needed to explain the population activity (Jazayeri and Ostojic, 2021). The linear dimensionality is often referred to as the *embedded dimensionality*, and it is important because it describes the dimensionality of the system to linear readouts (a simplification often used for downstream brain regions) (Kaufman et al., 2014; Semedo et al., 2019; Jazayeri and Ostojic, 2021).

Method	Description	Source	EC	ST	I
		Correlation or variance methods			
Temporal variability	Standard deviation across time of single time series data.	McGinley et al. (2015a); Neske et al. (2019)	x	✓	✓
Population variance	Instantaneous variance across neurons.	Rowland*, Van der Plas*, Loidolt* et al. (2023)	x	x	✓
Across-trial variability	Instantaneous variance across trials.	Churchland et al. (2010)	x	x	✓*
Fano factor	Spike count covariance divided by spike count mean.	Cohen and Maunsell (2009); Churchland et al. (2010)	x	✓	✓*
Noise correlations	Pearson correlation between neurons after subtracting trial-average responses.	Cohen and Maunsell (2009); Rumyantsev et al. (2020)	✓	✓	x
		PCA eigenvalue methods			
Shared dimensionality	Number of PCs required for 95% explained variance of shared covariance matrix (determined by factor analysis).	Machens et al. (2010); Williamson et al. (2016)	✓	?	x
Shared variability	Magnitude of first eigenvalue (λ_1) from shared covariance matrix, obtained by factor analysis.	Huang et al. (2019)	✓	?	x
Participation ratio (PR) of PCA eigenvalues	PR $= \frac{1}{\sum_i \tilde{\lambda}_i}$ where $\tilde{\lambda}_i = \frac{\lambda_i}{\sum_i \lambda_i}$ and λ_i is the i$^{\text{th}}$eigenvalue.	Abbott et al. (2011); Mazzucato et al. (2016); Gao et al. (2017); Farrell et al. (2022)	✓	?[1]	✓*
Dimensionality gain (DG)	Compute 1) the PR of Eigenvalues (*i.e.*, the global linear dimensionality), 2) a local, nonlinear estimate of intrinsic dimensionality (ID) (using various methods). DG = PR / ID. This indicates how strongly the global representation is compressed locally.	Recanatesi et al. (2021)	✓	?	x
		Classification methods			
Shattering dimensionality	Number of binary classification theoretically possible with linear classifiers.	Rigotti et al. (2013); Fusi et al. (2016); Nogueira et al. (2023)	✓	x	x

[1] Mazzucato et al. (2016) applied this method to 1 second windows, consisting of 5 time points. This limits the upper bound of the dimensionality estimate when the number of neurons $N > 5$ (because 5 PCs can explain all covariance). (Mazzucato et al. (2016) used $2 < N < 10$.) Hence, the question mark for all PCA-based methods reflects the limited applicability to single-trial data, as the ratio of number of time points/number of neurons might be insufficient.

Bottleneck methods					
Nonlinear PCA†	5-layer autoencoder, where the 3rd layer is the bottleneck (and the minimal number of neurons is the dimensionality).	Camastra (2003); Camastra and Staiano (2016)	✓	x	x
Restricted Boltzmann Machines	2-layer network with a high-dimensional and low-dimensional layer (the minimal number of units in the low-dimensional layer is the effective dimensionality).	Van der Plas*, Tubiana* et al. (2023)	✓	x	x
Fractal methods					
Correlation dimension†	Start with a hypersphere of radius ϵ at data point $\mathbf{x}$. Grow ϵ and measure the number of other data points $N_{\mathbf{x}}$ that are within this hypersphere. This typically grows as a power law and the correlation dimensionality d as $C \propto \epsilon^d$, after averaging over $\mathbf{x}$.	Camastra (2003); Strogatz (2015); Camastra and Staiano (2016)	✓	x	x
Two nearest neighbour method†	For each data point $\mathbf{x}$, find the difference to the two nearest neighbours. The ratio between these two is distributed according to a known (Pareto) distribution, with one free parameter, the intrinsic dimension. This can therefore be inferred.	Facco et al. (2017); Ansuini et al. (2019)	✓	x	x
Very high temporal resolution methods					
Normalised synchrony	Standard deviation across time of the instantaneous population firing rate, normalised by shuffled estimates.	Reato et al. (2023)	x	✓	x
Spectral power	Power estimates of 300 ms windows of relevant spectral bands (α, β, γ).	Van Vugt et al. (2018)	x	✓	✓
Weighted permutation entropy	A single time series is converted to a series of ranked motifs, which are used to calculate the (variance-weighted) entropy of a running window of 200 ms, compared against the past 30 s.	Waschke et al. (2019)	x	x	✓

Table A.1: Table 1: Overview of methods used in neuroscience to quantify the variability of a neural population (also see Waschke et al. (2021)). The last three columns indicate whether a method can be used (✓: yes, ?: unknown, x: no) at the following three resolutions: experimental condition (EC), single trial (ST) and instantaneous (I)), see Figures 1.5, 1.4. For I, ✓* indicates 'yes, across-trials only'. † indicates that, to the best of our knowledge, this method has not been applied to neuroscience data (yet).

B

Derivation of optimal loss for RNN prediction task

In the following, we derive the optimal (*i.e.*, lowest) loss for the prediction task of the RNNs (as plotted in Figures 4.1D, 4.7).

First, we realise that the only time point that cannot be predicted with certainty (*i.e.*, zero loss) is the first S_2 stimulus (see Equation 4.3). Hence we only need to compute the optimal loss at this time point to determine the optimal loss across an entire trial. We will only consider the DMC task where $S_2 = B_\beta$, noting that DMS is equivalent.

First case: 2-sample, one-hot

Here, the optimal loss can be calculated as follows. Let us denote the true probability of stimulus B_β with p_β. If $S_1 = A_1$, then:

$$p_{\beta=1} = 0.75, \quad p_{\beta=2} = 0.25. \tag{B.1}$$

and vice versa ($\alpha = 1$ and $\alpha = 2$ occur with equal probability so we will only consider one case). The loss function l, used to train the model parameters $\theta = \{U, W, V\}$, is the sum of

a cross-entropy and L1 regularisation term:

$$l_k = \sum_{t'=5}^{16} \sum_i (-y_{k,t',i} \log \hat{y}_{k,t',i}) + \lambda \cdot ||\theta||_1 \tag{B.2}$$

where $\lambda = 0.005$, and t' defines the trial times of the prediction task (5, 6, 7, ... 16) (*i.e.*, the A_α stimuli are not taken into account). Note that the minimum for the cross-entropy loss H of a binomial probability p is found by:

$$H = -p \log \hat{p} - (1-p) \log(1-\hat{p}) \Longrightarrow \frac{dH}{d\hat{p}} = -\frac{p}{\hat{p}} + \frac{(1-p)}{(1-\hat{p})} = 0$$

$$\frac{(1-p)}{(1-\hat{p})} = \frac{p}{\hat{p}} \iff (1-p) \cdot \hat{p} = p \cdot (1-\hat{p}) \iff \hat{p} = p \tag{B.3}$$

The only time point that cannot be predicted with certainty (*i.e.*, zero cross-entropy loss) is the first S_2 stimulus (S_2^1) at the β-predicting elements of the stimulus vector. Hence, the loss becomes:

$$\begin{aligned} l_k &= \sum_{t'=5}^{16} \sum_i (-y_{k,t',i} \log \hat{y}_{k,t',i}) = \sum_{t \in S_2^1} \sum_{i \in \beta} (-y_{t',i} \log \hat{y}_{t',i}) \\ &= -p \cdot \log \hat{p} - (1-p) \cdot \log(1-\hat{p}) = -p \cdot \log p - (1-p) \cdot \log(1-p) \end{aligned} \tag{B.4}$$

which is approximately 0.562335 for $p = p_{\beta 1} = \frac{3}{4}$.

Second case: 4-sample, periodic

This derivation is a little more involved. We consider the *non*-rotated task, noting that the rotated task is equivalent. If $S_1 = A_0$ (where the other 3 possibilities ($\pi/2$, π, $3\pi/2$) are equivalent and equally probable.

$$p_{\beta=0} = \frac{3}{4}, \quad p_{\beta=\pi/2} = p_{\beta=\pi} = p_{\beta=3\pi/2} = \frac{1}{12} \tag{B.5}$$

For this task, the mean squared error loss function MLS was used (section 4.1.4):

$$MLS = (y_{\text{true}} - y_{\text{estimated}})^2 \tag{B.6}$$

To calculate the loss, we sum over time points and vectors, using again that the first S_2 time point is the only time point with a non-zero optimal loss. In the previous, one-hot encoded

2-sample, prediction task, the S_β stimulus $\begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$ (see Figure 4.1A) could be either $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ or $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Now, in the periodic 4-sample task, there are four possible states: $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ with $p_{\beta=0} = \frac{3}{4}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ with $p_{\beta=\pi/2} = \frac{1}{12}$, $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$ with $p_{\beta=\pi} = \frac{1}{12}$ and $\begin{pmatrix} 0 \\ -1 \end{pmatrix}$ with $p_{\beta=3\pi/2} = \frac{1}{12}$. To calculate the combined loss, we denote the optimal solution as $\begin{pmatrix} x \\ y \end{pmatrix}$. Then:

$$MLS_{x,y} = \frac{1}{12}\left(9\begin{pmatrix} 1-x \\ -y \end{pmatrix}^2 + \begin{pmatrix} -x \\ 1-y \end{pmatrix}^2 + \begin{pmatrix} -1-x \\ -y \end{pmatrix}^2 + \begin{pmatrix} -x \\ -1-y \end{pmatrix}^2\right) \tag{B.7}$$

Simplify to:

$$MLS_{x,y} = \frac{1}{12}\begin{pmatrix} 9(1-x)^2 + x^2 + (-1-x)^2 + x^2 \\ 9\cdot y^2 + (1-y)^2 + y^2 + (-1-y)^2 \end{pmatrix} \tag{B.8}$$

$$MLS_{x,y} = \frac{1}{12}\begin{pmatrix} 9(x^2-2x+1) + x^2 + (x^2+2x+1) + x^2 \\ 9\cdot y^2 + (y^2-2y+1) + y^2 + (y^2+2y+1) \end{pmatrix} \tag{B.9}$$

$$MLS_{x,y} = \frac{1}{12}\begin{pmatrix} 12x^2 - 16x + 10 \\ 12y^2 + 2 \end{pmatrix} \tag{B.10}$$

The MLS loss sums the vector elements:

$$MLS_{x,y} = \frac{1}{12}\left(12x^2 - 16x + 12 + 12y^2\right) \tag{B.11}$$

$$MLS_{x,y} = \frac{1}{3}\left(3x^2 - 4x + 3 + 3y^2\right) = \frac{1}{3}\left(3(x^2+y^2+1) - 4x\right) \tag{B.12}$$

$$MLS_{\text{optimal}} = \min_{x,y}\frac{1}{3}\left(3(x^2+y^2+1) - 4x\right) \tag{B.13}$$

The minima can be found separately:

$$\begin{pmatrix} x \\ y \end{pmatrix}_{\text{optimal}} = \begin{pmatrix} \text{argmin}_x\left(3x^2 - 4x + 3\right) \\ \text{argmin}_y\, 3y^2 \end{pmatrix} = \begin{pmatrix} \text{argmin}_x\left((x-\frac{2}{3})^2 + \frac{5}{9}\right) \\ 0 \end{pmatrix} \tag{B.14}$$

which leads to:

$$\begin{pmatrix} x_{\text{optimal}} \\ y_{\text{optimal}} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \\ 0 \end{pmatrix} \quad \text{and} \quad MLS_{\text{optimal}} = \frac{5}{9} \tag{B.15}$$

*: co-first author.
†: co-senior author.

Bibliography

Abbott, L. F., Rajan, K., and Sompolinsky, H. (2011). Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. In Ding, M. and Glanzman, D. L., editors, *The dynamic brain: an exploration of neuronal variability and its functional significance*, chapter 4, pages 65–82. Oxford university press. pages 90

Abdelfattah, A. S., Kawashima, T., Singh, A., Novak, O., Liu, H., Shuai, Y., Huang, Y.-C., Campagnola, L., Seeman, S. C., Yu, J., et al. (2019). Bright and photostable chemigenetic indicators for extended in vivo voltage imaging. *Science*, 365(6454):699–704. pages 78

Alexander, W. H. and Brown, J. W. (2018). Frontal cortex function as derived from hierarchical predictive coding. *Scientific reports*, 8(1):1–11. pages 16

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. pages 91

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366. pages 11

Bar, M. (2009). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1235–1243. pages 16

Barak, O. (2017). Recurrent neural networks as versatile tools of neuroscience research. *Current opinion in neurobiology*, 46:1–6. pages 4, 17, 84

Bartolo, R., Saunders, R. C., Mitz, A. R., and Averbeck, B. B. (2020). Information-limiting correlations in large neural populations. *Journal of Neuroscience*, 40(8):1668–1678. pages 11

Bell, P. and Renals, S. (2015). Regularization of context-dependent deep neural networks with context-independent multi-task training. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4290–4294. IEEE. pages 60

Ben-David, S. and Schuller, R. (2003). Exploiting task relatedness for multiple task learning. In *Learning theory and kernel machines*, pages 567–580. Springer. pages 84

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. pages 84

Camastra, F. (2003). Data dimensionality estimation methods: a survey. *Pattern recognition*, 36(12):2945–2954. pages 91

Camastra, F. and Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41. pages 91

Carpenter, R. and Reddi, B. (2012). *Neurophysiology: a conceptual approach*. CRC Press. pages 1, 5

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75. pages 60, 62, 84

Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., and Wang, X.-J. (2017). Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron*, 93(6):1504–1517. pages 17

Chen, J. L., Voigt, F. F., Javadzadeh, M., Krueppel, R., and Helmchen, F. (2016). Long-range population dynamics of anatomically defined neocortical networks. *Elife*, 5:e14679. pages 4, 18, 77

Chen, T.-W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300. pages 19, 34, 35, 57, 58

Chettih, S. N. and Harvey, C. D. (2019). Single-neuron perturbations reveal feature-specific competition in v1. *Nature*, 567(7748):334–340. pages 7, 18, 19, 34, 57

Churchland, M. M., Byron, M. Y., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369–378. pages 12, 13, 14, 90

Cohen, M. R. and Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600. pages 1, 2, 9, 11, 12, 13, 90

Condylis, C., Lowet, E., Ni, J., Bistrong, K., Ouellette, T., Josephs, N., and Chen, J. L. (2020). Context-dependent sensory processing across primary and secondary somatosensory cortex. *Neuron*. pages 4, 77

Cunningham, J. P. and Byron, M. Y. (2014). Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509. pages 88

Dalgleish, H. W., Russell, L. E., Packer, A. M., Roth, A., Gauld, O. M., Greenstreet, F., Thompson, E. J., and Häusser, M. (2020). How many neurons are sufficient for perception of cortical activity? *Elife*, 9:e58889. pages 2, 8, 9, 18, 20, 36, 40

Davis, Z. W., Muller, L., Martinez-Trujillo, J., Sejnowski, T., and Reynolds, J. H. (2020). Spontaneous travelling cortical waves gate perception in behaving primates. *Nature*, 587(7834):432–436. pages 82

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215. pages 16, 85

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624. pages 85

Denève, S. and Machens, C. K. (2016). Efficient codes and balanced networks. *Nature neuroscience*, 19(3):375–382. pages 5

Diamond, M. E., Von Heimendahl, M., Knutsen, P. M., Kleinfeld, D., and Ahissar, E. (2008). 'where' and 'what' in the whisker sensorimotor system. *Nature Reviews Neuroscience*, 9(8):601–612. pages 3, 4, 19

Doiron, B., Litwin-Kumar, A., Rosenbaum, R., Ocker, G. K., and Josić, K. (2016). The mechanics of state-dependent neural correlations. *Nature neuroscience*, 19(3):383–393. pages 11

Driscoll, L., Shenoy, K., and Sussillo, D. (2022). Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *bioRxiv*, pages 2022–08. pages 17

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., and Ostojic, S. (2022). The role of population structure in computations through neural dynamics. *Nature neuroscience*, 25(6):783–794. pages 17

Duncker, L., Driscoll, L., Shenoy, K. V., Sahani, M., and Sussillo, D. (2020). Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in neural information processing systems*, 33:14387–14397. pages 17, 69, 85

Ebrahimi, S., Lecoq, J., Rumyantsev, O., Tasci, T., Zhang, Y., Irimia, C., Li, J., Ganguli, S., and Schnitzer, M. J. (2022). Emergent reliability in sensory cortical coding and inter-area communication. *Nature*, 605(7911):713–721. pages 1, 77

El-Boustani, S., Sermet, B. S., Foustoukos, G., Oram, T. B., Yizhar, O., and Petersen, C. C. (2020). Anatomically and functionally distinct thalamocortical inputs to primary and secondary mouse whisker somatosensory cortices. *Nature communications*, 11(1):3342. pages 4

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99. pages 84

Engel, T. A. and Steinmetz, N. A. (2019). New perspectives on dimensionality and variability from large-scale cortical dynamics. *Current opinion in neurobiology*, 58:181–190. pages 88

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8. pages 91

Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., and Shea-Brown, E. (2022). Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. *Nature Machine Intelligence*, 4(6):564–573. pages 12, 13, 14, 90

Freedman, D. J. and Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–88. pages 84

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879. pages 16, 85

Fusi, S., Miller, E. K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74. pages 90

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., and Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, page 214262. pages 88, 90

Grady, C. L. and Garrett, D. D. (2018). Brain signal variability is modulated as a function of internal and external demand in younger and older adults. *Neuroimage*, 169:510–523. pages 9

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258. pages 85

Hennequin, G., Agnes, E. J., and Vogels, T. P. (2017). Inhibitory plasticity: balance, control, and codependence. *Annual Review of Neuroscience*, 40:557–579. pages 5

Hennequin, G., Vogels, T. P., and Gerstner, W. (2014). Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406. pages 5, 17

Houweling, A. R. and Brecht, M. (2008). Behavioural report of single neuron stimulation in somatosensory cortex. *Nature*, 451(7174):65–68. pages 18

Huang, C., Pouget, A., and Doiron, B. (2022). Internally generated population activity in cortical networks hinders information transmission. *Science Advances*, 8(22):eabg5244. pages 5, 79

Huang, C., Ruff, D. A., Pyle, R., Rosenbaum, R., Cohen, M. R., and Doiron, B. (2019). Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101(2):337–348. pages 2, 5, 11, 90

Jacobs, E. A., Steinmetz, N. A., Peters, A. J., Carandini, M., and Harris, K. D. (2020). Cortical state fluctuations during sensory decision making. *Current Biology*, 30(24):4944–4955. pages 82

Javadzadeh, M. and Hofer, S. B. (2022). Dynamic causal communication channels between neocortical areas. *Neuron*, 110(15):2470–2483. pages 12, 13, 14, 77

Jazayeri, M. and Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70:113–120. pages 89

Kanashiro, T., Ocker, G. K., Cohen, M. R., and Doiron, B. (2017). Attentional modulation of neuronal variability in circuit models of cortex. *Elife*, 6:e23978. pages 5

Kao, T.-C., Sadabadi, M. S., and Hennequin, G. (2021). Optimal anticipatory control as a theory of motor preparation: a thalamo-cortical circuit model. *Neuron*, 109(9):1567–1581. pages 5

Kaufman, M. T., Churchland, M. M., Ryu, S. I., and Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. *Nature neuroscience*, 17(3):440–448. pages 89

Knoblich, G., Ohlsson, S., Haider, H., and Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, memory, and cognition*, 25(6):1534. pages 85

Kwon, S. E., Yang, H., Minamisawa, G., and O'Connor, D. H. (2016). Sensory and decision-related activity propagate in a cortical feedback loop during touch perception. *Nature neuroscience*, 19(9):1243–1252. pages 4, 18, 77, 81

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436. pages 17

Libby, A. and Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, pages 1–12. pages 2, 16, 84, 86

Liu, R., Yang, B., and Hauptmann, A. G. (2020). Simultaneous bearing fault recognition and remaining useful life prediction using joint-loss convolutional neural network. *IEEE Transactions on Industrial Informatics*, 16(1):87–96. pages 60

Machens, C. K., Romo, R., and Brody, C. D. (2010). Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *Journal of Neuroscience*, 30(1):350–360. pages 90

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press. pages 71

Maestú, F., de Haan, W., Busche, M. A., and DeFelipe, J. (2021). Neuronal excitation/inhibition imbalance: core element of a translational perspective on alzheimer pathophysiology. *Ageing Research Reviews*, 69:101372. pages 5

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84. pages 1, 13, 17

Marshel, J. H., Kim, Y. S., Machado, T. A., Quirin, S., Benson, B., Kadmon, J., Raja, C., Chibukhchyan, A., Ramakrishnan, C., Inoue, M., et al. (2019). Cortical layer–specific critical dynamics triggering perception. *Science*, 365(6453):eaaw5202. pages 2, 5, 6, 7, 8, 18, 40, 79

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., and Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature neuroscience*, 22(7):1159–1167. pages 17, 69, 84

Mastrogiuseppe, F. and Ostojic, S. (2018). Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623. pages 17

Mazzucato, L., Fontanini, A., and La Camera, G. (2016). Stimuli reduce the dimensionality of cortical activity. *Frontiers in systems neuroscience*, 10:11. pages 13, 14, 90

McGinley, M. J., David, S. V., and McCormick, D. A. (2015a). Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron*, 87(1):179–192. pages 44, 81, 90

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., Tolias, A. S., Cardin, J. A., and McCormick, D. A. (2015b). Waking state: rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161. pages 9

Meredith, M. A. and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology*, 56(3):640–662. pages 1

Migault, G., van der Plas, T. L., Trentesaux, H., Panier, T., Candelier, R., Proville, R., Englitz, B., Debrégeas, G., and Bormuth, V. (2018). Whole-brain calcium imaging during physiological vestibular stimulation in larval zebrafish. *Current Biology*, 28(23):3723–3735. pages 1

Mohan, K., Zhu, O., and Freedman, D. J. (2021). Interaction between neuronal encoding and population dynamics during categorization task switching in parietal cortex. *Neuron*, 109(4):700–712. pages 84

Nagai, Y. (2019). Predictive learning: its key role in early cognitive development. *Philosophical Transactions of the Royal Society B*, 374(1771):20180030. pages 16

Nelson, S. B. and Valakh, V. (2015). Excitatory/inhibitory balance and circuit homeostasis in autism spectrum disorders. *Neuron*, 87(4):684–698. pages 5

Neske, G. T., Nestvogel, D., Steffan, P. J., and McCormick, D. A. (2019). Distinct waking states for strong evoked responses in primary visual cortex and optimal visual detection performance. *Journal of Neuroscience*, 39(50):10044–10059. pages 81, 90

Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J., and Cohen, M. R. (2018). Learning and attention reveal a general relationship between population activity and behavior. *Science*, 359(6374):463–465. pages 2, 9, 11, 12, 13, 88

Ni, J. and Chen, J. L. (2017). Long-range cortical dynamics: a perspective from the mouse sensorimotor whisker system. *European Journal of Neuroscience*, 46(8):2315–2324. pages 3, 4

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25. pages 58

Nicole, O. and Pacary, E. (2020). Camkiiβ in neuronal development and plasticity: An emerging candidate in brain diseases. *International Journal of Molecular Sciences*, 21(19):7272. pages 34

Nikolenko, V., Watson, B. O., Araya, R., Woodruff, A., Peterka, D. S., and Yuste, R. (2008). Slm microscopy: scanless two-photon imaging and photostimulation using spatial light modulators. *Frontiers in neural circuits*, 2:393. pages 7

Nogueira, R., Rodgers, C. C., Bruno, R. M., and Fusi, S. (2023). The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26(2):239–250. pages 90

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*. pages 16

Orhan, A. E. and Ma, W. J. (2019). A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nature neuroscience*, 22(2):275–283. pages 17, 67, 84

Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D., and Ferster, D. (2009). Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–592. pages 5

Pachitariu, M., Stringer, C., Dipoppa, M., Schröder, S., Rossi, L. F., Dalgleish, H., Carandini, M., and Harris, K. D. (2017). Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *Biorxiv*, page 061507. pages 35, 57

Packer, A. M., Russell, L. E., Dalgleish, H. W., and Häusser, M. (2015). Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nature methods*, 12(2):140–146. pages 4, 7, 18, 19, 34

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830. pages 37, 58, 75

Pollock, E. and Jazayeri, M. (2020). Engineering recurrent neural networks from task-relevant manifolds and dynamics. *PLoS computational biology*, 16(8):e1008128. pages 17

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87. pages 85

Reato, D., Steinfeld, R., Tacão-Monteiro, A., and Renart, A. (2023). Response outcome gates the effect of spontaneous cortical state fluctuations on perceptual decisions. *eLife*, 12:e81774. pages 5, 12, 13, 82, 91

Recanatesi, S., Farrell, M., Lajoie, G., Deneve, S., Rigotti, M., and Shea-Brown, E. (2021). Predictive learning as a network mechanism for extracting low-dimensional latent space representations. *Nature communications*, 12(1):1–13. pages 16, 89, 90

Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. D. (2010). The asynchronous state in cortical circuits. *science*, 327(5965):587–590. pages 5

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590. pages 90

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nature neuroscience*, 20(1):107–114. pages 5

Rowland*, J. M., van der Plas*, T. L., Loidolt*, M., Lees, R. M., Keeling, J., Dehning, J., Akam, T., Priesemann, V., and Packer, A. M. (2023). Propagation of activity through the cortical hierarchy and perception are determined by neural variability. *Nature Neuroscience*, 6:1584–1594. pages vi, vi, vi, vi, 12, 13, 31, 34, 44, 46, 90

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. pages 84

Rumyantsev, O. I., Lecoq, J. A., Hernandez, O., Zhang, Y., Savall, J., Chrapkiewicz, R., Li, J., Zeng, H., Ganguli, S., and Schnitzer, M. J. (2020). Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105. pages 11, 90

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, 13(9):e1005768. pages 16, 85

Sadeh, S. and Clopath, C. (2021). Inhibitory stabilization and cortical computation. *Nature Reviews Neuroscience*, 22(1):21–37. pages 5

Samaha, J., Iemi, L., Haegens, S., and Busch, N. A. (2020). Spontaneous brain oscillations and perceptual decision-making. *Trends in cognitive sciences*, 24(8):639–653. pages 82

Sanzeni, A., Akitake, B., Goldbach, H. C., Leedy, C. E., Brunel, N., and Histed, M. H. (2020). Inhibition stabilization is a widespread property of cortical networks. *Elife*, 9:e54875. pages 5

Schneidman, E., Berry II, M. J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007 – 1012. pages 1

Semedo, J. D., Zandvakili, A., Machens, C. K., Byron, M. Y., and Kohn, A. (2019). Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259. pages 77, 89

Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron*, 103(5):934–947. pages 17

Stephani, T., Hodapp, A., Jamshidi Idaji, M., Villringer, A., and Nikulin, V. V. (2021). Neural excitability and sensory input determine intensity perception with opposing directions in initial cortical responses. *Elife*, 10:e67838. pages 82

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365. pages 14, 88

Strogatz, S. H. (2015). *Nonlinear dynamics and chaos, with applications to physics, biology, chemistry, and engineering*. Westview press, 2nd edition. pages 91

Strutz, T. (2011). *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*, volume 1. Springer. pages 37

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press, 2nd edition. pages 89

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189–208. pages 15

Umakantha, A., Morina, R., Cowley, B. R., Snyder, A. C., Smith, M. A., and Byron, M. Y. (2021). Bridging neuronal correlations and dimensionality reduction. *Neuron*, 109(17):2740–2754. pages 10, 11, 81, 82, 88

Urai, A. E., Doiron, B., Leifer, A. M., and Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, pages 1–9. pages 88

Van der Plas, T. L., Manohar†, S. G., and Vogels†, T. P. (2022). Predictive learning enables neural networks to learn complex working memory tasks. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pages 518–531. PMLR. pages vi, vi, vi

Van der Plas*, T. L., Tubiana*, J., Le Goc, G., Migault, G., Kunst, M., Baier, H., Bormuth†, V., Englitz†, B., and Debrégeas†, G. (2023). Neural assemblies uncovered by generative modeling explain whole-brain activity statistics and reflect structural connectivity. *Elife*, 12:e83139. pages 13, 91

Van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726. pages 5

Van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., and Roelfsema, P. R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, 360(6388):537–542. pages 77, 81, 82, 91

Vogels, T. P., Rajan, K., and Abbott, L. F. (2005). Neural network dynamics. *Annu. Rev. Neurosci.*, 28:357–376. pages 5

Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062):1569–1573. pages 5

Vyas, S., Golub, M. D., Sussillo, D., and Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275. pages 17

Wan, Q., Cai, Y., Samaha, J., and Postle, B. R. (2020). Tracking stimulus representation across a 2-back visual working memory task. *Royal Society open science*, 7(8):190228. pages 84

Waschke, L., Kloosterman, N. A., Obleser, J., and Garrett, D. D. (2021). Behavior needs neural variability. *Neuron*, 109(5):751–766. pages 1, 9, 10, 40, 81, 91

Waschke, L., Tune, S., and Obleser, J. (2019). Local cortical desynchronization and pupil-linked arousal differentially shape brain states for optimal sensory performance. *Elife*, 8:e51501. pages 2, 5, 9, 12, 13, 15, 82, 83, 91

Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., Rae, J., Mirowski, P., Leibo, J. Z., Santoro, A., et al. (2018). Unsupervised predictive memory in a goal-directed agent. *arXiv preprint arXiv:1803.10760*. pages 16, 84, 85

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40. pages 85

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2020). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263. pages 16

Williamson, R. C., Cowley, B. R., Litwin-Kumar, A., Doiron, B., Kohn, A., Smith, M. A., and Yu, B. M. (2016). Scaling properties of dimensionality reduction for neural populations and network models. *PLoS computational biology*, 12(12):e1005141. pages 90

Woolsey, T. A. and Van der Loos, H. (1970). The structural organization of layer iv in the somatosensory region (si) of mouse cerebral cortex: the description of a cortical field composed of discrete cytoarchitectonic units. *Brain research*, 17(2):205–242. pages 3

Yamashita, T. and Petersen, C. C. (2016). Target-specific membrane potential dynamics of neocortical projection neurons during goal-directed behavior. *Elife*, 5:e15798. pages 18, 77

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature neuroscience*, 22(2):297–306. pages 17, 69, 84, 85

Yang, H., Kwon, S. E., Severson, K. S., and O'connor, D. H. (2016). Origins of choice-related activity in mouse somatosensory cortex. *Nature neuroscience*, 19(1):127. pages 4, 77

Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O'Shea, D. J., Sohal, V. S., Goshen, I., Finkelstein, J., Paz, J. T., et al. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature*, 477(7363):171–178. pages 5, 19, 34, 57

Yu, J. and Jiang, J. (2016). Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *EMNLP 2016: Proceedings of the Conference on Empirical Methods in Natural Language Processing: Austin, Texas, November 1-5*, pages 236–246. Association for Computational Linguistics. pages 60

Zhang, Y., Rózsa, M., Liang, Y., Bushey, D., Wei, Z., Zheng, J., Reep, D., Broussard, G. J., Tsang, A., Tsegaye, G., et al. (2023). Fast and sensitive gcamp calcium indicators for imaging neural populations. *Nature*, 615(7954):884–891. pages 78

Zhou, J., Jia, C., Montesinos-Cartagena, M., Gardner, M. P., Zong, W., and Schoenbaum, G. (2021). Evolving schema representations in orbitofrontal ensembles during learning. *Nature*, 590(7847):606–611. pages 84