# Testing Theories of Computation and Learning in the Visual Cortex


**Sarah L. Armstrong**

Department of Experimental Psychology

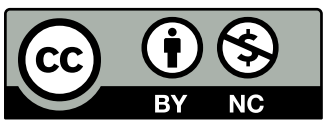
St Edmund Hall

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Trinity 2024




Preferred citation: Armstrong, SL. *Testing Theories of Computation and Learning in the Visual Cortex*. (2024) D.Phil. thesis. University of Oxford.

# Abstract

Modern systems neuroscience provides experimental access to the dynamics of neural systems over different timescales. Many studies of sensory systems in the brain do not exploit this and have instead focused on on static features measured from stimulus-evoked responses. In this thesis, we show that observation of dynamics – changes in neural variables over time – can offer a means to distinguish theoretical models relating to key unsolved questions in neuroscience. We apply this approach to two types of dynamical systems models. In the first study, we analyse the dynamics of firing rates in model recurrent networks. We offer routes to experimentally testing particular implementations of theories of sensory processing such as efficient and predictive coding. In the second study, we propose that learning dynamics – the evolving patterns of neural changes across time compared within and across areas – can serve as a crucial hallmark to test theories of learning in the brain. We devise a multi-stage behaviour and neuronal imaging paradigm in mice to experimentally test key predictions of the deep learning theory of perceptual learning and ultimately find that several key predictions were not met. Taken together, the work in this thesis serves as a demonstration of how experiment and theory in neuroscience can be combined with careful consideration. Through testing a popular theory, it also highlights key challenges in this approach.

# Acknowledgements

Adam, your anything-is-possible attitude, ability to make things happen, quick-thinking, data-hawking and pragmatism have driven this work forward and shaped me into an independent researcher. This turned out to be useful a little sooner than expected. I'm grateful that you took this project on, and supported me in building each facet from the ground-up in your lab despite the many cutting-edge, borderline sci-fi projects happening in every corner. Your motivational "don't think, just do!" post-it is probably the only reason there are any results, and will be stuck to my monitor indefinitely.

Andrew, you took me in for a master's rotation as a lowly biologist, shared my excitement with testing theories, and nurtured my learning until ill-defined ideas on occasion became actual maths. I'll always be grateful for this generosity, and for your belief in me. Going from postdoc to full professor during my tenure as a student is no mean feat, and also serves as a major sign that I need to graduate. I'm excited to see the many secrets of learning be unlocked as your lab continues to flourish.

Jimmy and Rob, working alongside you both was a true delight, and I hold many great memories of that golden area of the Packer lab. You both taught me an incredible amount, punctuated by balcony beers, ranging from imaging, brain surgery and python to the nuances of optimal toastie cooking. Only with your company could a week spent in the dark basement of a midwest microscope factory have been fun. I'm grateful too for the fine company of the Sherrington penthouse over the years - Sheltron, Thijs, Gabi, Huriye, David, Prajay, Caitlin, Antara, Sam, Chiara, Ivan and Blake. Rodrigo, it's been a joy to collaborate with someone so brilliant this past year, and your invaluable support has kept me just about sane with the exception of Cosyne.

Last but not least, thank you to my family for putting up with me, and for your love and support. To George, thanks for waiting for me, for the dishev desk-rescues and late nite stubo jams. To my mum especially, for your dedication to our family, your unconditional care, and the cups of tea that mysteriously appear during hyper-focussed coding sessions.

# Author contributions

I declare that this thesis is my own and that attribution has been given where information was derived from other sources. I implemented experimental hardware and protocols, performed all experimental procedures, data collection, and established data pre-processing, analysis and visualisation. This was achieved with supervision from **Adam Packer** and **Andrew Saxe**, who together helped conceptualise both studies included in the thesis. Behavioural data analysis in 3 has relied on key contributions from **Rodrigo Carrasco Davis**.

# Contents

# List of Selected Abbreviations

| | |
|---|---|
| $\Delta F/F$ | Normalised fluorescence activity (from the calcium reporter). |
| **1D, 2D, ..** | 1-dimensional, 2-dimensional, .. |
| **FOV** | Field of view. |
| **ITI** | Inter-trial interval. |
| **OD** | Orientation discrimination. |
| **CW** | Clockwise. |
| **CCW** | Counter clockwise. |
| **OTC** | Orientation tuning curve. |
| **FWHM** | Full width at half maximum. |
| **GPR** | Gaussian process regression. |
| **NP** | Non-performing. |

# 1 Introduction

## 1.1 The Dynamic Brain

The environment is fundamentally dynamic, and brains have evolved as a control center to flexibly adapt animal behaviour to each moment in time. Perception, cognition and behaviour arise from complex interactions between components of the nervous system, integrating signals from the body and environment. Specialised sensory systems are driven by continuous external inputs, whilst motor systems drive goal-directed behaviour in a constantly changing world. Through learning, neural circuits are refined by experience throughout life (Whitlock *et al.* 2006; Holtmaat & Svoboda 2009; May 2011).

Dynamical systems theory provides a mathematical framework for understanding temporal evolution, or change, in physical and abstract systems (Katok *et al.* 1995). This approach employs differential equations to model how system variables change over time, enabling the description, analysis, and prediction of both linear and nonlinear dynamics (Katok *et al.* 1995; Strogatz 2024). A key strength of this framework lies in its ability to reduce high-dimensional complexity through state-space representations that facilitate interpretation of system behavior (Sorzano *et al.* 2014; Cunningham & Yu 2014). A canonical example of this is the motion of a

pendulum. While measured in a three-dimensional space of space and time, the motion of a simple pendulum can be completely characterised in a two-dimensional state space defined by angular position and velocity.

Beyond descriptive modelling, dynamical systems theory offers powerful tools for system identification - the process of inferring internal structure from observable behaviour (Kalman 1962; Ho 1966; Nise 2019). In linear systems, linear functions map internal system states and inputs to outputs. Here, impulse response analysis reveals how perturbations propagate through the system, exposing underlying connectivity and computational principles that remain hidden in static observations (Kailath 1980). For nonlinear systems, specialized approaches are necessary, including analysis of fixed points, limit cycles, and qualitative behavioral transitions such as bifurcations and chaos (Strogatz 2024).

The brain can be studied through the lens of dynamical systems as an investigative framework, and is considered by some to be a dynamical system (Van Gelder 1995; Favela 2021; Shine *et al.* 2019). This perspective treats the brain as a system where cognition and behaviour are instantiated through the continuous evolution of neural states rather than discrete symbolic operations. Dynamical systems tools have been successfully applied in neuroscience since the pioneering work of Hodgkin and Huxley (1952) in studying the membrane potential dynamics of a neuron. In recent decades, the framework has proven valuable in describing how populations of interacting neurons relate to behaviour (Favela 2021; Churchland *et al.* 2012; Shenoy *et al.* 2013; Vyas *et al.* 2020; Cunningham & Yu 2014; Sussillo 2014; Breakspear 2017).

Because of the brain's immense complexity, models are applied across multiple spatial and temporal scales depending on the target phenomenon. Across scale transitions, different details are abstracted away to make models both analytically and computationally tractable plus conceptually interpretable. For example, in systems neuroscience, the detailed biophysical mechanisms of action potentials - such as the specific ionic currents underlying spike generation - are often abstracted away in favour of higher-level models that describe neuronal activity in terms of firing rates (Dayan & Abbott 2005). Similarly, the complex structural and molecular machinery of synapses - including their precise dendritic locations, spine morphology, neurotransmitter

release dynamics, receptor kinetics, and postsynaptic integration - is frequently reduced to scalar weights that capture synaptic strength. This is the foundation of connectionist models in neuroscience and psychology (J. A. Feldman & Ballard 1982; Rumelhart, McClelland, *et al.* 1986; McClelland 1988; Doerig *et al.* 2023). These abstractions allow researchers to focus on how populations of neurons interact and influence each other's dynamics, rather than on the precise ionic, molecular, or structural mechanisms of individual spikes or synapses. Importantly, similar dynamical systems tools can be applied across different scales of organisation and to different aspects of neural function, whether the focus is on how neural activity evolves during behaviour or how synaptic connections change during learning.

The Hodgkin-Huxley model of action potentials serves as a foundational example in neuroscience (Hodgkin & Huxley 1952). At a molecular level of scale, the Hodgkin-Huxley model describes membrane potential dynamics, where state variables are ion channel conductances in a cell membrane that change with time and voltage. Here, dynamics are governed by a four-dimensional system of differential equations, with state variables representing membrane voltage and the gating variables that control sodium and potassium conductances. Hodgkin and Huxley's contribution was to demonstrate that the complex temporal dynamics of action potentials emerge from relatively simple rules governing voltage-dependent ionic conductances, providing the first mechanistic account of how macroscopic neuronal excitability arises from underlying biophysical processes. While the four-dimensional phase space of this model is challenging to interpret, the FitzHugh-Nagumo model was developed which simplified it to a two-dimensional system of excitability and refractoriness (FitzHugh 1961; Nagumo *et al.* 1962; Izhikevich 2007). This simpler model still captures the essential dynamics of neuronal excitation and, by being fully visualisable in two dimensions, revealed nonlinearities and qualitative behaviours such as thresholds and oscillations that were not as apparent in the higher-dimensional representation. This highlights how simplifying complex systems can reveal essential dynamics. Dimensionality reduction involves identifying the most important variables or combinations of variables that capture a system's essential behaviour while discarding less critical dimensions that may obscure interpretation. This approach has become valuable in understanding complex neural systems by revealing underlying structure and dynamics (Cunningham & Yu 2014).

Moving to a circuit or network level of scale, the state variables in populations of neurons are often neuronal firing rates, where dynamics are determined by synaptic connectivity patterns combined with transfer functions that translate input to output. A successful example of this approach is found in the study of cortical control of arm movements by Churchland and colleagues (2012). This study showed that single-neuron responses recorded in motor cortex during reaching are complex and do not map cleanly onto specific movement variables. However, when analysed at the population level, neural activity was described by low-dimensional, dynamical trajectories. This work demonstrates how dynamical systems approaches can reveal structured population dynamics that are hidden at the single-neuron level, supporting the argument that neural function is best understood through temporal evolution of network states (Shenoy *et al.* 2013).

Neural circuit function is constrained by structure (Honey *et al.* 2010; Gal *et al.* 2017; Real *et al.* 2017; Lappalainen *et al.* 2024). This means that neural activity patterns cannot be explained without a description of connectivity patterns, or how neurons connect to each other. At a 'macro' level of scale, neuroscientists have divided brains into areas linked to particular cognitive functions or behaviours (Verma & Kumar 2022; Kandel 2021). Each area receives inputs from and projects outputs to different brain areas, whilst some areas have recurrent or "self-to-self" internal connectivity (Harris & Mrsic-Flogel 2013). This macro-level organisation and connectivity structure is consistent across brains of a given species, having been established over evolutionary timescales (Assaf *et al.* 2020; Mars *et al.* 2016; Kaas 2020). Each brain area incorporates cell types with distinct functional characteristics, believed to further adapt circuits to their particular functions (Luo *et al.* 2018; Xing *et al.* 2023).

Macro-structural features can be thought of as analogous to the architecture of artificial neural network models, where structure demonstrably constrains function. The brain's structural architecture provides constraints and possibilities for the functions it can implement or learn, much like how a neural network's architecture limits the functions it can learn (Jacot *et al.* 2018) or implement (Montúfar *et al.* 2014; Shao & Shen 2023; Suárez *et al.* 2021; Kriegeskorte & Golan 2019). The modular and hierarchical organisation of the brain, with specialised regions and structured connectivity, shares some architectural principles with artificial neural networks - particularly the

multi-layered processing found in deep networks. The remarkable success of deep learning in engineering applications such as computer vision (Voulodimos *et al.* 2018) has led researchers to draw parallels with hierarchical processing in the brain, inspiring the adoption of multi-layered network models in neuroscience where different brain areas are conceptualised as different processing layers (Yamins *et al.* 2014; B. A. Richards *et al.* 2019). Some neuroscientists now use neural network models specifically to test hypotheses about brain function, exploring how architectural constraints shape computational capabilities in both biological and artificial systems (B. A. Richards *et al.* 2019; Kriegeskorte & Golan 2019; Suárez *et al.* 2021; Doerig *et al.* 2023).

Life experience is believed to shape neuronal connectivity at a finer level of scale. This refers to the details of which neuron connects to which through synapses, and the degree of functional influence it has or the strength of a synapse. This is believed to be achieved through various forms of neural plasticity, which is the capacity of neural circuits to modify their structure and function in response to activity and experience (Whitlock *et al.* 2006; D. E. Feldman 2009; Humeau & Choquet 2019). Synaptic plasticity encompasses changes in synaptic strength across multiple timescales, such as long-term potentiation (LTP), long-term depression (LTD) and short term facilitation, as well as structural modifications including the formation and elimination of synaptic connections (Citri & Malenka 2008; D. E. Feldman 2009). In artificial neural network models, plasticity is analogous to changes in weight matrices - the main parameters that define connectivity patterns between units and the strength of their functional interactions. These weights are initialised and subsequently modified through learning algorithms.

Experience-dependent plasticity during brain development is believed to be crucial for establishing the functional architecture of neural circuits, with synaptic connections being refined and stabilised based on patterns of neural activity driven by sensory experience and environmental statistics (G. Yang *et al.* 2009; Holtmaat & Svoboda 2009; Zhang & Poo 2001). Environmental statistics refer to the regularities and patterns present in sensory input - such as the spatial and temporal correlations in visual scenes, the frequency distributions of sounds, or the co-occurrence patterns of different sensory features (Simoncelli & Olshausen 2001; Lewicki 2002). During development, neural circuits are thought to adapt to these statistical properties through

activity-dependent mechanisms that can involve synaptic strengthening, new synapse formation, and synaptic elimination (Katz & Shatz 1996; Zhang & Poo 2001; Isaac *et al.* 1997). Furthermore, critical periods represent transient developmental windows during which sensory cortical areas exhibit heightened plasticity, with ocular dominance being a well-established example of experience-dependent reorganization (Wiesel & Hubel 1963; Hensch 2004). The relative contributions of genetic inheritance versus experience and activity-dependent plasticity represent an active area of research, with contemporary approaches emphasising probabilistic interactions between genetic, neural, and environmental factors (Stiles 2011).

Whilst development features critical windows that have long-lasting effects on function, activity-dependent processes continue to operate throughout life, enabling ongoing learning and adaptation in mature neural circuits (May 2011; Kerr *et al.* 2011; Holtmaat & Svoboda 2009). The ability to learn novel tasks, adapt to new environments, and recover from injuries throughout life is a remarkable property of animals. All of these processes can be encapsulated by the term learning. Learning likely involves multiple different processes and mechanisms that, while often studied separately, work together as an integrated system. The prevailing view in neuroscience is that synaptic connectivity - both the pattern and strength of connections - constitutes the primary substrate through which plasticity enables learning. While synaptic weights and connectivity patterns cannot easily be directly observed in the living brain, learning manifests as measurable changes in neuronal tuning properties - such as shifts in receptive field structure, stimulus selectivity, and response magnitude - which provide observable signatures of the underlying plasticity processes (A. Schoups *et al.* 2001; Khan *et al.* 2018; Recanzone *et al.* 1993; T. Yang & Maunsell 2004; Yan *et al.* 2014).

In efforts to quantify and formalise descriptions of neural processes, theoretical neuroscientists have attempted to describe neural dynamics, synaptic plasticity, and learning through normative principles and algorithmic rules (Dayan & Abbott 2005). For example, researchers seek objective functions that neural firing rate dynamics might optimise, such as efficient coding principles that minimise metabolic cost while maximising information transmission (Barlow 1961), or sparse coding that activates only a small fraction of neurons for any given input (Olshausen

& Field 1996). At the synaptic level, synaptic plasticity rules such as those relating to spike-timing dependent plasticity (STDP) describe how synaptic strength changes based on the precise temporal relationship between pre- and post-synaptic activity (Suvrathan 2019). In the domain of reward-based learning, theoretical frameworks attempt to link neuromodulatory signals like dopamine to learning rules, where dopamine is hypothesised to signal reward prediction errors that drive synaptic updates according to reinforcement learning algorithms (Schultz *et al.* 1997; Glimcher 2011; Liebana *et al.* 2025). In artificial neural networks, or connectionist models of the brain, differentiable objective functions are defined explicitly based on the goal of a task, and update rules for weights can be derived directly from the function in order to optimise task performance (Rumelhart, McClelland, *et al.* 1986). This type of optimisation is achieved through gradient descent, and underpins modern artificial intelligence and machine learning (LeCun, Bengio & G. Hinton 2015). This has inspired the framing of learning in brains in similar terms of learning rules and optimisation, naturally leading to the quest to uncover the brain's "learning rules" (Marblestone *et al.* 2016).

This raises a fundamental question: what would it mean for the brain to 'implement' a learning rule? One interpretation would require that learning rules exist as explicit symbolic structures that could be located and decoded somewhere in the brain, acting as high-level control programs that reconfigure neural function. However, this symbolicist perspective faces significant theoretical and empirical challenges (Van Gelder 1995; McClelland 1988). An alternative view is that rule-like behaviour emerges implicitly from the collective dynamics of neural circuits rather than being explicitly represented. Through evolution, neural circuits may have developed dynamics that cause connectivity to change in response to experience in ways that might be characterised by abstract learning rules that may ultimately relate to normative principles.

Uncovering such descriptive rules remains valuable for understanding learning mechanisms and developing theoretical frameworks, even if the rules themselves are not explicitly "represented" by the brain. This perspective allows us to distinguish between the algorithmic description of learning rules and their biological implementation - that is, between a mathematical characterisation of how learning proceeds and the specific cellular and molecular processes through which

it occurs (Marr 1982). Studying implementational details alone may not sufficiently constrain our understanding learning at an algorithmic level (Hennig 2023), suggesting that both levels of description are necessary for a complete account of neural learning.

In this thesis, I investigate theories of the systems underlying sensation and perception in the brain. Although neural circuit dynamics and learning are interdependent and operate simultaneously in the brain, they are often studied independently and have been the subject of distinct theoretical frameworks. In the first project, I examine theories of neural circuit dynamics underlying stimulus responses in sensory cortex, treating learning as effectively 'frozen' to isolate the dynamics of mature circuits. To model these phenomena, I analyse recurrent neural networks to generate predictions for empirical measurements recorded from neuronal populations in vivo. I explore whether targeted perturbations of neural activity could potentially test theories of sensory circuits by observing how the system responds to and recovers from these manipulations. In the second project, I investigate how neuronal responses to visual stimuli change through perceptual learning, with experiments designed to test predictions derived from deep gradient descent learning in artificial neural networks. While the first project provides theory and simulations, the second project additionally incorporates experiments carefully designed to test theoretical predictions. Both projects are united by their application of dynamical systems tools and their focus on investigating theories underlying stimulus responses during sensory stimulation and over the course of learning. In addition, both projects are driven by the goal of conducting theory-inspired experiments that can constrain theoretical models through testable predictions. This theory-first approach addresses several key challenges in systems and computational neuroscience. First, high-dimensional neuronal data can in theory make finding a post-hoc explanatory model trivial. Models with many free parameters can fit a wide range of data, even if they are not truly explanatory, meaning model fit alone is not a good measure of merit (Roberts & Pashler 2000). Furthermore, many theories are formulated at levels of abstraction that make it difficult to derive specific, testable predictions, and without such predictions, theories cannot be empirically falsified (Popper 1963). To address these issues, both projects begin with theoretical frameworks that make quantitative predictions for specific experiments, then design experiments to test them.

In the remainder of this introductory chapter, I introduce the background for each project individually.

## 1.2 Structure and function of neural circuits in sensory cortex

The cerebral cortex contains specialised areas devoted to processing each sensory modality. A dominant view in neuroscience holds that neurons within these areas extract meaningful features from sensory input, transforming raw sensory data into neural representations that can guide decisions and behaviour. According to this framework, perception emerges through hierarchical, distributed processing, where simple features detected at early stages are progressively combined into increasingly complex representations - from basic elements like edges and textures to complete objects with semantic meaning (Felleman & Van Essen 1991). This computational hierarchy is thought to map onto the brain's anatomical organisation, with different cortical areas implementing distinct processing stages as sensory information propagates from peripheral receptors through successive levels of cortical processing. Patterns of neural activity associated with particular stimuli or stimulus features are often termed neural codes (Harris & Mrsic-Flogel 2013). Understanding how these stimulus responses emerge from the interplay between sensory input, cortical cell types, synaptic connectivity, and neuronal integration has become a central goal in systems neuroscience. This mechanistic approach promises to bridge the gap between neural mechanisms and perceptual experience, while sensory systems provide an experimentally tractable domain where controlled stimuli can be systematically related to neural circuit properties. Success in deciphering sensory coding could provide fundamental principles for understanding more complex cognitive functions, and offer insights into how neural circuits become disrupted in neurological and psychiatric disorders.

Cortical neurons are broadly classified into two fundamental classes based on their neurotransmitter release: excitatory cells that release glutamate and inhibitory cells that release GABA. Excitatory neurons comprise approximately 80% of cortical neurons, while inhibitory neurons, often called interneurons, make up the remaining 20% (Harris & Mrsic-Flogel 2013). Within these

broad categories, there is diversity: neurons can be further subdivided into numerous subtypes based on their genetic markers, physiological properties, and connectivity patterns. Excitatory cortical neurons comprise primarily pyramidal cells and spiny stellate cells, named for their distinctive morphologies, whilst inhibitory neurons are commonly classified by their molecular markers into parvalbumin (PV), somatostatin (SOM), and vasoactive intestinal peptide (VIP) cells (Khan *et al.* 2018). Different neuronal subtypes are hypothesised to contribute distinct computational roles, which remains an active area of investigation (Harris & Shepherd 2015).

Connectivity between neurons is typically categorised by the anatomical relationship between the presynaptic and postsynaptic cells. Excitatory neurons form both long-range connections that project to different cortical areas and short-range connections within the same area or cortical layer, whilst inhibitory interneurons primarily form local connections within their vicinity (Harris & Mrsic-Flogel 2013). When connections between neurons contribute to self-excitation within a local circuit - whether through direct pyramidal-to-pyramidal connections or indirect pathways such as excitatory-inhibitory-excitatory loops - this constitutes recurrent connectivity (Nó 1933). This recurrent architecture is fundamental to neural population dynamics, as it enables sustained activity, amplification of signals, and complex temporal patterns that underlie sensory processing (Rubin *et al.* 2015; Hennequin *et al.* 2012; Carandini & Heeger 2012; Oldenburg *et al.* 2024). Understanding how this structured recurrent connectivity shapes neural population responses to sensory stimuli, and how it can be experimentally probed, is therefore central to understanding cortical computation (Mastrogiuseppe & Ostojic 2018; Duncker *et al.* 2020).

Extensive research into cortical circuit structure has revealed important principles governing how connectivity patterns shape neural dynamics and stimulus responses. Rather than being randomly organised, recurrent connectivity in sensory cortex exhibits highly structured patterns that directly influence functional properties. In layer 2/3 primary sensory cortex of mice, experiments have shown that physical distance predicts connection probability between neurons (Levy & Reyes 2012; Campagnola *et al.* 2022). More importantly, neurons that respond to similar stimulus features, such as orientation, are significantly more likely to be connected than those with dissimilar tuning (Ko *et al.* 2011; Cossell *et al.* 2015; Rossi *et al.* 2020). This principle

extends beyond excitatory connections: PV-expressing interneurons connect broadly to nearby pyramidal cells but with connection strengths structured according to functional similarity, creating organised subnetworks (Znamenskiy *et al.* 2024).

These structured connectivity patterns have direct consequences for neural population dynamics. Connected neurons exhibit higher activity correlations than unconnected pairs, demonstrating how synaptic organisation constrains which neuronal ensembles can be co-active (Ko *et al.* 2011; Cossell *et al.* 2015). From a theoretical perspective, the precise balance between excitatory and inhibitory connections is thought to play a crucial role in stabilising recurrent dynamics, preventing runaway excitation that would otherwise destabilise network activity (Rubin *et al.* 2015). Different cortical layers exemplify these structure-function principles. Neocortex typically has 6 layers (Kandel 2021). Layer 2/3 excitatory cells receive strong, non-selective inhibition and functionally specific excitation, resulting in sparse, selective responses, whilst layer 5 projection neurons integrate diverse excitatory inputs with weaker inhibition, producing broader, more sustained responses.

Neocortex exhibits a laminar organisation with distinct patterns of inter-areal and inter-laminar connectivity. This has led to the conception of an organisational principle called the 'canonical microcircuit', which proposes that a fundamental anatomical motif is repeated throughout neocortex with minor variations as a common strategy for information processing (Douglas *et al.* 1989; Bastos *et al.* 2012; Miller 2016; Harris & Shepherd 2015). This model suggests that the circuit comprises superficial (L2/3) and deep (L5/6) pyramidal cells, excitatory cells concentrated in layer 4, and a shared pool of inhibitory interneurons, all interconnected. The typical information flow initiates with thalamic inputs primarily targeting L4, then ascends via feedforward interlaminar connections from L4 to L2/3, and subsequently from L2/3 to L5 and L6, with descending feedback projections from L5/6 back to L4. According to this view, intracortical excitatory recurrent connections provide the majority of excitation - far exceeding initial thalamic inputs - with this amplification regulated by robust inhibition to maintain balance and prevent runaway activity (Miller 2016). The canonical microcircuit is thought to perform computations such as the feedforward establishment of neuronal selectivity and recurrent gain control, where

intracortical input largely cancels strong external input, explaining nonlinearities like surround suppression and normalisation (Miller 2016). While the basic connectivity motifs appear to generalise across cortical areas, there are often variations in structural and organisational details across different mammalian species and cortical regions (Miller 2016; Harris & Shepherd 2015; Scala *et al.* 2019). Despite these variations, the canonical microcircuit has become an influential organising principle for understanding cortical structure-function relationships.

### 1.2.1 Theories of sensory cortex

In Marr's hierarchical framework for analysis, understanding of information processing systems is organised into three levels: the computational level asks what problem needs to be solved, the algorithmic level asks what computational strategy is used to solve this problem, and the implementational level concerns how these strategies are physically realised in neural circuits (Marr 1982). The circuit details discussed above - cell types, connectivity patterns, and structure-function relationships - represent primarily implementational level understanding, as they describe the neural substrate that supports computation. We now turn to computational and algorithmic theories of sensory cortex, which address fundamental questions about what perceptual problems sensory systems need to solve, why particular computational strategies might be optimal, and how these strategies could be algorithmically implemented in neural circuits. These theories typically integrate normative and algorithmic perspectives, proposing that neural circuits implement optimal solutions to computational problems through specific mechanisms. Horace Barlow famously began a 1961 book chapter by noting that 'A wing would be a most mystifying structure if one did not know that birds flew' (Barlow 1961). This highlights the importance of understanding structure in the context of function. Extending this to sensory neural circuits, computational understanding may not emerge directly from connectivity details alone, suggesting that tailored experimental approaches are needed to distinguish between competing theoretical frameworks.

Barlow went on to propose one of the foundational theories in this domain - efficient coding - which proposes the normative principle that neural responses should maximise information

transmission while minimising redundancy and metabolic cost (Barlow 1961). Indeed, sensory information from natural environments contains high levels of redundancy and correlation, making it highly predictable and enabling significant compression (Simoncelli & Olshausen 2001; Simoncelli 2003; Ruderman 1994). Inspired by electrophysiological studies of retinal responses, Barlow proposed that sensory information could be economised through reducing the average frequency of impulses used to convey messages, allocating more resources to uncommon (i.e., unpredictable) sensory stimuli. Redundancy and predictability are two sides of the same coin - redundant information is by definition predictable, since knowing one part of a pattern allows inference of the remaining parts. Beyond reducing energy usage, Barlow proposed that efficient coding is a key organisational principle of the nervous system. By arranging neural messages 'according to their prior probabilities', the brain could build an internal model of its environment and focus resources on unpredictable sensory information to guide behaviour, rather than processing all inputs equally. This proposal provides key algorithmic concepts - prior probabilities, internal models, and prediction-based resource allocation - that relate to the broader framework of perception as inference, which will be discussed below. Barlow also made what he described as the "monstrous suggestion" that sensory information might ultimately be represented by only a few neurons at the highest level of processing - this introduced the notion of hierarchical representations. Overall, Barlow's 1961 paper anticipated several influential ideas that remain central to theoretical neuroscience today: efficient coding, hierarchical representations, and the notion that the brain maintains internal models based on environmental statistics.

Several researchers have extended and formalised efficient coding principles, producing more sophisticated theories involving constraints such as noise, variance metabolic costs (Simoncelli 2003). Efficient coding theory makes several predictions about neural response properties. First, neural responses should relate to the degree of unpredictability or 'surprise' of sensory stimuli - neurons should respond more strongly to unexpected events than to predictable ones. Second, these responses should adapt dynamically according to the statistical regularities of the environment and the history of recent sensory inputs, meaning that neural tuning should shift based on context and experience (Fairhall *et al.* 2001). Third, rather than simply relaying raw

sensory information, neurons should represent complex features that capture relationships and correlations between different input channels, reflecting efficient encoding of natural stimulus statistics (Baddeley *et al.* 1997; Vinje & Gallant 2000; Rieke *et al.* 1997). Analysis of firing rate distributions has offered empirical support for these ideas. For example, exponential firing rate distributions observed under naturalistic images in V1 and IT are consistent with efficient coding predictions for sparse, decorrelated neural responses (Baddeley *et al.* 1997). However, some studies found different firing rate distributions that initially appeared inconsistent with efficient coding (Treves *et al.* 1999). De Polavieja and colleagues (2002) addressed this discrepancy by developing a model that incorporated noise constraints, demonstrating that varying noise conditions can account for the observed diversity in firing rate distributions while maintaining consistency with efficient coding principles.

Olshausen and Field (1996) proposed that sparse coding - a strategy where only a small fraction of neurons are active at any time - provides an algorithmic implementation supporting efficient coding principles. Beyond efficiency, sparse coding offers several computational advantages: it improves memory storage by minimising pattern interference, reveals underlying structure in natural stimuli, facilitates downstream computation by reducing representational complexity, and decreases energy consumption through low population activity levels (Olshausen & Field 1996). Crucially, they demonstrated that these advantages emerge naturally from the statistics of natural images. When they trained a learning algorithm to optimise for sparse representations of natural image patches, the resulting receptive fields spontaneously developed the three key properties observed in primary visual cortex: spatial localisation, orientation selectivity, and bandpass filtering. This demonstrated how sparse coding as a representational strategy can account for detailed neural response properties while achieving computational efficiency.

At a broader computational level, perception has been conceptualised as Bayesian inference, where the brain solves the fundamental problem of interpreting uncertain and incomplete sensory information by maintaining probabilistic models of the world (Knill & W. Richards 1996; Aitchison & Lengyel 2017). This framework posits that perception fundamentally involves inferring the latent causes of sensory observations, using Bayes' rule to combine prior

expectations with incoming sensory evidence to compute posterior beliefs about the state of the world. These ideas can be traced as far back as Helmholtz's concept of unconscious inference (Helmholtz 1867), and have been formalised through modern probabilistic frameworks (Kersten *et al.* 2004; Knill & Pouget 2004). Evidence for Bayesian perception comes from multiple sources: visual illusions and psychophysical experiments demonstrate that perception is influenced by predictions and expectations (Lange *et al.* 2018). For example, studies have shown that learned stimulus associations can bias the perception of ambiguous motion (Chalk *et al.* 2010). Neural recordings further reveal that responses are modified by spatial and temporal context - for instance, V1 neurons respond to illusory contours that must be inferred from context rather than being physically present in the input (Grosof *et al.* 1993). Behavioural studies support Bayesian principles in decision-making, cue combination, and motor control (Ernst & Banks 2002; Wolpert *et al.* 1995; Beck *et al.* 2008). Crucially, this computational-level theory does not specify particular neural algorithms or implementations. Various algorithmic approaches could potentially implement Bayesian inference, including probabilistic population codes where firing rates represent parameters of posterior probability distributions (Pouget *et al.* 2013; Ma *et al.* 2006), sampling-based approaches (Orbán *et al.* 2016; Echeveste *et al.* 2020), direct coding of latent variables (Olshausen & Field 1996), or predictive coding mechanisms (Rao & Ballard 1999). This diversity of potential implementations means that empirically disproving one particular algorithmic approach is not likely to invalidate the broader computational framework.

Predictive coding offers a specific algorithmic implementation of Bayesian inference, proposing that the brain implements probabilistic inference through hierarchical prediction and error-correction mechanisms (Rao & Ballard 1999). This framework has gained considerable popularity, perhaps due to its intuitive appeal and theoretical elegance, as reflected in the high citation counts of key articles on the topic (Friston 2005). According to this theory, higher levels in the cortical hierarchy generate predictions about activity at lower levels, while lower levels compute prediction errors - the difference between predicted and actual sensory input. Only these prediction errors are transmitted upward, meaning that predictable components of the input are filtered out at each level. This approach directly links to efficient coding principles, as minimising prediction error is equivalent to removing statistical redundancy (Y. Huang & Rao

2011). What distinguishes predictive coding from other algorithms for Bayesian computations is its specific neural implementation: it proposes that neural responses explicitly represent prediction errors, with distinct populations of neurons encoding predictions versus errors that should be found at different levels of sensory systems. While this hypothesis makes it easier to conceive of experimental paradigms, predictive coding remains controversial due to limited empirical evidence that can distinguish it from other explanations of neural responses.

Several lines of empirical evidence have been interpreted as supporting predictive coding in sensory systems. Firstly, extra-classical receptive field effects in visual cortex, including end-stopping and surround suppression, exhibit response patterns consistent with predictive coding predictions: neural responses are reduced when centre and surround stimuli form coherent structures, potentially reflecting decreased prediction errors (Bolz & Gilbert 1986; Rao & Ballard 1999). Mismatch responses have been documented across multiple sensory modalities. In auditory cortex, oddball paradigms demonstrate enhanced responses to infrequent stimuli, termed 'mismatch negativity' in EEG recordings (Näätänen *et al.* 2007). Sensorimotor mismatch responses provide another line of evidence, studied using virtual reality paradigms where visual flow is coupled to locomotion. In visual cortex, violations of expected visual flow during locomotion produce neural responses that have been interpreted as visuomotor mismatch signals (Keller & Mrsic-Flogel 2018). Similar effects have been observed in auditory cortex, where distinct neuronal populations display activity patterns interpreted as prediction errors when expected auditory consequences of self-generated sounds are violated (Audette & Schneider 2023). Sensorimotor mismatch responses may represent a preprocessing mechanism that stabilises visual input by removing predictable self-motion effects. This phenomenon, while likely essential for sensory processing, does not necessarily address the core computational problems of perception as hierarchical Bayesian inference. However, recent work has proposed extending perception as inference to cross-modal predictions, with Mikulasch et al. (2023) formalising sensorimotor mismatch within a causal inference framework where motor and visual areas cooperate to jointly explain optic flow, generating both positive and negative mismatch responses.

However, alternative mechanisms may account for many sensory responses attributed to predictive coding. For example, divisive normalization may account for extra-classical receptive field effects through local inhibitory interactions that scale responses based on surrounding neural activity, without requiring prediction (Carandini & Heeger 2012). Furthermore, the phenomenon of balanced amplification in non-normal recurrent networks could account for the phenomenon of large transients in neuronal responses after changes in a sensory stimulus (Murphy & Miller 2009). Next, Muzzu and Saleem (2021) demonstrated that mismatch responses in visual cortex may have a simpler explanation: they found that neurons in mouse V1 showed enhanced responses when drifting gratings matching their preferred orientation ceased, with the strongest effects in neurons tuned to low temporal frequencies, suggesting these effects reflect basic feature selectivity rather than prediction error computations. Attentional mechanisms might also contribute to some observed effects: enhanced responses to unexpected stimuli may reflect attentional orienting rather than prediction error signalling (Saalmann *et al.* 2007; Bisley 2011).

Direct variable coding models of bayesian inference offer a further alternative explanation (Aitchison & Lengyel 2017). In these approaches, neural firing rates encode the inferred values of latent variables - hidden causes that generate sensory observations - rather than representing discrepancies between expected and actual inputs. In models of predictive coding where predictions and prediction errors are represented by distinct neuronal populations, the prediction neurons essentially follow a direct coding scheme. However, it is possible to construct a system where only the latent variables, and not the prediction errors, are reflected in neural responses, so that prediction error need not be explicit. Under one implementation of this direct coding framework, enhanced neural responses to unexpected stimuli could reflect increased uncertainty about latent variable values which could be misinterpreted as prediction error signals due to higher mean activity (Aitchison & Lengyel 2017; Orbán *et al.* 2016). This convergence of multiple plausible explanations for the same neural phenomena demonstrates the necessity for experimental approaches capable of generating distinguishing predictions between theoretical frameworks.

The hypothesis of predictive coding has led researchers to propose specific mappings onto cortical microcircuits. Bastos et al. (2012) formalised hierarchical predictive coding as a system that can be implemented through neuronal dynamics performing gradient descent on the sum of square prediction errors, where coupled prediction and prediction error units each integrate inputs according to a differential equation. The implementation uses coupled differential equations where prediction units integrate recurrent population dynamics, error signals from same-level prediction error units, and error signals from higher-level prediction error units, whilst prediction error units integrate bottom-up sensory inputs and top-down predictions to compute residual errors. To map these computations onto cortical anatomy, Bastos and colleagues argued that cortical columns integrate feedforward and feedback inputs that operate at different frequency bands, with feedforward connections conveying high-frequency (gamma) signals and feedback connections using lower-frequency (alpha/beta) oscillations. This spectral segregation corresponds to the laminar organization of cortex, where superficial layers exhibit gamma-band activity and deeper layers show alpha/beta-band activity, reflecting their roles in feedforward and feedback processing respectively.

In this proposed mapping onto the canonical microcircuit, three distinct computational quantities are distributed across different excitatory cell populations within cortical layers. Prediction errors from the previous hierarchical level are conveyed to layer 4 excitatory cells, where sensory afferents canonically converge. These signals are then fed to superficial layer 2/3 prediction units, which include both excitatory interneurons with local connectivity that remain within the cortical column and pyramidal cells that can project beyond the column. Layer 2/3 prediction error units compute the residual error between the current layer's prediction and the previous layer's input, then broadcast these errors through feedforward connectivity to the next highest level in the hierarchy. Simultaneously, the prediction units in L2/3 transmit their predictions to the previous hierarchical level via deeper layer excitatory pyramidal neurons, which possess the feedback connectivity necessary for transmitting top-down signals.

In this mapping onto cortical layers, signals are distributed across multiple cell types, but fundamentally, there are three distinct quantities that are proposed to be represented in excitatory

neurons: the previous layer's prediction error (L4 excitatory cells), the current layer's prediction (L2/3 and deeper layer excitatory neurons), and the residual error between the current layer and the previous layer (L2/3 excitatory pyramidal cells). The predictive coding model equations predict that two types of neuronal dynamics should broadly be observed: transient increases in activity in prediction neurons when predictions fail (layer 2/3), and sustained responses representing continuous activity that relates to prediction neurons' encoding of current beliefs (deeper layers). Broadly consistent with this, stimulating excitatory neuronal populations in cortical layer 5 compared to layer 2/3 was shown to recruit qualitatively different local network dynamics: stimulating L5 excitatory neuronal populations produced sustained network activity dynamics, whilst stimulating L2/3 populations conversely did not effectively recruit local excitation, possibly reflecting increased inhibition (Beltramo *et al.* 2013). This proposed mapping to cell types in cortical layers is particularly relevant for experimental validation, as in vivo systems neuroscience experiments are well-positioned to measure activity in functionally-defined excitatory cells, for example using transgenic mice that express calcium sensors targeted to cortical excitatory neurons. Whilst in vivo cellular imaging approaches are traditionally best suited to recording upper cortical layers 2/3 and 4, recent advances in imaging may further allow deeper layers to be measured (Prevedel *et al.* 2025).

Beyond identifying neural representations of prediction errors and latent variables, a crucial aspect of predictive coding that has received limited experimental attention is its fundamental requirement for recurrent dynamics between dynamically coupled populations of prediction and error neurons, where predictions must be continuously updated through recurrent neural interactions. This dynamical systems perspective offers an alternative approach to empirically testing predictive coding theories by examining the precise quantitative interrelations governing these coupled neural dynamics. The empirical paradigms discussed above typically involve behavioural contexts where animals are actively moving and making decisions whilst processing external sensory stimulation, with analyses correlating firing rates to external behavioural and sensory variables. Such experimental approaches, while informative, are not tightly controlled in the sense that may evoke diverse neural processes that could facilitate spurious correlations between subsets of neural activity and external variables. An alternative approach is to use

targeted neural perturbations to directly probe the putative dynamical system underlying predictive coding. This method enables specific predictions about how different neuronal populations within the hypothesised circuit should respond to controlled manipulations, offering a more rigorous test of whether the system operates according to the dynamical principles proposed by models of predictive coding in neural circuits. This motivates our theoretical work investigating whether targeted neural perturbations can provide the specificity needed to distinguish between theoretical frameworks

### 1.2.2 Targeted neuronal activity perturbations

Techniques for targeted neural perturbations that could be leveraged for testing theories of cortical function have become feasible through advances in optogenetic technology. In 2005, a landmark paper demonstrated that Channelrhodopsin 2 (ChR2) could be used to control mammalian neuronal activity with millisecond precision (Boyden *et al.* 2005). The following two decades have seen optogenetics techniques burst into prominence, applied in thousands of research papers (Deisseroth 2015). This new era has resulted from concurrent advances in molecular genetics, opsin development, and optical methods. The ability to manipulate the activity of genetically defined cell types and projections, during sensation and behaviour, and with fast temporal precision has allowed experimenters to probe their causal role in neural circuits underlying diverse functions from homeostasis to cognition. Beyond establishing mesoscale causality relationships, optogenetics has also been instrumental in mapping functional pathways across brain regions (Kress *et al.* 2013; Hunnicutt *et al.* 2014)

External sources of light must be delivered to brain tissue in order to activate opsin proteins used for optogenetic control of neurons. The lack of spatial precision in traditional light-delivery techniques means that large populations of cells must be manipulated simultaneously in 'bulk', limiting our ability to investigate fine-scale details of circuit dynamics. Whilst powerful for establishing broad causal relationships, inferring precise causal roles of brain areas from bulk manipulations presents significant challenges because brain regions are highly interconnected. Disrupting a large neural population in one area can lead to 'off-target' effects on computation in

downstream areas due to acutely altered levels of drive from input pathways (Otchy *et al.* 2015). While informative, population-level manipulations leave many detailed questions about neural network dynamics, computational mechanisms, and the specific roles of functionally-defined neurons unanswered. This is particularly the case for testing theories that make predictions about the connectivity or response properties of individual cells within populations, such as those with different stimulus tuning properties or distinct functional characteristics. To address these limitations, finer-scale tools are necessary to elucidate the role of individual neurons and small ensembles within larger circuit contexts.

In the last decade, the required optical tools have been developed and combined with 2-photon calcium imaging approaches to allow simultaneous manipulation and read-out of neural activity at single-cell resolution in mouse cortex (Packer *et al.* 2015)). Using a spatial light modulator, individual cells can be targeted for optogenetic stimulation based on their spatial location and functional properties, such as tuning to specific sensory stimuli. Resultant activity in the surrounding neural population can then be measured with high spatial and reasonable temporal resolution, through different cortical layers and even across different brain areas. This 'all-optical' approach represents a significant technological advance, enabling researchers to bridge the gap between population-level manipulations and single-cell precision.

These tools hold particular promise for exploring neural dynamics in ways that may not be accessible through natural sensory stimulation alone. Sensory input arrives at cortical areas via many layers of processing. Targeted perturbations could potentially allow neural dynamics to be pushed 'off-manifold', exploring subspaces of activity that are permitted by local connectivity patterns yet difficult to access by experimentally manipulating an animal's external environment. This capability opens new experimental possibilities for testing theoretical predictions about circuit function and connectivity.

Experiments applying this all-optical approach have explored several fundamental questions in neuroscience. Some studies have focused on behavioural manipulation, asking what is the minimum number of neurons required for behavioural report in visual and somatosensory cortex (Marshel, Kim, *et al.* 2019; Dalgleish *et al.* 2020). Others have used targeted photostimulation

to probe the circuits underlying persistent activity in anterior lateral motor cortex (Daie *et al.* 2019). Further studies have aimed to dissect the functional connectivity of circuits in visual cortex, finding that the effect of perturbing excitatory neurons in layer 2/3 mouse primary visual cortex often causes suppression of other neurons, and can be predicted as a function of physical distance and similarity of visual feature tuning (Chettih & Harvey 2019; Oldenburg *et al.* 2024). More recently, studies have found that the propagation of perturbation-evoked activity depends on both the state of sensory cortex and the behavioural state of mice, offering novel insights into principles of information transmission in sensory cortices (Rowland *et al.* 2023; Russell *et al.* 2024).

Despite these technological advances and initial experimental successes, the field faces theoretical challenges in leveraging these capabilities effectively to build understanding of the neural basis of behaviour. Most centrally, it remains unclear exactly how one would systematically test particular theories of neural computation using activity perturbations. Specifically: how would a neural population respond to targeted perturbations under different theoretical frameworks? Which crucial predictions would distinguish between competing theories? How do these predictions depend on various experimental parameters and design choices?

Some pioneering theoretical studies have begun to address these questions, particularly focusing on how the properties of inhibition-stabilised network models may be inferred using neuronal perturbations (Rubin *et al.* 2015; Sadeh & Clopath 2020; Murphy & Miller 2009; Palmigiano *et al.* 2020; Palmigiano *et al.* 2023). These studies have provided valuable insights into the effects of perturbations under the assumptions of one prominent mechanistic theory of cortical response dynamics, but they leave open the broader question of how perturbations would affect predictions under other theoretical frameworks, such as the algorithmic implementations of bayesian inference such as predictive coding models that have been discussed here.

Furthermore, when modelling of experimental data is conducted post-hoc - building theory from intuitions about data without having first formalised theoretical hypotheses - results can be informative but may not represent the most scientifically rigorous approach. Multiple models could potentially be constructed to account for any particular observation, with the choice often influenced by the particular theoretical lens through which researchers approach

the data. A more robust approach requires theoretical frameworks to guide experiments and inform interpretation of results. Specifically, we must generate falsifiable predictions from competing theories and demonstrate how experimental outcomes would distinguish between different theoretical possibilities. Developing theories prior to experiments may also lead to more targeted experimental designs and reveal that existing data already constrains theoretical possibilities in ways not previously recognised.

In chapter 2, we develop a theoretical framework to explore the utility of targeted neuronal activity perturbations to reveal principles of sensory computation in model networks, and derive testable experimental predictions in a number of applications such as predictive coding theory. We ask what can direct circuit manipulations tell us that classic sensory stimulation experiments cannot? And how could the two approaches be combined?

## 1.3 Learning in sensory systems

### 1.3.1 Perceptual learning in the brain

Several forms of learning are thought to take place within sensory systems across all modalities. Some forms of learning and plasticity involve adaptation, where temporary changes in sensitivity to stimuli occur in response to constant, repeated, or high-intensity stimulation (Kohn 2007). Habituation occurs when behavioral responses to stimuli are reduced following repeated exposure (Worden & Marsh 1963; Randlett *et al.* 2019). Additionally, homeostatic processes are thought to regulate the excitability of sensory circuits to maintain stable activity levels (Davis & Bezprozvanny 2001). Beyond these short-term adaptive mechanisms, associative learning involves the formation of associations between stimuli and various outcomes, such as reward, and encompasses classical and operant conditioning paradigms (McGann 2015).

Perceptual learning manifests as behavioral improvements in sensory discriminations and pattern recognition following training or experience (Seitz 2017). This form of learning is evident in numerous everyday contexts across sensory modalities: radiologists learn to recognize subtle

abnormalities in x-rays that would be imperceptible to laypeople, musicians develop the ability to distinguish subtle tonal differences, and wine-tasting training can refine one's palate for detecting complex flavor profiles. To study perceptual learning in controlled experimental settings, researchers typically train participants to discriminate between simple stimulus features over periods ranging from days to months, with task difficulty gradually increasing as perceptual thresholds improve. Example stimuli include motion coherence, orientation of sinusoidal gratings, visual textures, and auditory tone discrimination Law & Gold 2008; A. Schoups *et al.* 2001; Vogels & Orban 1985; Fiorentini & Berardi 1981; Karni & Sagi 1991; Demany 1985). Vernier discrimination tasks require participants to detect whether two visual elements (such as line segments) are perfectly aligned or slightly offset from each other (Herzog & Fahle 1997). Studies have demonstrated that sleep between consecutive training days is necessary, possibly due to consolidation processes (Miyamoto *et al.* 2016; Karni, Tanne, *et al.* 1994).

Psychophysics is the scientific study of the relationship between physical stimuli and perceptual experience, using controlled experiments to quantify how changes in stimulus features (such as contrast, orientation, or spatial frequency) affect behavioral responses and perceptual thresholds. In the context of perceptual learning research, psychophysical methods involve repeated measurements of participants' performance on specific visual tasks over multiple training sessions, allowing researchers to track improvements in sensitivity and establish the fundamental principles governing how practice changes perception (Kingdom & Prins 2016).

Psychophysics experiments have led to several principles of visual perceptual learning being proposed. One of these is specificity: perceptual learning has been shown to be specific to trained stimulus features such as orientation, spatial frequency, size, motion direction and retinotopic spatial location (Fahle & Edelman 1993; Fiorentini & Berardi 1981; A. A. Schoups *et al.* 1995; A. Schoups *et al.* 2001; Ghose *et al.* 2002; T. Yang & Maunsell 2004; Vogels & Orban 1985; Ball & Sekuler 1982; Crist *et al.* 1997). For example, improvements in orientation discrimination obtained over several weeks did not transfer to different spatial regions when participants' eyes remained fixed on a specific target throughout training (A. A. Schoups *et al.* 1995). Similarly, in monkeys trained on a three-dot bisection task at a fixed spatial location, no transfer occurred to

a three-dot vernier task despite both tasks relying on identical stimuli (Fahle & Edelman 1993). Such findings have been taken to suggest that plasticity can occur at relatively early stages of hierarchical visual processing, where neurons have restricted spatial receptive fields and exhibit selectivity for specific feature dimensions, before integration across feature dimensions occurs to form more complex responses and receptive fields develop greater invariance (DiCarlo *et al.* 2012). This specificity might also imply that perceptual learning is guided by task demands rather than passively extracting statistical regularities from all encountered stimuli. If learning were driven solely by statistical exposure, improvements might be expected across any stimulus dimension encountered during training.

Two broad categories of algorithms for sensory learning are often studied as distinct phenomena: statistical learning and task-driven learning. Statistical learning involves extracting regularities from sensory input, such as learning that certain textures tend to co-occur, that objects have predictable spatial relationships, or that motion patterns follow certain statistical distributions (Aslin 2017; Schapiro *et al.* 2016). This is commonly conceptualised as a passive and unsupervised form of learning that occurs through repeated exposure, in contrast to forms of learning that might be driven by explicit feedback based on task performance or reward signals. Task-driven learning is conceptually distinct from statistical learning because it requires different learning algorithms, typically involving error-corrective processes. Since higher-order brain areas outside of early sensory cortices are thought to facilitate the cognitive processes involved in perceptual decision-making and evaluating the consequences of actions, task-driven learning necessitates feedback signals to be communicated across brain areas. Because statistical learning does not require task feedback, it could potentially be achieved through more local plasticity mechanisms in sensory systems, while the hippocampus has been implicated in the extraction of higher-order environmental regularities (Schapiro *et al.* 2016).

However, given that perceptual learning occurs as a consequence of practicing specific tasks, an alternative possibility is that feedback signals during or after task performance drive learning. Some studies have demonstrated that perceptual learning is regulated by the nature of feedback provided (Aberg & Herzog 2012; Fahle & Edelman 1993; Herzog & Fahle 1997). In

particular, one study of vernier discrimination in humans found that the presence of feedback on performance enhanced perceptual learning, and furthermore, making feedback uncorrelated with the participant's response prevented learning entirely (Herzog & Fahle 1997). Other studies have presented conflicting evidence: for example, Watanabe and colleagues found that subjects exposed to a task-irrelevant visual stimulus below the threshold of conscious perception improved performance in a later test of motion direction discrimination, which might argue for a statistical learning explanation (Watanabe *et al.* 2001).

Several key ideas have been proposed for how neural circuits could implement perceptual learning. First, representation modification through reparameterisation of neural tuning curves would alter the response properties of individual neurons, such as their selectivity or sensitivity to specific stimulus features (Petrov *et al.* 2005). Another motif of representation modification is changes in population tuning, where learning involves transformations that affect how populations of neurons collectively represent stimuli (Failor *et al.* 2025). Selective reweighting is another theory, where adjusting the inputs to decision areas could leave sensory representations unchanged while modifying how these signals are combined and weighted for perceptual decisions (Petrov *et al.* 2005). These potential mechanisms are not mutually exclusive and may operate at different stages of sensory processing or be differentially engaged depending on specific task conditions.

Evidence for tuning curve reparameterisation has yielded what have often been thought of as conflicting results, exemplified by comparing two methodologically similar studies in macaque monkeys. Schoups et al (2001) trained macaque monkeys to fixate on a target, then report through saccades whether a gaussian-windowed sinusoidal grating was oriented clockwise or counterclockwise relative to an implicit (unshown) reference orientation, starting from ±20 degrees. The grating had a fixed spatial frequency and a randomised phase across trials. Over the course of training, task precision was increased, and monkeys were ultimately able to report approximately 1.5 degree rotations with >80% accuracy. Electrophysiological recordings revealed that orientation tuning curve slopes in area V1 after orientation discrimination training, specific to the trained stimulus location (A. Schoups *et al.* 2001). Specifically, neurons whose

preferred orientation was ±16 degrees from the reference angle displayed the highest slope changes. A similar study by Raiguel *et al.* 2006 recorded neurons in area V4, finding that neurons with preferred orientation offset by ±22 to 67 degrees changed the most. The fact that the changes were most prominent in neurons whose preferred orientation was slightly offset from the trained stimulus orientation indicates that the most informative neurons changed the most: those neurons will change their firing rates most for small orientation differences at the trained stimulus orientation (Raiguel *et al.* 2006)

However, another study found no evidence for V1 or V2 changes in bandwidth or amplitude of tuning curves after training, even at the trained stimulus location (Ghose *et al.* 2002). This study used a different experimental paradigm. Firstly, it employed a match-to-stimuli paradigm where the monkey had to report whether two oriented gratings shown in quick succession were different through pushing a lever. Secondly, both the phase and spatial frequency of gratings were randomised across trials. From the same laboratory, Yang and Maunsell (2004) later employed a similar match-to-stimuli paradigm, but recorded area V4, where they found orientating tuning curves bandwidths decreased.

Attempting to reconcile these disparate findings, one theory is that subtle task details account for the differences, namely that the grating stimuli had either fixed or varying spatial frequency (Vogels 2010). V1 neurons are often sensitive to the spatial frequency of gratings (Marshel, Garrett, *et al.* 2011), whereas spatial frequency invariance and other feature pooling occurs higher in sensory hierarchies (DiCarlo *et al.* 2012). From this perspective, training with random phase may have been insufficient to consistently engage the same V1 neurons, while training with fixed phase may have targeted the same V1 neurons in each trial more reliably. Task precision is also thought to influence the site of plasticity. Indeed, monkeys in the Schoups (2001) study (0.5-1.2 degrees) were ultimately trained to perform finer orientation discriminations than in the Ghose (2002) study (3.3-7.3 degrees), offering another potential explanation for the disparities. This illustrates how subtle differences in experimental design can drive different neural outcomes, something that theories must account for.

A study in mice by Failor et al. (2025) emphasises the role of population-level modifications. Here, they investigated how visuomotor task training alters visual cortical stimulus coding, simultaneously sampling thousands of V1 neurons as oriented gratings were shown. The study found that diverse changes in individual tuning curves could be described by a simple and plausible transformation at the level of population responses that effectively suppressed activity in neurons with low orientation selectivity to the task stimuli. While this transformation did affect measured orientating tuning curves, those effects were seemingly complex. Conversely, at the level of the neural population, a straightforward principle emerged: representations of the discriminated stimuli became more distinct, reflected through an increase in the angle between population response vectors in high-dimensional neural activity space, termed orthogonalisation. The degree of sparsening and orthogonalisation appeared to be dynamic across trials, which implied that long-term representation modifications were not responsible. However, this study used a detection task involving sensorimotor association learning rather than a classical perceptual learning paradigm, where mice learned to associate distinct oriented gratings separated by a 45 degree angle with motor actions without further changes in task precision. It therefore remains unclear whether this class of behavioural paradigm reveals mechanisms relevant to perceptual learning, which may involve longer-term changes.

Evidence for selective reweighting comes from both theoretical and empirical considerations. The reweighting theory posits that perceptual learning primarily reflects plasticity in the weighting of inputs from different levels of sensory systems to optimize the performance of a decoder in decision areas, without requiring changes to neuronal tuning curves in early sensory areas (Dosher & Lu 1999; Dosher, Jeter, *et al.* 2013). To test the reweighting theory empirically, Dosher et al. (1998) conducted a psychophysical experiment in which participants practiced an orientation discrimination task while varying amounts of Gaussian noise were added to the stimuli. They measured how performance improved across different noise levels and used mathematical modeling to determine what type of neural change could explain the observed pattern of learning. Their results were consistent with selective reweighting of neural inputs to decision areas, rather than changes in stimulus representations or changes in nonlinearities in the signal processing pathway. Further lines of evidence from large-scale neuronal recordings

in mice could be taken to support this theory. Large populations of V1 neurons have been shown to be highly informative about stimulus features. Ideal decoders trained on the activity of thousands of individual V1 neurons could achieve 0.35 degree discrimination thresholds when discriminating stimulus orientation, without any behavioural training (Stringer *et al.* 2021). They also recorded from neighbouring visual area LM, thought to be the mouse analog of area V2 (Wang, Gao, *et al.* 2011), finding narrowly higher discrimination thresholds of 0.37 degrees. With *neurometric* thresholds lower than perceptual thresholds reported from psychophysics behavioural experiments, this raises the question as to why V1 or LM/V2 tuning curve changes would ever be necessary. However, ideal decoders assume access to all available information, whereas the brain faces constraints such as the accessibility of information through connectivity bottlenecks, noise, and competing computational demands. This implies that learning may still be required in higher-order areas to improve read-outs of such information, even when such information can theoretically be decoded with perfect access to earlier brain areas. Stringer also re-analysed a dataset from Steinmetz *et al.* 2019 where mice were trained to indicate which of two gratings had higher contrast, and argued that observed trial-to-trial variability in the (97.5% accurate on average) decoder performance of V1 responses did not predict behavioural performance variability. This perspective further questions the notion that representational changes in early sensory areas may not be necessary for improvements in perceptual thresholds for simple stimulus features. However, these results indicate that the visual system is already well-optimised for the relatively simple discrimination or detection tasks employed in these mouse paradigms, leaving open the possibility that more challenging perceptual tasks - though difficult to implement in mice - would reveal learning-dependent changes.

### 1.3.2 Learning in artificial neural networks

As discussed earlier, artificial neural networks in machine learning offer an attractive way to model learning in the brain, assuming that connection weights are the primary means through which learning changes the brain. There are three broad categories of learning algorithm that are used to train artificial neural networks: supervised learning, self-supervised learning and

unsupervised learning. With supervised learning, networks are trained using input-output pairs where the desired response is explicitly provided, allowing the network to learn mappings between stimuli and correct responses through error-driven weight updates - analogous to learning with explicit feedback or instruction. Self-supervised learning involves training networks to predict missing or future parts of the input from other parts, such as predicting the next word in a sentence or reconstructing masked portions of an image, thereby learning useful representations from the statistical structure of the data itself without external labels. Finally, unsupervised learning discovers patterns and structure in data without any target outputs, such as clustering similar inputs or learning efficient representations that capture the underlying data distribution - similar to how sensory systems might extract statistical regularities from environmental input without explicit teaching signals.

Deep learning is a type of supervised learning that is applied to multi-layer networks. This was made possible through the backpropagation algorithm, first described by Rumelhart, Hinton, and Williams in 1986 (1986). This breakthrough was historically significant because it enabled networks to automatically discover useful internal representations in hidden layers that capture regularities in the task domain, distinguishing it from earlier, simpler methods such as the perceptron-convergence procedure. The algorithm works by computing the gradient of the loss function with respect to each weight using the chain rule of calculus, propagating error derivatives backward through the network layers to update connection strengths proportional to their contribution to the overall error. Deep learning is theoretically powerful due to its expressivity - the ability of deep networks to approximate arbitrarily complex functions through hierarchical composition of nonlinear transformations - and its capacity for representation learning, where intermediate layers automatically discover useful feature representations from raw data. The foundational architecture for deep learning is the deep feedforward network, also called multilayer perceptrons (MLPs), which forms the basis of many commercial applications and serves as a conceptual stepping stone to more complex architectures (Goodfellow *et al.* 2016). For visual processing tasks, convolutional neural networks represent a specialised type of feedforward network that incorporates spatial structure through convolution operations, where small kernels or filters are applied across the spatial dimensions of the input to detect local

features, with weight sharing enabling the same feature detectors to operate at different spatial locations (LeCun & Bengio 1995). More recently, transformer architectures - another class of deep networks based on attention mechanisms - have revolutionised natural language processing and are increasingly transforming diverse fields from scientific research to everyday applications (Vaswani *et al.* 2017). Deep learning has been the basis of remarkable advances in machine learning applications, revolutionising fields from natural language processing to robotics. In computer vision, deep convolutional networks have achieved superhuman performance on tasks such as image classification, object detection, and medical image analysis, with architectures like ResNet setting new benchmarks across diverse visual recognition problems (Voulodimos *et al.* 2018).

### 1.3.3 The deep learning theory of perceptual learning

Deep neural networks represent a paradigm shift in computational neuroscience models because they can learn end-to-end from naturalistic sensory inputs - such as natural images - without researcher intervention in specifying the computational details (A. Saxe *et al.* 2021). This approach suggests that the sophisticated neural representations observed in biological visual systems might emerge from relatively simple learning principles when applied to sufficiently rich environmental data, rather than requiring explicit engineering of neural computations.

The idea that brains use deep learning has gained empirical support from studies demonstrating that task-driven deep convolutional neural networks (CNNs) can develop intermediate representations that are predictive of neural responses across layers of the biological visual system. Yamins et al. (2014) provided seminal evidence for this correspondence by examining thousands of convolutional neural network architectures that incorporated visually-inspired operations such as linear filtering, pooling, and normalisation. Rather than fitting models directly to neural data, they optimised networks based on object categorisation performance using recognition tasks. They observed that models achieving better categorisation performance were also more accurate at predicting neural responses in primate inferior temporal (IT) cortex, despite neural data not being used during training. Their best-performing model - which achieved human-level recognition performance - showed predictive accuracy for IT neural responses and predicted

responses in V4 cortex through its intermediate layers. This hierarchical correspondence between network layers and visual cortical areas suggested that performance optimisation in artificial networks might reflect similar computational principles to those operating in biological visual representations. These findings led researchers to propose that deep learning-like processes could contribute to the development of visual cortex representations, where complex neural computations emerge organically through interaction between simple learning rules and environmental structure. Extending this work across the human ventral stream, Güçlü and van Gerven (2015) demonstrated that deep learning models could predict functional magnetic resonance imaging voxel responses across several early visual areas, addressing the existence of a gradient in neural complexity throughout the hierarchy. Complementing these findings, Khaligh-Razavi and Kriegeskorte (2014) used representational similarity analysis - which compares dissimilarity matrices between artificial and biological activity patterns elicited by different stimuli - to show that CNNs trained with supervised, but not unsupervised, algorithms learned features similar to those in different visual cortical areas, with higher-order area IT achieving the best object categorisation performance, consistent with hierarchical processing principles.

However, this idea has been highly controversial, primarily because the algorithm through which credit is assigned appears inconsistent with established principles of inter-areal connectivity between brain areas (Lillicrap, Santoro, *et al.* 2020). The backpropagation algorithm requires distinct symmetric feedforward and feedback connections that share identical weights between brain areas, with feedback weights carrying the gradient signals necessary for weight updates. This presents multiple biological implausibilities: the weight transport problem demands symmetric learning rules that do not exist in biology, and such symmetric reciprocal connectivity patterns are not observed in known neuroanatomy. Additionally, backpropagation requires signed error signals and assumes that feedback connectivity does not functionally influence neural responses to feedforward signals in targeted neurons - contrary to established understanding of how feedback connections operate in the brain.

Since the development of backpropagation, researchers have proposed an expanding set of biologically plausible candidate mechanisms that circumvent the requirement for symmetric

connectivity. These approaches demonstrate that mechanisms inspired by realistic principles of neural circuit architecture can enable neurons and neural circuits to compute, or approximate, the gradients required for supervised learning without backpropagation (Lillicrap, Santoro, *et al.* 2020; Guerguiev *et al.* 2017; Amit 2019; Bengio *et al.* 2016; Pozzi *et al.* 2019; Lillicrap, Cownden, *et al.* 2016; Detorakis *et al.* 2019; Whittington & Bogacz 2019; Konishi *et al.* 2023; Max *et al.* 2024). Because convolution operations invoke weight sharing across space, CNNs have been deemed implausible, despite the computational improvements they can offer over fully-connected neural networks. However, Pogodin and colleagues (2021) proposed a scheme through which convolutional-like behavior - one of the neural network operations previously assumed to be highly implausible - could be biologically implemented. To overcome the implausibility of weight sharing among neurons, they devised a network incorporating lateral connectivity and Hebbian learning during a sleep-like learning phase in which subgroups of neurons share their activity, achieving "nearly convolutional" performance. While these studies demonstrate that deep networks can capture static representations across the visual hierarchy, understanding how perceptual learning occurs requires examining how these representations change through training. The reverse hierarchy theory (RHT) offers a framework for understanding how perceptual learning progresses through hierarchical neural circuits. The RHT proposed that perceptual learning follows a gradual process in reverse-hierarchical direction, where modifications progress to earlier levels of the hierarchy only if higher-level changes are insufficient for task mastery (Ahissar & Hochstein 2004). Many of the RHT's key insights and proposals have since been formalised and grounded in deep learning theory, making it relevant to understand this earlier perspective.

The RHT's central premise is that perceptual improvement stems from a gradual top-down-guided enhancement in the utilization of task-relevant information, beginning at high-level areas and progressing to lower levels. Learning initially occurs at high-level areas of the visual system, and only when these prove inadequate does it extend backwards to lower levels, which provide better signal-to-noise ratios. The RHT offers to resolve conflicting findings regarding the neural substrates of perceptual learning by suggesting that task conditions determine plasticity location through this hierarchical progression. For instance, the discrepancy between studies finding V1

changes after orientation discrimination training (A. Schoups *et al.* 2001) versus those finding no V1 modifications (Ghose *et al.* 2002) may reflect differences in task precision and stimulus variability during training. The RHT would predict that when spatial position varied during training, learning should remain at higher cortical levels and show greater transfer, whereas fixed spatial positions during fine discrimination training should eventually recruit lower-level areas, producing more specific learning with measurable V1 changes.

Critically, the RHT proposes that while early sensory areas contain detailed information necessary for fine discriminations, higher-level areas do not always have access to these finer details of lower-level representations. Learning involves making this information progressively more accessible through training. However, unlike selective reweighting theories that propose changes only in decision areas while leaving sensory representations unchanged, the RHT explicitly involves modifications to neural representations throughout the hierarchy. The theory suggests that learning involves a sequential process of modifications that amplify task-relevant information while diminishing irrelevant signals, with these changes occurring in stages starting at the highest levels and gradually extending to lower levels when improved precision is required. At higher levels, neural representations adapt by strengthening responses to task-relevant features while weakening responses to irrelevant ones, directly altering stimulus representation. When learning extends to lower levels, it can produce changes such as more selective tuning curves in V4 or enhanced orientation selectivity in V1, representing direct modifications to sensory encoding rather than merely improved readout. This attention-mediated learning process enables higher-level areas to guide the progressive recruitment of lower-level areas that can provide increasingly precise information.

However, the RHT is not a quantitative theory, and does not provide candidates for a learning algorithm to account for how this reverse hierarchical learning cascade occurs. Deep learning models provide a concrete framework for understanding such processes, as learning in deep networks similarly involves reverse hierarchical propagation of error signals through gradient descent algorithms. This parallel suggests that deep learning principles may offer quantitative

predictions for how perceptual learning unfolds across the visual hierarchy, bridging the gap between the RHT's conceptual framework and mechanistic understanding of neural plasticity.

To formalise a deep learning theory of perceptual learning, Saxe developed a quantitative analytical framework by examining how learning unfolds in deep linear neural networks (A. M. Saxe 2015). This is described here in detail, at a high level, because it directly inspires the experimental design in the study in chapter 3. Using analytically tractable deep linear networks enabled the characterisation of learning dynamics, while avoiding the complexities of nonlinear activation functions. Although the learned function in deep linear networks can ultimately be collapsed to a single linear transformation through matrix multiplication of the weight matrices, their learning dynamics exhibit nonlinear dynamics that mirror those in more complex networks (A. M. Saxe *et al.* 2014).

The model architecture incorporates multiple processing stages: an input stage taken to represent early sensory areas, intermediate stages corresponding to visual areas like V1 and V4, and a 'decision' stage representing higher-order areas involved in perceptual choices. The inputs were one-hot encoded representations rather than pixel-level inputs that complex networks are commonly trained on, allowing for precise analytical treatment. Learning occurs through task-driven gradient descent optimisation of network weights, minimising the squared error across training exemplars. A key derivation involved mathematically reducing the complex learning dynamics to two scalar variables: one variable captured the magnitude of modifications in the input stage and another represented changes in the decision stage. The effect of learning could be captured as an increase in the magnitude of a difference vector that captures the distinction between neural response patterns to the two stimulus orientations in the discrimination task.

The analytical derivation expresses the network weight matrices in terms of these two scalar variables and the difference vector. Through substitution into the gradient descent differential equations and subsequent algebraic manipulation, the complex multi-dimensional dynamics collapse to two coupled scalar differential equations. These equations contain parameters that characterise the discrimination problem: one parameter measures the overlap between the difference vector and input responses, encoding task difficulty, while another measures the magnitude

of the difference vector itself. Under the assumption of homogeneous tuning properties across neural populations (equal tuning curve bandwidths at initialisation), these parameters become related, allowing the dynamics to depend on a single difficulty parameter where small values represent challenging discriminations and large values represent easy discriminations.

For early learning stages, decision-layer changes grow much more rapidly than input-layer changes, demonstrating that higher-level areas undergo plasticity before lower-level areas. For more advanced learning stages, decision layer changes remain consistently larger than input-layer changes by a predictable factor. Task difficulty determines the plasticity distribution across network layers: easy discriminations concentrate changes primarily in decision-layer connections with minimal sensory involvement, while more challenging discriminations produce modifications across both layers, though decision-layer changes remain predominant.

This work established some qualitative hallmarks of learning dynamics in a deep, feedforward and fully connected linear neural network trained with supervised gradient descent learning that could be experimentally tested. Because these hallmarks relate to the dynamics of gradient descent learning, a system where neurons are able to somehow compute these gradients might display the same learning dynamics, even if the backpropagation algorithm was not used directly.

In section 1.3.5, we list these predictions as applied to an orientation discrimination paradigm. First, a study of perceptual learning in a complex, nonlinear neural networks is discussed, to demonstrate that many of the same insights and qualitative hallmarks can extend to this setting. Then, the rationale for using the mouse visual system as a model is set out, before describing predictions applied to this system.

A study from Wenliang and Seitz modelled perceptual learning in a CNN (Wenliang & Seitz 2018). Here, they utilised a modified version of an influential network architecture in computer vision, AlexNet (Krizhevsky *et al.* 2012). Briefly, the network contained 5 convolutional layers and a final fully connected decision layer, and employs nonlinear activation functions. The network layers had been pre-trained on an image classification task. The network was then trained to perform discrimination of images of oriented sinusoidal gratings for angle separations

between 0.5 and 10 degrees, reporting whether a stimulus was clockwise or anticlockwise relative to a reference. To achieve this, two separate, identical 'streams' of the network were set up to process the two stimuli, both converging on a single decision layer. Firstly, training on higher precision discriminations caused less transfer, whilst low precision tasks caused higher transfer to novel stimuli. This model also reproduced the finding that low precision tasks are learned faster, i.e. with fewer training iterations. The prediction that learning high precision tasks requires greater change of earlier areas also held true in the CNN - the layers corresponding to V1-2 did not change unless the discrimination threshold was relatively small, whilst the area corresponding to V4 changed in all conditions. This supports the notion that differences in task precision, and learned perceptual thresholds, pushes plasticity to earlier layers and accounts for the differences in the site of plasticity found in the electrophysiological studies in monkeys described earlier (A. Schoups *et al.* 2001; Ghose *et al.* 2002; T. Yang & Maunsell 2004). Overall, this model is in accordance with both the RHT and many of the qualitative hallmarks of deep perceptual learning described by Saxe (2015).

### 1.3.4 The mouse visual system

We chose to design experiments for mice, because of the unparalleled experimental tools available for recording large populations of functionally defined cortical neurons. Access to a multi-layered hierarchical system is required to test the deep learning theory of perceptual learning. Amongst sensory systems in mice, somatosensation is likely to be a highly salient modality due to the large somatotopic whisker representation present on the dorsal cortex. Whisking is essential for nocturnal animals living in tunnels, such as mice. The somatosensory cortex is also easily accessible experimentally. However, although a primary and a secondary somatosensory cortex (S1/S2) has been defined in mice, we also required knowledge of, and ability to experimentally control, the stimulus features determining responses in individual neurons. This relates to spatio-temporal patterns across of motion across whisker combinations in whisker S1, but is poorly understood (Laboy-Juárez *et al.* 2019).

The visual modality has been historically well-researched in multiple species, partially because it is salient to humans. Visual cortex is relatively more well-defined in mice compared to the somatosensory cortex. Orientation discrimination was chosen for the task because orientation-selective neuronal responses are well-established in mice, combined with there being a substantial body of background literature on these paradigms. Orientation tuning of visual neurons has been studied for decades across many mammalian species (Rose & Blakemore 1974; Hubel & Wiesel 1968; Chapman & Stryker 1993; Marshel, Garrett, *et al.* 2011). Up to 16 visual cortical areas in mice have a distinct retinotopic map of space (Zhuang *et al.* 2017). Orientation-tuned responses are ubiquitous across different cortical layers of mouse visual areas (Kreile *et al.* 2011). Prior studies have also confirmed that orientation tuning responses are present in higher order mouse visual areas, with some variation in properties such as spatial and temporal frequency preferences and proportion of cells displaying orientation responses (Marshel, Garrett, *et al.* 2011; de Vries *et al.* 2020).

We asked whether the notion of a visual hierarchy in mice was supported by both functional and/or anatomical evidence and identified areas to target from this prior research. Firstly, V1 is the primary target of geniculo-cortical projections so is assumed to be the first level of processing. Overall, V1 projects to many higher order visual areas, meaning that there is no perfect hierarchy (Wang & Burkhalter 2007; Froudarakis *et al.* 2019). Area LM is well-established to be the equivalent of V2 in mice - V1 inputs shows a dense, retinotopic projection to LM (Wang & Burkhalter 2007). In addition, LM projections showed the biggest density in areas LI, P, POR, as well as back to V1 (Wang, Gao, *et al.* 2011). Recordings from anaesthatised mice have indicated receptive field sizes are consistent with layers of a hierarchy (D'Souza *et al.* 2022). Two key further studies have provided evidence for hierarchical organization of mouse visual areas following from V1. These studies also suggest the presence of two functional streams, as is posited by the two-streams hypothesis (DeYoe & Essen 1988; Saleem 2020). (D'Souza *et al.* 2022) found evidence through estimation of hierarchical distances with beta regression on anatomical tracing results. The order of mouse visual areas forming a hierarchy was suggested here to be V1, then LM and RL, with area LI at the top. Next, neurons recorded after learning a perceptual decision making task formed two functional clusters based on their changes, and these clusters

were in line with areas deemed to ventral versus dorsal stream associated in mice, suggesting areas LM, LI, P and POR were in the ventral stream and AL, RL plus AM in the dorsal stream, which did not display learning-induced changes (Goltstein *et al.* 2021). Finally, acute optogenetic suppression of areas V1, LM and AL were shown to impair performance in an orientation discrimination task in mice (LI was not tested), suggesting that they are relevant to orientation discrimination task performance (Jin & Glickfeld 2020, although see Otchy *et al.* 2015. Taken together, there is some evidence that lateral visual areas LM, LI and POR form a hierarchy linked to the concept of a ventral stream, whilst area AL could be taken to be in a distinct stream, with V1 as a common input. We chose to record orientation-evoked responses in neural populations from areas V1, LM, LI and AL, representing 3 levels of a putative ventral-stream associated hierarchy, and one level of a putative dorsal-stream associated pathway for comparison.

### 1.3.5 Experimental predictions to test the deep learning theory of perceptual learning

The theoretical work by Saxe (2015) implies that different theories such as unsupervised Hebbian learning and supervised gradient descent make distinct predictions about learning dynamics. Applied to a perceptual learning paradigm, this work predicts that the order of representational changes (which layers change when) and the nature of changes in tuning curves, are distinct under different theories. In networks trained with gradient descent, changes in deeper layers ('higher' cortical areas) precede those in lower layers. Coarse discriminations transfer more effectively to untrained locations than fine discriminations, as they depend on modifications in higher cortical areas that are more invariant to location. In addition, the tuning curves of the most informative neurons within a layer changed the most, whilst the least informative layer changes the most.

Applying this work to the domain theory of perceptual learning and the mouse visual system, several specific predictions are outlined below for an orientation discrimination task with sinusoidal grating stimuli.

1. Learning-induced changes in orientation tuning progresses in a reverse-hierarchical order.

   We aim to investigate the timing and magnitude of changes in neuronal orientation tuning at several stages of learning. Increasing task difficulty would move the locus of change to areas lower in the hierarchy. We predict that shifts in neural tuning will progress in a reverse hierarchical order, with higher-order visual areas LI then LM changing first, followed by changes in V1.

2. The least informative neurons across areas will change the most.

   Considering the layers of an artificial neural network as surrogates for different areas in the visual system, theory from A. M. Saxe 2015 predicts that the least informative layers, those corresponding to areas that do not encode task-relevant information, will change the most. This is due to the hierarchical structure of the learning system, where layers at lower levels must encode task information to inform higher layers. In this work, relevant task information is orientation tuning to the given stimulus.

3. The most informative neurons within an area will change the most.

   When applied to an orientation discrimination paradigm, learning should target the neurons with the steepest slope in their orientation tuning curve around the trained orientation, which would be 0 degrees. These neurons are the most informative because their responses change the most with slight variations in orientation around the steepest part of the slope. Depending on the tuning curve's bandwidth, these neurons are distinct from those that are most active or prefer the trained orientation. However, if common unsupervised plasticity mechanisms are employed, neural tuning changes will primarily target the most active neurons rather than the most informative ones.

4. Spatial transfer decreases with orientation discrimination difficulty

   As the difficulty of the orientation discrimination task increases, the transferability of learning across space will decrease. This decrease arises from the prediction that higher-order visual areas, which have greater spatial pooling and translation invariance, change earlier, meaning that the effects of learning are less local in space. However, with increasing

task difficulty, engagement of lower-order areas like V1 becomes necessary, and these areas exhibit reduced spatial pooling, leading to diminished transfer of learning across space. The directional hypothesis anticipates a negative correlation between task difficulty and the extent of spatial transfer.

Behaviourally, we will test spatial transfer through adding sparse probe trials, in which the stimulus – a gaussian-windowed sinusoidal grating – is centered at randomly selected positions from a coordinate grid. This session will occur only 3 times per mouse, following expert performance at task difficulties of 45 degrees, 35 degrees, and 25 degrees. Each location will be sampled only once for both clockwise and counterclockwise orientations.

5. Spatially local neural changes

   This prediction helps differentiate between locally connected and effectively convolutional neural network architectures. In convolutional architectures, synaptic weight changes are shared across the visual field, meaning that a single filter can be applied across visual space. As discussed earlier, biologically plausible routes to achieving effectively convolutional architecture have been proposed (Pogodin *et al.* 2021). In this experiment, we will map the spatial receptive field of each recorded neuron. If changes in orientation tuning predominantly target neurons with receptive fields at the trained stimulus location, it would challenge the hypothesis that weight sharing is implemented in the mouse visual system.

# 2

# Distinguishing theories of sensory processing using neuronal activity perturbations

## 2.1 Summary

The dynamics and connectivity of neural circuits integrate and transform multi-modal information facilitating perception, cognition and movement. Understanding the principles that underlie neural responses in the brain is therefore a primary goal in neuroscience. A classic paradigm is to correlate external sensory stimuli with firing rate responses in a brain area. However, when a space of potential theories could account for such responses, we need to devise specific predictions that would distinguish and constrain them.

A variety of theories have been posited to explain neuronal responses in early sensory cortices including efficient coding, predictive coding, and supralinear stabilised networks (SSNs). Often these theories predict similar evoked sensory responses, raising a key theoretical question: what experiments could distinguish between these alternatives, revealing the underlying algorithm and circuit connectivity? Predictions generally depend on connectivity and dynamics. Efforts to estimate connectivity from sensory-evoked responses are hindered by the problem that

activity can be correlated without direct connectivity (Das & Fiete 2020). Cellular connectomics approaches using fixed brain tissue offer a route to estimating structural connectivity, but the link between structure and the rich dynamic function of neural circuits in behaving animals is immature. It has long been thought that direct circuit manipulations *in vivo* will help to dissect the intricate circuitry underlying neural computation. Now that these experimental tools have come to fruition (Packer *et al.* 2015) and are rapidly becoming more accessible, we need theoretical frameworks to guide experiments and derive understanding from data.

Here, we develop a framework to explore the utility of neuronal activity perturbations to reveal principles of sensory computation in model networks, and derive testable experimental predictions in a number of applications. We ask what can direct circuit manipulations tell us that classic sensory stimulation experiments cannot? And how could the two approaches be combined?

Although multiple neurons can be targeted simultaneously in experiments, we focus on the simple case of single-neuron perturbations like in the experiments performed by Chettih & Harvey 2019. In order to be able to make useful inferences from perturbations, we begin with a simplifying assumption about the tuning of sensory neurons, namely that weight matrices are circulant. Under this assumption of approximately regular tuning of neurons to stimulus features in sensory cortices, we derive explicit 'influence functions' that describe how the network response to a single-neuron perturbation depends on tuning coefficients for a range of cortical theories and network architectures.

The influence function quantifies how perturbing a single neuron affects the activity of other neurons across the population-essentially mapping the 'ripple effect' of targeted perturbations through the network. We demonstrate that these influence functions reveal fundamentally different information than sensory responses. Crucially, perturbations expose the nature of recurrent connectivity independently of feedforward connectivity or input structure, whereas sensory responses depend on both feedforward and recurrent components. In strictly feedforward networks, influence does not spread beyond the targeted neurons, with the perturbation effect remaining localised. However, in recurrent networks, the influence pattern reflects the underlying connectivity structure and can be mathematically inverted to reconstruct the

recurrent tuning properties that generated the observed perturbation response. This provides a principled method for experimenters to infer circuit architecture from perturbation data.

We demonstrate analytically that there exists a space of different network architectures, employing different connectivity motifs, that produce identical firing rate responses to any sensory stimulus. We then ask whether single neuron perturbations would distinguish these networks. We find that perturbation responses differ amongst networks that respond identically to a sensory stimulus, reflecting the underlying recurrent connectivity motifs, which we show can be reconstructed by reversing the influence function. This reveals that while many theories are not identifiable from sensory-evoked steady state responses alone, they make experimentally distinctive predictions when both perturbation- and sensory-evoked responses are available.

We next demonstrate applications of these general results, showing how both normative and mechanistic principles could be tested using perturbations. We describe a test of a direct latent variable coding scheme, where neuronal firing rates directly represent latent variables that predict the sensory input. These variables are updated through firing rate dynamics that minimise an input reconstruction objective, meaning their values are only updated by unpredictable, or surprising, events. For this reason, we refer to this scheme as efficient coding. Here, we identify a novel link between influence functions and feedforward sensory tuning that is a hallmark of this strategy. We extend this to the nonlinear case of sparse efficient coding, which further incorporates the representational efficiency constraints proposed in efficient coding frameworks through sparsity constraints that reduce redundancy and metabolic cost. In this case, we find that predictions can still be tested if we observe the temporal dynamics of the response, which becomes more informative than steady-state responses. We show that prediction units in a predictive coding network follow the same dynamics as the efficient coding units, but recurrently coupled prediction error units have qualitatively distinct dynamics. We analytically describe the effect of perturbing 'error' and 'state' neurons in a predictive coding network, demonstrating precise quantitative interrelations of influence between 'prediction' and 'error' neurons. This offers a route to probe the dynamical system at the core of this popular circuit model, showing how these cell types must be dynamically coupled to optimise the input prediction objective.

Our results make a case for the usefulness of targeted perturbations guided by careful theoretical considerations, whilst acknowledging how the inferences made from results can change depending on assumptions about network architecture and nonlinearities. Our results show how a combination of perturbations, sensory responses, and theoretical considerations can help constrain long-standing debates about the function of early sensory cortices.

### 2.1.1 A circulant linear recurrent network model

We develop a theory of single neuron activity perturbations in linear recurrent firing rate networks with dynamics

$$\tau \frac{d}{dt} r = -r + Wx + Mr + Bu, \tag{2.1}$$

where neuron-like variables $r \in \mathbb{R}^N$ represent firing rates, $W \in \mathbb{R}^{N \times N}$ specifies feedforward weights, $x \in \mathbb{R}^N$ is a sensory input and $M \in \mathbb{R}^{N \times N}$ specifies recurrent weights. We allow neurons to be perturbed directly through the term $Bu \in \mathbb{R}^N$ where $u$ is a unit step function and $B$ is a matrix that defines which neurons can be targeted through perturbations. Although multiple neurons can be targeted simultaneously in experiments, we focus primarily on the case of single-neuron perturbations like in the experiments performed by Chettih and Harvey (2019). Specifically, we model a single-neuron perturbation by setting the $Bu$ input equal to a vector with a single element set to 1, a delta function in orientation tuning space that is constant in time for t>0. The Fourier transform of $Bu$ is then a constant vector in the spatial frequency domain, suitably scaled: $\frac{1}{\sqrt{N}}\mathbf{1}$. We also treat the sensory input $x$ as constant in time.

We allow neuronal populations to have different tuning curves for feedforward and recurrent connectivity. However, with no assumptions on network connectivity, perturbations of a subset of neurons reveal little about network connectivity. We therefore constrain the form of the connectivity. Inspired by sensory receptive fields which regularly tile a stimulus dimension, we study perturbations in firing rate networks under the assumption of *circulant* stimulus and recurrent tuning. This signifies that for each weight matrix, each neuron has the same shape tuning function but with a different stimulus preference, such as preferred

orientation (2.1A). Functionally, circulant weight structure corresponds to convolution of a tuning curve function across all of input space. With circulant weights, each row is a circular shift of the above row (Gray 2006).

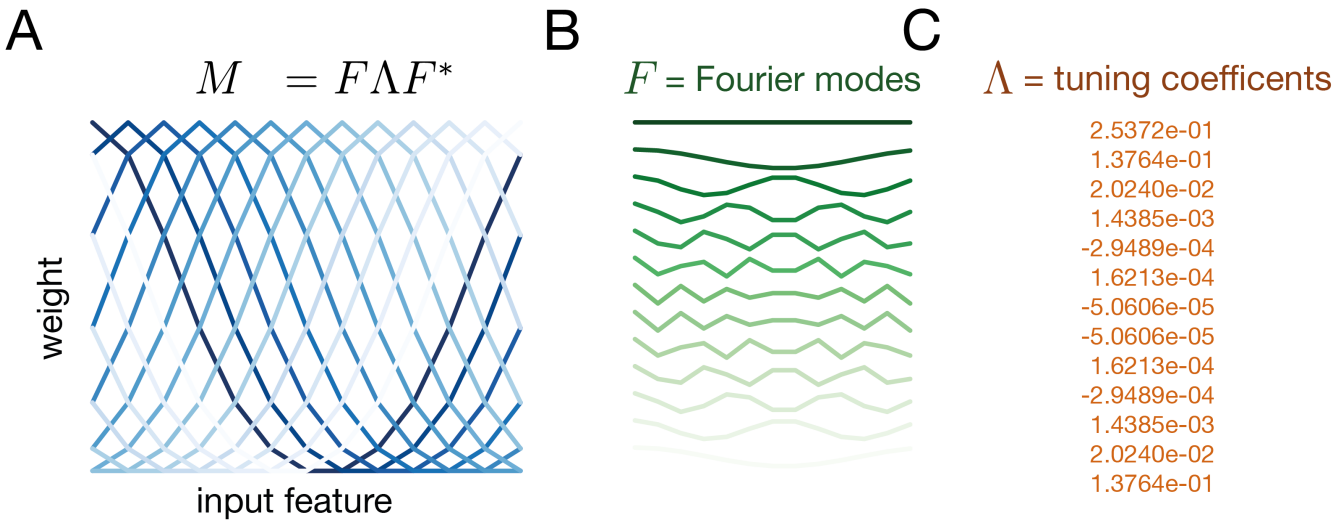



**Figure 2.1: A** An example of a circulant weight matrix $M \in \mathbb{R}^{13\times13}$, where each shade of blue corresponds to a different row of the matrix, which in turn corresponds to the tuning curve of an individual neuron to some input feature, e.g. visual stimulus orientation. These tuning curves use a Gaussian function, but can take any form. This matrix can be decomposed into the product of 3 matrices $F\Lambda F^*$. **B** The eigenvectors of a circulant matrix correspond to a discrete Fourier transform matrix, $F$. Individual Fourier modes (real part only) are depicted in different shades of green. **C** Tuning coefficients (eigenvalues of $M$) that form the diagonal of $\Lambda$ define the mixture of Fourier components that specify the shape of the tuning curve.

All circulant matrices have an orthogonal set of eigenvectors corresponding to Fourier modes (figure 2.1B). In the linear recurrent firing rate network, this means that feedforward and recurrent weight matrices each share the exact same set of eigenvectors. Those eigenvectors have a specific form: a discrete Fourier transform (DFT) matrix $F \in \mathbb{C}^{N\times N}$ where the multiplication $F^*v$ is the DFT of a signal $v \in \mathbb{R}^N$. $F$ is normalised such that $F^*F = I$ where $^*$ denotes the conjugate transpose. Assuming tuning functions are symmetric and even, this drastically reduces the number of parameters needed to fully describe weight matrices to $N/2-1$ for even N or $(N-1)/2+1$ for odd N. The clearest example of potential circulant structure would be in visual networks with orientation tuned neurons, where each neuron is positioned on a virtual ring (Ben-Yishai *et al.* 1995; Rubin *et al.* 2015). Another example might be head-direction tuned cells.

Weight matrices in equation 2.1, when circulant, can be decomposed as:

$$W = F\Sigma F^*$$
$$M = F\Lambda F^* \tag{2.2}$$

Where $\Sigma \in \mathbb{R}^{N\times N} = diag(\sigma_1, ..., \sigma_N)$ and $\Lambda \in \mathbb{R}^{N\times N} = diag(\lambda_1, ..., \lambda_N)$. Here $\sigma_i$ and $\lambda_i$ are eigenvalues of feedforward and recurrent weight matrices respectively, which we call 'tuning coefficients' because they specify a linear combination of frequency components that forms a given tuning curve function. The specific mixture of frequency coefficients specifies determines factors such as the bandwidth of each curve.

Substituting the decomposition of $W$ and $M$ , we can re-express the steady-state firing rates as follows

$$r^{ss} = F\Sigma(I - \Lambda)^{-1}F^*x + \frac{1}{\sqrt{N}}F(I - \Lambda)^{-1}\mathbf{1} \tag{2.3}$$

Where the first term is the effect of any sensory input and the second term is the effect of a single-neuron (delta function) perturbation. This shows that the steady-state firing rates $r^{ss}$ are simply a scaling of any inputs in the frequency domain by a linear function of the tuning coefficients.

This setting still allows diverse possible tuning profiles which express a range of theories of sensory processing. Prior models of cortical orientation tuning have used this assumption through using specific recurrent tuning functions such as Gaussian (Rubin *et al.* 2015) or cosine (Ben-Yishai *et al.* 1995) over preferred orientation, which says that neurons are more strongly connected to other neurons sharing similar orientation preference. In experimental studies, neuronal stimulus tuning is often analysed as an average curve centered around a preferred orientation which implicitly assumes repetitive tuning. Measurements of orientation tuning in visual cortical neuron populations seem to be consistent with circulant tuning (Rossi *et al.* 2020).

In this framework, we are able to solve directly for the effect of perturbations and derive specific experimental tests for theories of cortical function. We next provide examples for testing efficient coding and classical predictive coding.

### 2.1.2 Influence functions describe the effect of perturbing a single neuron on the rest of the network

We firstly define 'influence functions', which describe the effect of perturbing a single neuron $r_i$ on activity in the rest of the network $r$ at steady-state. This is analogous to current experimental paradigms, where targeted neuronal perturbations and neural population activity recordings are typically performed in the same cortical area and layer (Chettih & Harvey 2019).

Influence $d$ represents the activity difference between perturbed and unperturbed trials, accounting for the effect of any sensory input present. In the recurrent network (equations 2.1 and 2.3), this is:

$$d = F(I-\Lambda)^{-1}(\Sigma F^* x + F^* B u) - F(I-\Lambda)^{-1}(\Sigma F^* x) \tag{2.4}$$

$$= F(I-\Lambda)^{-1} F^* B u. \tag{2.5}$$

Equation 2.5 demonstrates that the response to perturbation $d$ depends on a scaling in the frequency domain by the diagonal matrix $(I-\Lambda)^{-1}$ whereas the sensory component of the response depends on scaling by $\Sigma(I-\Lambda)^{-1}$ in the frequency domain. From this we can see that perturbations reveal the nature of recurrent connectivity, and are unaffected by feed-forward connectivity or the structure of the input. Conversely the sensory component of the response depends on both feedforward and recurrent connectivity. In a strictly feedforward network, the influence does not spread beyond the targeted neurons; sensory influence depends on $\Sigma$ whilst perturbation influence amounts to $d = Bu$, i.e. there is no transformation of the perturbation input; the activity difference is equal to the perturbation itself.

We now compute the influence of perturbing just a single neuron on the rest of the network. We assume that $Bu$ is a vector with a single element equal to one, that is, a delta function perturbation. The Fourier transform of a delta function is the constant function, suitably scaled,

$$F^* B u = \frac{1}{\sqrt{N}} \mathbf{1} \tag{2.6}$$

where **1** is a vector of ones. Substituting this into equation 2.5, we have a closed form expression for the influence of a single neuron in a recurrent network:

$$\begin{aligned} d &= \frac{1}{\sqrt{N}} F(I - \Lambda)^{-1}\mathbf{1} \\ &= Fg \end{aligned} \tag{2.7}$$

where $g$ is a vector with elements $g_i = \frac{1}{\sqrt{N}(1-\lambda_i)}$.

Hence the influence function is the inverse Fourier transform of a spectrum that depends only on the recurrent connectivity coefficients $\lambda_i$. Moreover, the magnitude of the influence depends on the number of neurons in the population, with the effect of perturbation decreasing as the population size grows. A purely feedforward network conversely has zero influence beyond the perturbed neuron itself; when all $\lambda_i$ are set to 0, the influence evaluates to $Bu$ which is the delta function. Because of the assumed structure of connectivity, influence can be predicted by the difference in stimulus preference between a measured and perturbed neuron, and influence is greatest in similarly tuned neurons.

### 2.1.3 Perturbations specifically reveal the structure of recurrent connectivity

What do perturbations tell us about this network, and can we learn anything different compared to sensory-evoked responses alone?

Firstly, the influence of perturbing any single neuron measured in the rest of the network can be reversed to reconstruct the recurrent weight matrix $M$ completely. This is a consequence of the circulant structure. This requires the stimulus tuning of each recorded neuron to be known, which is experimentally possible. Furthermore, the activity difference in neurons at the perturbed orientation must be included in the influence measurement. To calculate the recurrent tuning coefficients, one must take the Fourier transform of the influence vector $d$ to recover the spectrum $g$: $F^*d = g$. The free parameters of $M$, which are the recurrent tuning

coefficients can then each be recovered for $i = 1, ..., N$

$$\lambda_i = 1 - \frac{1}{\sqrt{N} g_i} \tag{2.8}$$

Whilst perturbation responses depend only on recurrent tuning (equation 2.7), sensory responses depend on a mixture of feedforward and recurrent weights. Specifically, 'sensory influence' $q$ is:

$$q = \frac{1}{\sqrt{N}} F\Sigma(I - \Lambda)^{-1} F^* x \tag{2.9}$$

Here, the contributions from feedforward and recurrent components can only be disentangled by comparing with perturbation responses.

### 2.1.4 Networks that respond identically to any sensory stimulus can have different perturbation responses

Networks can compute similar functions of their sensory input, using different blends of recurrent and feedforward connectivity. Here we describe a family of networks that have identical steady state responses to any sensory input.

Any two recurrent networks, 'A' and 'B', with tuning coefficients $\Lambda^A, \Sigma^A$ and $\Lambda^B, \Sigma^B$ will have identical responses to any sensory input $x$ if

$$\frac{\sigma_i^A}{1 - \lambda_i^A} = \frac{\sigma_i^B}{1 - \lambda_i^B} \tag{2.10}$$

for all $i$.

Suppose we use as reference a purely feed forward network ($\lambda_i^A = 0$) with tuning coefficients $\sigma_i^A$ (figure 2.2A). Then another network with prescribed recurrent connectivity $\lambda_i^B$ (figure 2.2B) would need feed forward drive of

$$\sigma_i^B = \sigma_i^A (1 - \lambda_i^B), \tag{2.11}$$

$$\tag{2.12}$$

to have the same steady state. Here we must have $\lambda_i^2 < 1$ for stability. From this, we can see that the feed forward connectivity cannot flip sign. Figure 2.2C shows the surface in frequency coefficient space that would align these two networks to produce identical steady state firing rates. Example sensory responses of a feedforward and a recurrent network 'functionally aligned' to a point on this surface is plotted in figure 2.2D.

Whilst sensory-evoked responses would not disambiguate models in this space, perturbation influence functions - which are specifically diagnostic of recurrent connectivity - would differ between models. Conversely, for any two models to respond identically to a perturbation, they must share the same recurrent connectivity.

These results demonstrate the utility of activity perturbations: if we can only observe steady state responses to stimuli, all models in this subspace would be indistinguishable. It also points to the limitations of activity perturbations: if we can only observe steady state responses to perturbations, then we cannot predict the response to different stimuli.

## Comparing sensory and perturbation evoked responses provides a test of efficient coding in recurrent networks

We now ask, can this general theory can be used to test a specific theory?

Efficient coding, a prominent theory of sensory processing, posits that sensory systems should optimally encode maximal sensory information subject to resource constraints.

To formalise this, we define an objective to minimise

$$\|x - W^T h\|^2 \tag{2.13}$$

which corresponds to optimal input reconstruction or prediction. The following firing rate dynamics in a linear recurrent network model of will minimise this objective through gradient descent

$$\begin{aligned} \tau h &= -h + (WW^T + I)h + Wx \\ &= -h + Mh + Wx \end{aligned} \tag{2.14}$$

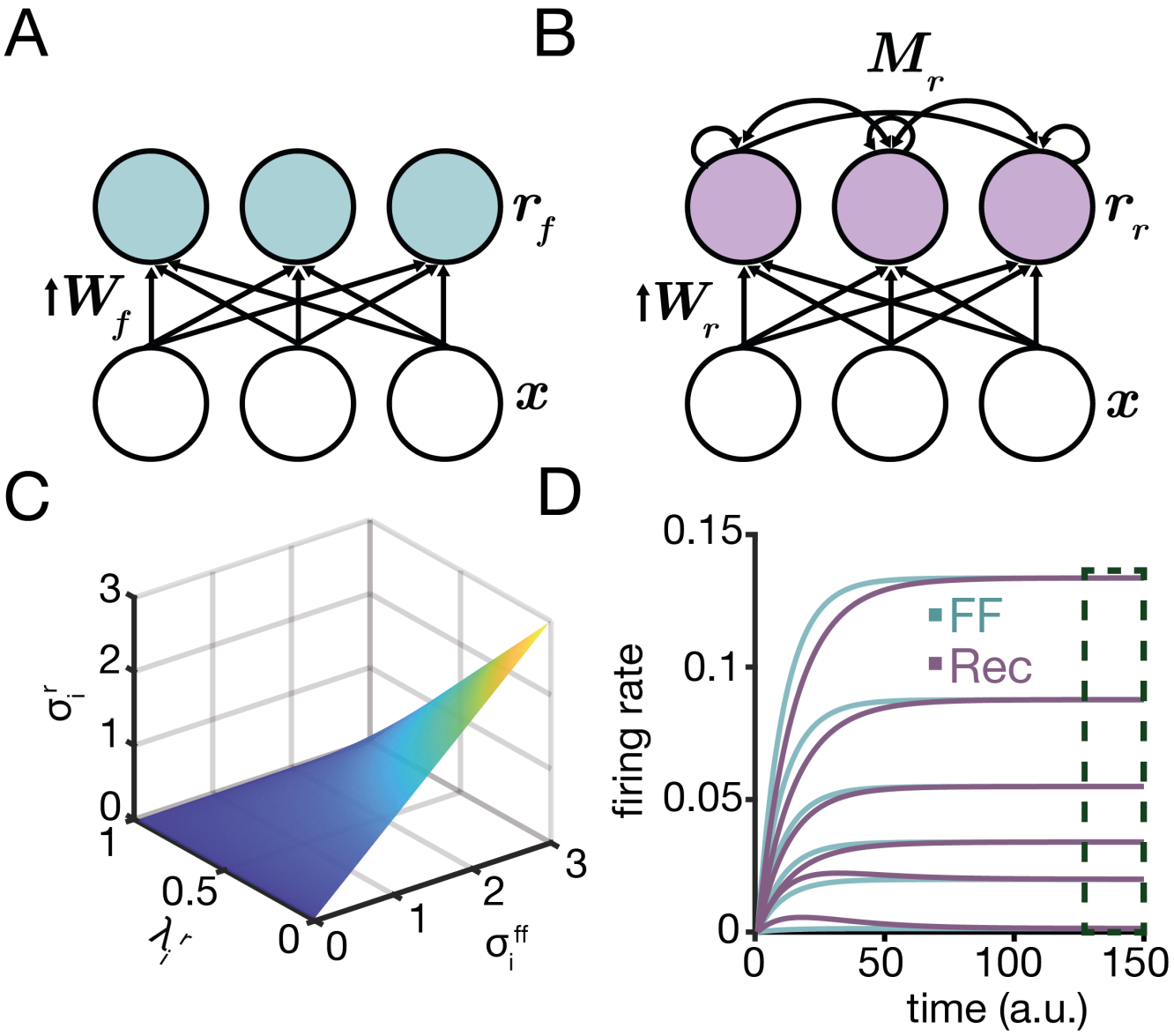



**Figure 2.2:** Examples of 2 possible neural network architectures whose firing rate dynamics arise from different connectivity motifs. **A** a purely feedforward network where $\tau r_f = -r_f + W_f x$ **B** a recurrent network where $\tau r_r = -r_r + M_r h + W_r x$; layer $x$ conveys sensory input. **C** A surface in weight matrix frequency coefficient space functionally aligns linear networks in A to produce identical steady-state firing rates for any sensory input. The required relationship between a given frequency coefficient of $W_f$ ($\sigma_f$) in a feedforward network and the corresponding frequency coefficients of recurrent $M_r$ ($\lambda_r$) and $W_r$ ($\sigma_r$) matrices are plotted. We can derive this relationship for diverse network structures. **D** Simulated firing rates of the networks in A and B with 6 units in response to a random sensory input vector $u$, demonstrating that networks converge to an identical steady-state (region in dashed box).

We note that these equations describe firing rate dynamics only, and we do not address learning here. This means that the weight matrices, fixed here, determine the minimum reconstruction error that can be achieved. Here, the recurrent weight matrix, $M = -WW^T + I$, is a function of the feedforward tuning $W$. This makes $M$ a circulant matrix with frequency coefficients $\lambda_i = 1 - \sigma_i^2$. Feedforward and recurrent weights therefore have to be related in a precise quantitative manner in order to minimize the objective. As a consequence, influence and sensory-evoked responses should be interrelated. Whilst the previously defined general influence function remains valid, influence in this network can be also predicted as a function of feedforward tuning coefficients $\sigma_i$ alone, despite the fact that the feedforward pathway does not directly contribute to influence. Here $d_{EC}$ signifies the influence function in an efficient coding network

$$d_{EC} = \frac{1}{\sqrt{N}} F\Sigma^{-2}\mathbf{1} \tag{2.15}$$

Specifically, perturbation influence should be predictive of responses to stimulation of a single-neuron in the afferent sensory input pathway, which we call 'sensory influence', $s$. In a recurrent network, $s$ is the inverse Fourier transform of a spectrum $q$ that depends on *both* feedforward and recurrent connectivity

$$\begin{aligned} s =& \frac{1}{\sqrt{N}} F\Sigma(I - \Lambda)^{-1}\mathbf{1} \\ =& Fq \end{aligned} \tag{2.16}$$

With efficient tuning principles, sensory influence $s_{EC}$ can additionally be predicted directly from the feedforward tuning coefficients $\sigma_i$ alone

$$s_{EC} = \frac{1}{\sqrt{N}} F\Sigma^{-1}\mathbf{1} \tag{2.17}$$

In this case, if one measures perturbation and sensory influence in the network, plotting the Fourier transform of the sensory influence $q_i$ versus the square root of the Fourier transform of the perturbation influence $\sqrt{g_i}$ should reveal a linear relationship. Alternatively, if one can

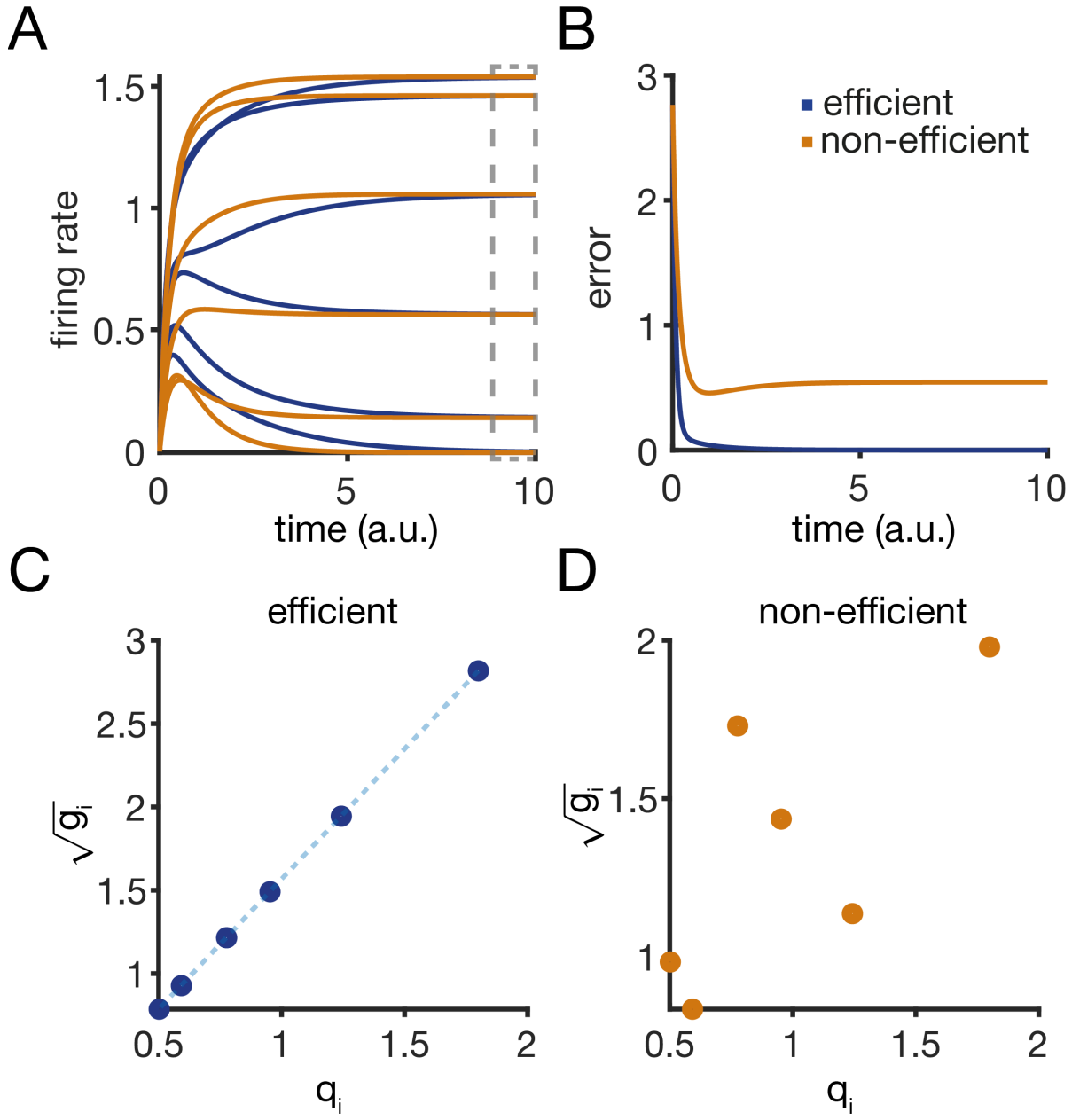



**Figure 2.3:** **A** Simulated firing rates of example 'efficient' and 'non-efficient' networks that are aligned to converge to the same steady-state firing rates (region in dashed box) in response to any sensory input vector $u$. **B** The input reconstruction error (equation 2.13) is minimised through the firing rate dynamics in efficient but not non-efficient networks. **C** In networks with efficient tuning principles, the square-root of the Fourier transform of the influence $d$ from a single-neuron perturbation ($\sqrt{g}$; equation 2.7) and the Fourier transform of the sensory influence $s$ from stimulation of the sensory input pathway ($q$; equation 2.17) follow a linear relationship. **D** Conversely, in a network with unrelated feedforward and recurrent tuning that poorly minimises the input reconstruction objective, this relationship is less linear.

directly measure the feed forward connectivity, for instance by measuring the activity derivative right as a stimulus is presented, this presents another route to testing the theory, described below.

Just as a sensory input is applied to a quiescent network, we have

$$\tau r = -r + Mr + Wx \tag{2.18}$$

$$= Wx. \tag{2.19}$$

We can thus measure the initial rate of change in activity and use this to derive an estimate of $\sigma_i$ by taking the Fourier transform, and plot this against $1/\sqrt{g_i}$, attained from the perturbation-evoked influence function above.

We contrast the efficient coding network with a network that is 'functionally aligned' to converge to identical steady-state firing rates, as described earlier, but has *unrelated* circulant feedforward and recurrent weight matrices (Figure 2.3A).

We refer to this network as 'non efficient', as its firing rate dynamics do not minimise the reconstruction objective through gradient descent. Example dynamics are shown in figure 2.3B. We observed that recurrent tuning curves look qualitatively similar between these networks, despite difference in their function.

Experimentally, one idea to test these principles could be to record both primary and secondary visual cortex (V1/V2), treating V1 as the input layer and V2 as the recurrent layer. Then perturbations of neurons in V1 matching the orientation preference of perturbed neurons in V2 could be used as the delta function sensory input. Alternatively, one could assume that presenting a simple visual stimulus such as an oriented grating, spatially restricted to a very small portion of the visual field, might be equivalent to providing this kind of input to a V1 neuron, but experimental verification would be required.

### 2.1.5 Early response dynamics are informative about recurrent tuning in sparse coding networks

So far we have considered linear models, whereas many theories of sensory processing invoke non-linearity. We sought to understand whether our linear analysis might extend to some nonlinear settings. Experimental and theoretical results suggest that sensory cortical responses are sparse, representing sensory information using a minimal number of neurons (Olshausen & Field 2004). Sparse coding is thought to confer several advantages such as energetic efficiency, increased storage capacity, and enhanced read-out ability.

To model this, we add a term to the objective function to impart a sparseness constraint

$$\begin{aligned} &\text{minimise } \|x - W^T r\|^2 + \alpha |r|_1 \\ &\text{subject to } r \geq 0 \end{aligned} \tag{2.20}$$

Where the scalar $\alpha$ sets the strength of sparseness - equivalent to L1 regularization of firing rates.

Performing gradient descent on this objective translates to the following dynamics in a recurrent network

$$\tau r = f(-WW^T r + Wx) \tag{2.21}$$

Where $f$ is a nonlinear function applied element-wise to a vector

$$f(h) = max(0, h - \alpha) \tag{2.22}$$

Making $r$ threshold-linear units, with a threshold of $\alpha$.

With this objective and dynamics, the optimal response to a sparse, single-neuron perturbation will itself be a sparse vector, a delta function, and the influence $d$ will be a constant vector in the frequency domain. Therefore with this non-linearity, steady-state responses are no longer informative about network weights, or the relationship between feedforward and recurrent tuning. However, we observed that the early transient part of the population response,

immediately after a perturbation or sensory stimulus, behaves close to linearly before diverging. The transient dynamics are therefore informative.

We asked how the early activity transient, just after a stimulus or perturbation is presented, can be used to test for efficient tuning principles, and derived a new test. To achieve this, one must measure the difference in activity derivatives $r$ on perturbed versus unperturbed trials in the network. The Fourier transform $F^*r$ is a vector $\bar{g}$. We then measure the difference in activity derivatives for sensory stimulated versus un-stimulated trials, and the Fourier transform of this is a vector $\bar{q}$. The linear firing rate equations coupled with the relationship between recurrent and feedforward weight matrices predicts that early in the transient response to stimuli, i.e. for small $t$, if the dynamics are initially linear, then there should be a linear relationship between $\sqrt{1-\bar{g}_i}$ and $\bar{q}_i$ (figure 2.4B). $\bar{g}$ is the Fourier transform of the firing rate derivative difference on perturbed-unperturbed trials during a delta function perturbation, whilst $\bar{q}$ is the Fourier transform of the firing rate derivative difference on sensory stimulated-unstimulated trials. The number of neurons N does not affect this relationship, neither does the magnitude of the perturbation.

We show through simulation that the linearity of this relationship is well-preserved in early population responses. This was quantified as variance explained ($R^2$) by a linear fit to the data points (figure 2.4). In recurrent networks with efficient tuning only, $R^2$ was closest to 1 immediately after the onset of the perturbation and decayed whilst approaching steady-state.

These results demonstrate how linear analysis can inform predictions for nonlinear networks, and suggest that in sparse coding networks, the initial rate of change of activity can be more informative than steady-state responses. With a different non-linearity such as a power-law however, testing this theory becomes further complicated.

### 2.1.6 Prediction and error neurons have distinct yet interrelated influence functions in a predictive coding network

We now turn to predictive coding networks. This is a popular theory of sensory processing but has proven difficult to test at neuronal level. In the model used here, the firing rate dynamics

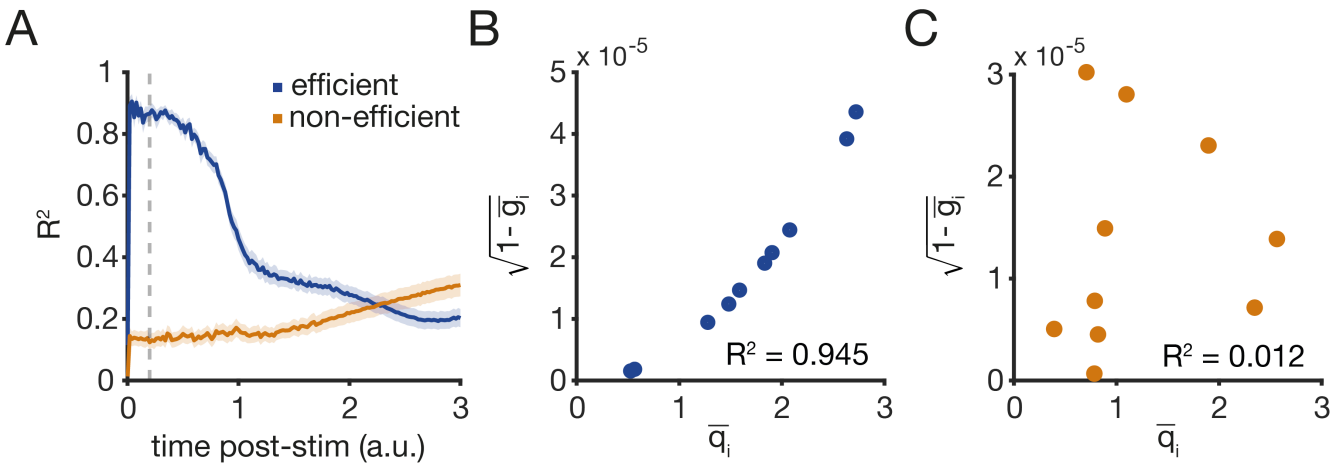



**Figure 2.4:** **A** The early firing rate derivative with respect to time in nonlinear sparse efficient coding networks behaves close to linearly, and contains information that can be used to test for efficient tuning principles. $\bar{g}$ is the Fourier transform of the firing rate derivative difference on perturbed-unperturbed trials during a delta function perturbation, whilst $\bar{q}$ is the Fourier transform of the firing rate derivative difference on sensory stimulated-unstimulated trials. The $R^2$ value from linear regression of these transformed vectors of derivatives is close to 1 for sparse efficient networks but is smaller for non-efficient networks in the early response. The accuracy of this test decays with time after stimulus onset. The plot shows the mean $R^2$ over 100 different weight matrix initializations in a network with 10 units and and $\alpha$ = 0.001. Shading indicates 95% confidence intervals. **B** Example data from a simulation of the network from A recorded shortly after the stimulus onset (dashed line in A), showing a close to linear relationship in the sparse efficient coding network but not the nonlinear non-efficient network shown in **C** where recurrent and feedforward tuning are random circulant matrices and $R^2$ is much lower.

perform gradient descent on the same input prediction objective as before but have dynamically coupled sub-populations of 'prediction' neurons $p$ and 'error' neurons $e$.

$$\begin{aligned} \tau e &= -e + x - W^T p + B_e u \\ \tau p &= We + B_p u \end{aligned} \tag{2.23}$$

Error neurons explicitly represent prediction error, and this is passed to 'prediction' neurons that integrate this input to update the internal state providing predictions. An identical computation to the efficient coding network described earlier is now implemented distributed across multiple neuronal sub-types. Here, error and prediction populations are reciprocally connected through a symmetric connection $W$, which is circulant and has the decomposition $W = F\Phi F^*$ with tuning coefficients $\Phi = \text{diag}\,(\phi_1, ..., \phi_N)$. We allow each sub-population to be perturbed independently through targeting matrices $B_e$ and $B_p$. This could be achieved experimentally through genetic

targeting methods, or through utilizing potential idiosyncrasies in the laminar localisation of prediction and error neurons, outlined in the thesis introduction (Bastos *et al.* 2012).

This predictive coding network yields a set of 4 influence functions with precise quantitative interrelations, describing the expected behaviour when perturbing and measuring the different sub-populations. For the influence functions $d$, we denote the perturbed population with the subscript $d_{e/p}$ and the measured population with the superscript $d^{e/s}$.

Firstly, perturbing error neurons yields a zero influence function across the population of error neurons at steady-state

$$d_e^e = \mathbf{0} \tag{2.24}$$

Intuitively, this is because error responses are 'predicted away' by prediction neurons, whilst they may still display transient responses before convergence. This means that perturbing error must have a positive signed influence on the population of prediction neurons. This positive influence depends on the tuning coefficients of $W$

$$d_e^p = \frac{1}{\sqrt{N}} F\Phi^{-1}\mathbf{1} \tag{2.25}$$

Furthermore, when prediction neurons are themselves perturbed, they suppress error neurons, yielding the same influence function with its sign flipped

$$d_p^e = -\frac{1}{\sqrt{N}} F\Phi^{-1}\mathbf{1} \tag{2.26}$$

Unlike perturbing error neurons, perturbing prediction neurons strongly affects other prediction neurons

$$d_p^p = \frac{1}{\sqrt{N}} F\Phi^{-2}\mathbf{1} \tag{2.27}$$

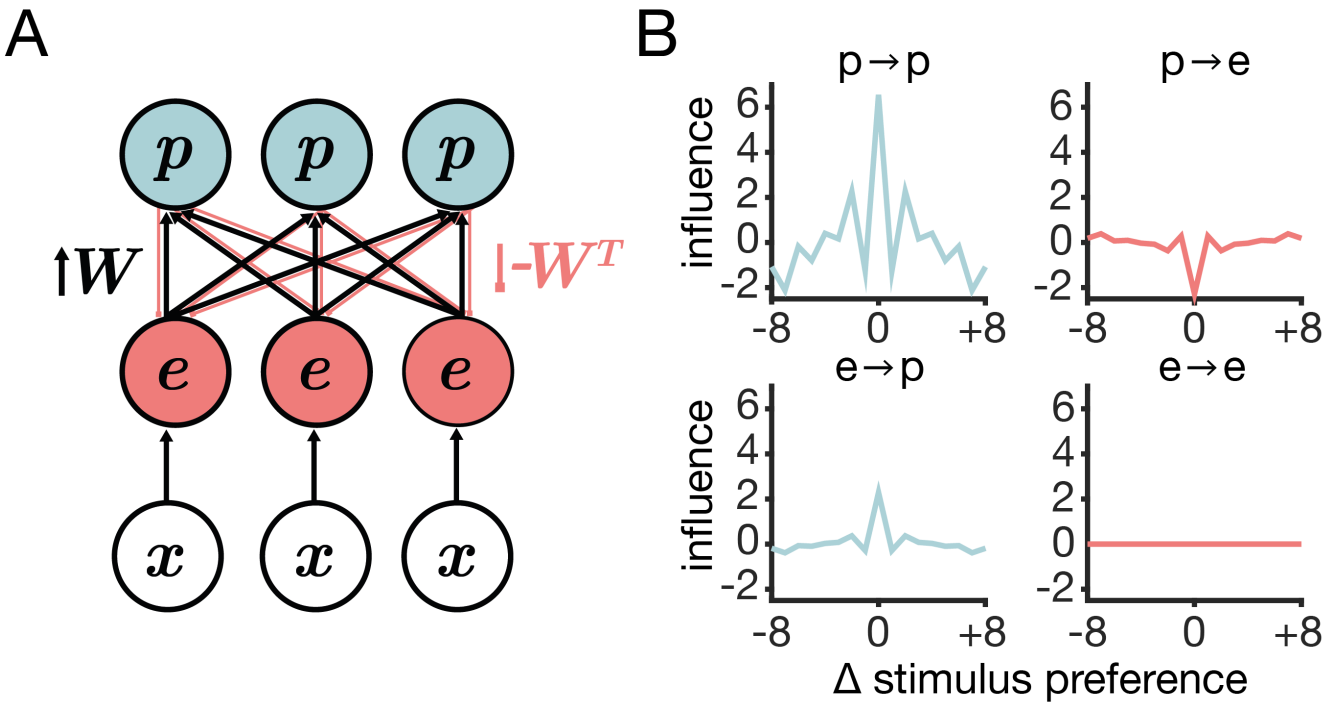



**Figure 2.5: A**Wiring diagram for a predictive coding network. The network has separate sub-populations of 'prediction' ($p$) neurons encoding an input prediction and 'error' ($e$) neurons encoding prediction error. The layer $x$ conveys sensory input. **B** Example influence functions demonstrating differential effects of perturbing dynamically coupled 'prediction' and 'error' neurons on 'prediction' and 'error' neuron responses which distinguish predictive coding networks. Each layer had N=17 units and frequency coefficients for $W$ were randomly drawn.

Knowing any one of $d_p^p$, $d_p^e$ and $d_e^p$ would predict the others. Combined with the proposed mapping of prediction and error units onto neuronal subtypes and cortical layers, our predictions propose a route to directly and quantitatively testing the dynamical system that is at the heart of this model. Crucially, this concerns the causal interrelations between activity in different cell populations in the circuit. This could be complimentary to approaches that correlate neuronal responses with sensory experience during behavioural paradigms that manipulate an animal's expectations.

### 2.1.7 Optimal patterned perturbations

So far, we have considered the case of single-neuron activity perturbations, whereas 'patterned', multi-neuron perturbations are possible experimentally. To approach this, one idea is to

use control theory and optimization approaches to find 'optimal' perturbation patterns to achieve certain objectives, such as to evoke the maximum norm response across a neural population. We might also consider how patterned perturbations align with different directions or trajectories of neural activity in neural state space, and how this could be used to test theories of computation through dynamics.

What perturbation pattern would yield the largest measured steady state difference in firing in the efficient coding network? We wish to solve

$$\max_u ||d||_2^2 \tag{2.28}$$

$$\text{subject to } ||u||_2^2 = 1. \tag{2.29}$$

Substituting the expression for $d$ we have

$$\max_u u^T B^T F (I - \Lambda)^{-2} F^* B u \tag{2.30}$$

$$\text{subject to } ||u||_2^2 = 1. \tag{2.31}$$

Next we assume that all neurons may be perturbed such that $B = I$. The optimal solution to this problem is the eigenvector associated with the maximum eigenvalue of $F(I - \Lambda)^{-2} F^*$. This will be a sinusoid over tuning space with frequency determined by the recurrent connectivity coefficient $\lambda_i$ which is closest to 1. If there are multiple closest coefficients, then any blend of of those sinusoids which sums to one will be optimal. The data in Chettih and Harvey (2019) could be taken to suggest that the optimal perturbation is likely high frequency in tuning space in primary visual cortex - approximately the same as the single neuron perturbation.

Another objective could be to find a perturbation pattern that maximizes the difference in expected responses between two competing network models (similar to the approach of Golan *et al.* 2020). In non-linear networks or with non-convex problems, simulation could be used to identify these solutions.

**Conclusion**

To summarise, we showed that a circulant assumption on weight matrices permits an analytic description of influence functions arising from single neuron perturbations in linear networks.

We showed how this allows us to estimate recurrent tuning curves from responses measured in a neural population. We demonstrate, with experiments in mind, how this could be used to test theories of efficient and predictive coding. We finally showed that, in the simplest nonlinear case, measuring the early temporal dynamics of neuronal responses might still permit the theory tests defined.

# 3

# An experimental test of the deep learning theory of perceptual Learning

## 3.1 Summary

This study aims to test for hallmarks of deep gradient descent learning in the brain. Deep learning algorithms have proven remarkably effective in optimizing multi-layer networks for perceptual tasks, leading to the hypothesis that brains can use similar learning rules to optimise their capabilities. Studies have found similarities between stimulus-evoked representations in intermediate layers of trained deep networks and brains. However, the hypothesis remains controversial because more direct evidence is lacking, and implementations of deep learning often require biologically unrealistic patterns of neural connectivity.

Learning dynamics – the evolving patterns of neural changes across time compared within and across areas – can serve as a crucial hallmark of deep gradient descent learning. For example, for differentiating between deep and shallow models, as well as gradient descent and algorithms such as correlational Hebbian learning. We have applied this framework to the domain theory

of visual perceptual learning and have devised specific experimental predictions inspired by mathematical analysis of deep linear networks (Saxe, 2015) combined with simulations in large, nonlinear networks (Wenliang & Seitz 2018).

In these experiments, we trained mice in a longitudinal, multi-stage orientation discrimination task and recorded orientation tuning curves in thousands of individual neurons across four areas of visual cortex in vivo. These areas, thought to comprise a functional hierarchy, were examined at various stages of learning, from an initial naive state to different levels of expertise (fine discrimination). We also aimed to assess spatial transfer of the visual task will be behaviorally assessed at three learning stages.

This investigation aims to contribute to our understanding of learning mechanisms in the brain, and also address long-standing questions regarding the neural mechanisms of perceptual learning.

## 3.2 Experimental approaches

We now describe the experimental approaches taken. We measured changes in neural population representations of oriented grating stimuli in awake mice across 4 visual cortical areas. Specifically, we measured high-resolution orientation tuning curves that were sampled over small increments of orientation space. We fit tuning curve functions to these responses, from which we measured parameters such as slope and bandwidth. Bandwidth refers to how broadly or narrowly tuned a neuron is to its preferred stimulus feature, quantified thorough measuring the width of the tuning curve at 50% of its peak response amplitude. Measurements were performed whilst mice were task-naive yet water-deprived, then over the course of long-term behavioural training after defined stages were achieved. Mice were trained behaviourally in a novel perceptual learning behavioural paradigm that was designed with several stages encompassing task acquisition then task refinement. Sampling of orientation responses was performed immediately after a session of behaviour task performance. Overall, this approach allowed us to ask whether

the timing and nature of learning-induced changes in neuronal orientation representations across multiple levels of the mouse visual hierarchy match the predictions made earlier.

### 3.2.1 Recording neuronal orientation tuning with 2-photon calcium imaging

We chose the approach of 2 photon imaging to record the neural dataset (Denk *et al.* 1990). Mice were genetically modified to express a fluorescence calcium sensing protein, GCaMP6s (Grienberger & Konnerth 2012; L. Huang *et al.* 2021). GCaMP6s was expressed under the control of CAMKII, which restricted expression to excitatory neurons. This protein undergoes a conformational change when bound to calcium. Cytosolic calcium concentration increases during action potentials (Berridge 1998), meaning fluorescence of a given cell reflects changes in its spike rate. We used a variant of GCaMP with relatively slow kinetics - this maximises sensitivity to neuronal spiking when recording precise temporal dynamics are not important (L. Huang *et al.* 2021). Here, temporal dynamics of responses were not considered as we averaged stimulus responses in a short window after stimulation.

This fluorescent activity sensor combined with 2-photon imaging allowed us to monitor fluorescence in layer 2/3 of neocortex in awake mice viewing orientation stimuli. We note that many previous studies studied the visual system in anaesthetised mice which, through its influence on neural excitatory-inhibitory balance amongst other factors, is known to alter sensory responses (Sengpiel & Bonhoeffer 2002). 2-photon imaging overcomes the issue of biological tissue scattering light which degrades the spatial resolution of an image (Denk *et al.* 1990). This relies on the principle of multi-photon absorption. Here, high-power lasers with deep-penetrating near-infrared wavelengths are scanned across a tissue. Multi-photon absorption to excite fluorescent calcium sensors is highly likely to occur at the focal point of the excitation beam only, hence fluorescence excitation is restricted to a small volume in neural tissue, achieving optical sectioning. Because imaging deep in cortical tissue is made possible through optical sectioning, the technique is minimally invasive, thus recordings using appropriate laser power do not damage tissue. For these reasons, 2-photon imaging permits observation of neural activity in large populations of individual neurons distributed within and across cortical areas.

Crucially, it allows repeated 'chronic' observations across several months. This allowed us to track changes over long-term learning.

## 3.3 Results

### 3.3.1 Establishing a multi-stage visual perceptual learning task for mice

Perceptual learning is defined as a refinement in perceptual discrimination abilities due to training in a perceptual task. To study this in mice, perceptual decision-making (PDM) paradigms are employed. These tasks are generally head-fixed to enable precise control of sensory stimulation in space relative to an animal's sensory organs. This requires mice to learn a sensory-motor association and report their decisions through a motor output such as licking, running or turning a wheel. Previous studies typically compared neural responses in animals in a naive state to animals that have acquired the task. There are likely many distinct learning processes involved in acquiring a PDM task, in particular, learning the sensory-motor association, and learning task contingencies. The former can in theory utilise pre-existing sensory representations without the need to refine them. Here, we designed a task with the aim of partially disentangling sensory-motor association from perceptual learning, a separating task acquisition from task refinement. The multi-stage task also served the goal of observing the learning dynamics of the target neural variables for testing the deep learning theory of perceptual learning.

#### 3.3.1.1 Task design and curriculum

Here, a novel 2-alternative forced choice (2AFC) orientation discrimination (OD) task was designed to assay visual perceptual learning in mice. The task precision, that is the difficulty of the discrimination, increased with each stage, with the goal being to ultimately push mice to the limit of their perceptual abilities. This was designed with the goal of separating perceptual learning from other aspects of task-learning. This task was adapted from the design of a previous two-alternate choice visual discrimination task where a steering wheel was used to move the

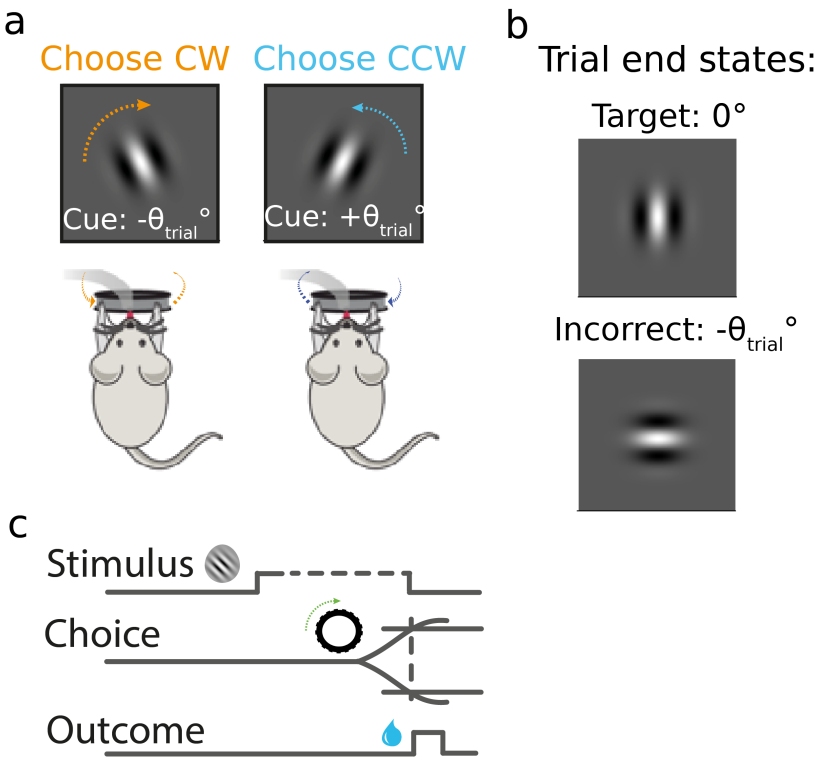



**Figure 3.1:** Schematic of the orientation discrimination task. (**a**) at the start of each trial, a Gaussian-windowed sinusoidal grating was shown at an angle $\theta_{trial}$ either CW (positive $\theta_{trial}$) or CCW (negative $\theta_{trial}$) relative to 0°. The angle $\theta$ shown determined the difficulty or precision of the task, and was determined by the curriculum stage (ranging from 45° to 20°). Mice had to move the steering wheel, rotating the stimulus to meet the target orientation of 0°. Once the grating reached 0° or $-\theta_{trial}$°, the trial ended (**b**). Trial structure is demonstrated in (**c**), and is described in more detail in methods section 3.4.4.1. The dashed line indicates the 'closed loop' part of the trial, where wheel and grating orientation were coupled until a choice was made.

azimuth position of a grating to the center of the screen from its starting position on the left or right side (Burgess *et al.* 2017). The structure of each trial is detailed in methods section 3.4.4.1.

Water-deprived, head-fixed mice learned to turn a steering wheel clockwise (CW) or counterclockwise (CCW) to rotate the orientation of a grating of angle $\theta_{trial}$ from the starting (cue) orientation to the target 'trained' orientation of 0° (vertical) in order to receive a 2µl water drop reward (figure 3.1a). The grating was a spatially-local Gaussian-windowed sinusoidal grating, located at 0° altitude and 20° azimuth in the left visual field. The standard deviation of the Gaussian window was 12 degrees, whilst the spatial frequency of the grating was 0.045 cycles per visual degree. Once the grating reached 0° or $-\theta$°, the grating angle was fixed in place and the trial ended (figure 3.1b). Incorrect choices resulted in a brief auditory white-noise stimulus and additional

time was added to the inter-trial-interval. A basic schematic of trial structure is shown in figure 3.1c, and the detailed temporal structure of each trial is described in methods, section 3.4.4.1.

We intuited that coupling steering wheel position to visual stimulus orientation would make the task more intuitive for mice, speeding up the formation of the sensory-motor association. Therefore the wheel position was coupled to the stimulus orientation, making the task dynamic. Wheel gain was adjusted for different discrimination angles, meaning the movement in degrees did not change with task precision. This was to ensure the motor output required to respond to a trial did not change, in case increasing motor precision changed performance as a function of task precision. This introduced an extra temporal dimension that could increase our ability to distinguish different learning theories. For example, we can distinguish between the neural representation of the initial cue stimulus that guides behaviour, and the trial-end stimulus that is seen after a correct trial (rewarded) or at the end of an incorrect trial (non-rewarded).

To introduce several stages to the training process, a staircase curriculum was implemented. The training began with the easiest $\theta_{trial} = 45$ degrees. When a mouse exceeded 70% trials correct for 3 consecutive sessions, 2-photon imaging with full-field gratings was performed immediately and the task precision in the next session was increased by reducing $\theta_{trial}$ by 10 degrees, then 5 degrees for subsequent stages. To enable psychometric functions, sparse probe trials (20% of trials) covering a range of angles were included after initial proficiency at the 45° stage was achieved. The curriculum was designed based on estimates of orientation discrimination thresholds observed in previous studies, which reported thresholds as high as 30°. In pilot experiments, we had observed that increasing the difficulty by too large a step perturbed the overall performance of the mice, even for the 45° probe trials.

The wheel gain $\beta$ was adjusted for each stage of task precision $\theta_{trial}$ so that the displacement of the wheel required to make a response was held constant at 28 degrees. This decision was made so that the movement required to make a response did not change as the perceptual difficulty of the task increased as the curriculum progressed. Otherwise, lower values of $\theta_{trial}$ would have necessitated finer movements which may have made choices either easier or more difficult, as well as potentially the physical effort expended.

Behavioural training was done in 'offline' in 'behaviour boxes' so that 4 mice could be trained in parallel, increasing the throughput of the training. Imaging was performed immediately after a proficient training session by moving the entire behaviour apparatus to the 2-photon microscope. This enabled imaging to be performed in the same familiar 'context' as the behaviour task without risk of changing the mouse's position or view of the screen. The screen was fixed in place throughout, so each mouse always had the same view of the screen for both behaviour training and orientation tuning experiments.

#### 3.3.1.2 Characterisation of behavioural performance

Overall, 13 mice went through the final behaviour and imaging pipeline, each being trained for between 2 and 3 months. For each cohort, training was stopped after a maximum 14 weeks of water restriction. 53% of mice acquired the task in the allocated time. 7 mice learned the first stage of the task, and at each subsequent training stage (35, 30, 25 and 20 degrees precision) a total of 5, 4, 4 and 2 mice respectively achieved proficiency, undergoing imaging after each stage. On average, it took 42 sessions for learning mice to acquire the first stage of the task. The drop in mouse numbers for late training stages occurred when mice took longer to acquire the first stage. Mice that did not improve in performance (referred to as non-performing mice throughout) were trained for the same duration of time in parallel to the learning mice in their cohort. Behavioural data for 6 non-performing mice was acquired, whilst 3 of these non-performing mice were imaged for a second time at the end of the allocated training period for their cohort.

Figure 3.2 illustrates the dynamics of the stimulus orientation on single trials for different stages of learning in 3 training sessions for one learning mouse (SA033).Figure 3.2a illustrates that at the start of training, initial reaction times (RTs; the time between the stimulus appearing and choice being made) were high, and wheel movements were unskilled. As the mouse progressed through training (figure 3.2b and c), wheel skill increased, reaction times decreased, and the proportion of correct (blue) trials increased (82% correct for sessions shown). Some mice made more rapid choices one one side - this is evident in figure 3.2b where RTs were consistently lower for CCW choices. This reflects our observation that motor strategies for moving the wheel were

diverse across mice. For example, mice could hold on to certain parts of the behaviour apparatus or their angled head-fixation device with one hand, allowing them to rapidly move the wheel in one direction, but more slowly in the other. Other mice were balanced in their wheel control.

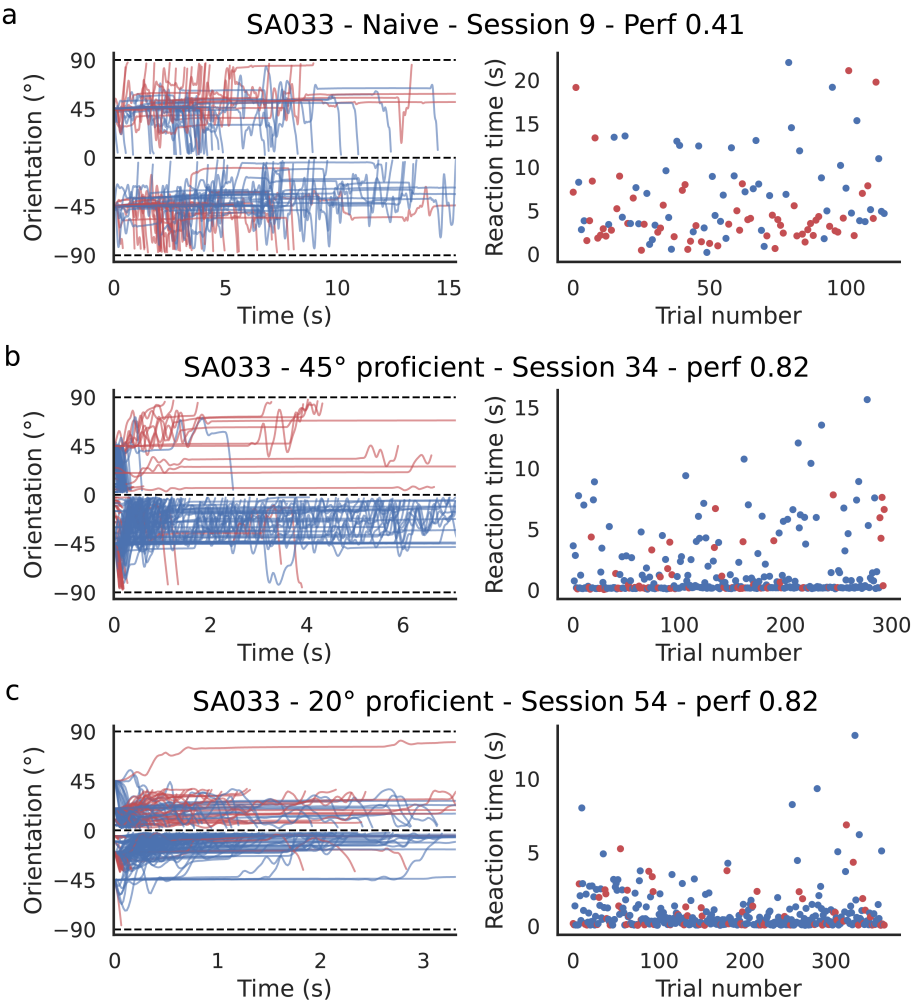



**Figure 3.2:** Within-trial dynamics of grating stimulus orientation in behaviour sessions for mouse SA033. Blue/red colour indicates correct/incorrect trials. Note that the horizontal axis (time) scale changes across rows. (**a**) shows a session from the naive training stage, where 293 trials were completed performance was below chance (41% non-repeat trials correct).(**b**) shows the final 45° stage session before imaging was performed (309 trials). Wheel movement were more ballistic and RTs were lower. (**c** shows the 20° stage proficient session before imaging (405 trials).

Figure 3.3 demonstrates the summary across behavioural sessions for the same example mouse (SA033). Performance, quantified as proportion correct on non-repeat trials, started below chance, rose to chance level which was sustained for some time, then eventually transitioned to proficiency by session 34 (figure 3.3a).

This mouse displayed a CCW (left) side bias on average, which displayed high variability across sessions (figure 3.3b). Bias is defined as the probability of choosing CW at 0° minus 0.5. Therefore unbiased mice have bias 0, CW/right biased mice have positive bias, and vice versa for negative/CCW bias. For some sessions, behaviour was balanced across sides, and for others, high side bias was apparent. Changes in bias did not relate to performance changes in an obvious manner. As session number increased, the cumulative correct responses or reward over trials increased as did the number of trials completed (figure 3.3c). Psychometric curves were plotted, split by training stage. The initial, long training period up to 45° proficiency was split into quintiles (figure 3.3d). Curves for later training stages are plotted in (figure 3.3e).

Ultimately, this mouse achieved the highest all-time performance score across mice, achieving 75% correct over the first session (445 trials completed) where the modal task precision was 15°. After this session, performance dropped gradually back to chance level, a strong left side bias returned, and despite training for an additional 16 sessions, the mouse never met proficiency at this stage. This implies that the performance drop was not due to perceptual inability. Instead, it may relate to effort, motivation, aging, or other factors.

To ask whether perceptual learning was occurring over the course of training, perceptual thresholds were obtained through inverse interpolation of psychometric curves, identifying the stimulus orientation at which 70% correct performance was achieved. Because some mice displayed imbalanced behaviour, we computed perceptual thresholds for left and right sides of the psychometric curve slope separately (figure 3.3f). Change trajectory was indeed asymmetric from left and right thresholds in the example mouse shown (SA033), with right threshold decreasing much more than left, although both did improve. The same was done for psychometric curve slope values, showing that sensitivity increased from session 30 onwards, steeply up to task acquisition. The left slope continued to increase for further training stages indicating that learning continued to take place with left bias (figure 3.3g).

The Shannon entropy of the stimulus orientation, averaged across trials, decreased rapidly at the time of learning, and continued to decrease throughout further training (equation 3.2, figure 3.3h). This reduction reflects smoother movements and a decrease in the variability

of stimulus positions during the closed-loop phase, indicating improved wheel control when making choices. Wheel position entropy over entire sessions was relatively low in comparison and did not decrease concurrently with stimulus orientation entropy for this example mouse. This disparity shows that the mouse enhanced its control over the stimulus during the closed-loop periods specifically, and task learning led to more effective manipulation of the stimulus. Finally, RTs decreased substantially over the first 20 sessions of training, long before proficiency was reached. Interestingly, RTs dropped long before a decrease in stimulus orientation entropy was observed. The latter only occurred around the time of proficiency in the 45° training stage. This dissociation suggests that the sensory-motor association was learned early on, reducing latency to initiate wheel movement after the visual stimulus appeared. However, the precision required to consistently reduce stimulus variability during the closed-loop (reflected by stimulus orientation entropy) likely developed later as the mouse refined its motor control skills and task strategy. The same figure for each individual mouse is shown in the appendix, section .1.

#### 3.3.1.3 Summary of behavioural performance metrics across mice and learning status

A cross-mouse summary of the average learning dynamics of behavioural variables discussed above is shown in figure 3.4. Here, normalized sessions are shown on the horizontal axes to account for the different number of sessions spent in each stage by each mouse. Average performance trajectories (P(correct)) split by learning status are shown in figure 3.4a. Non-performing mice, by definition, never exceeded 50% correct. Learning mice all began performing below 50% correct. This arose because mice began with a side bias yet had to repeat incorrect trials. Each mouse then ascended to a phase of performing at 50% correct, showing that they learned to move the wheel both ways to make CW and CCW choices. This transition did not occur in non-performing mice, who remained more side-biased on average (figure 3.4b). Overall, a generally left/CCW side bias was prominent in all mice (3.4b). Bias was variable across sessions. Some bias reductions were observed up to initial learning, but the average magnitude of bias then increased. The sign of the bias was generally consistent throughout training, but for one learning mouse, there was a strong bias that switched sign.

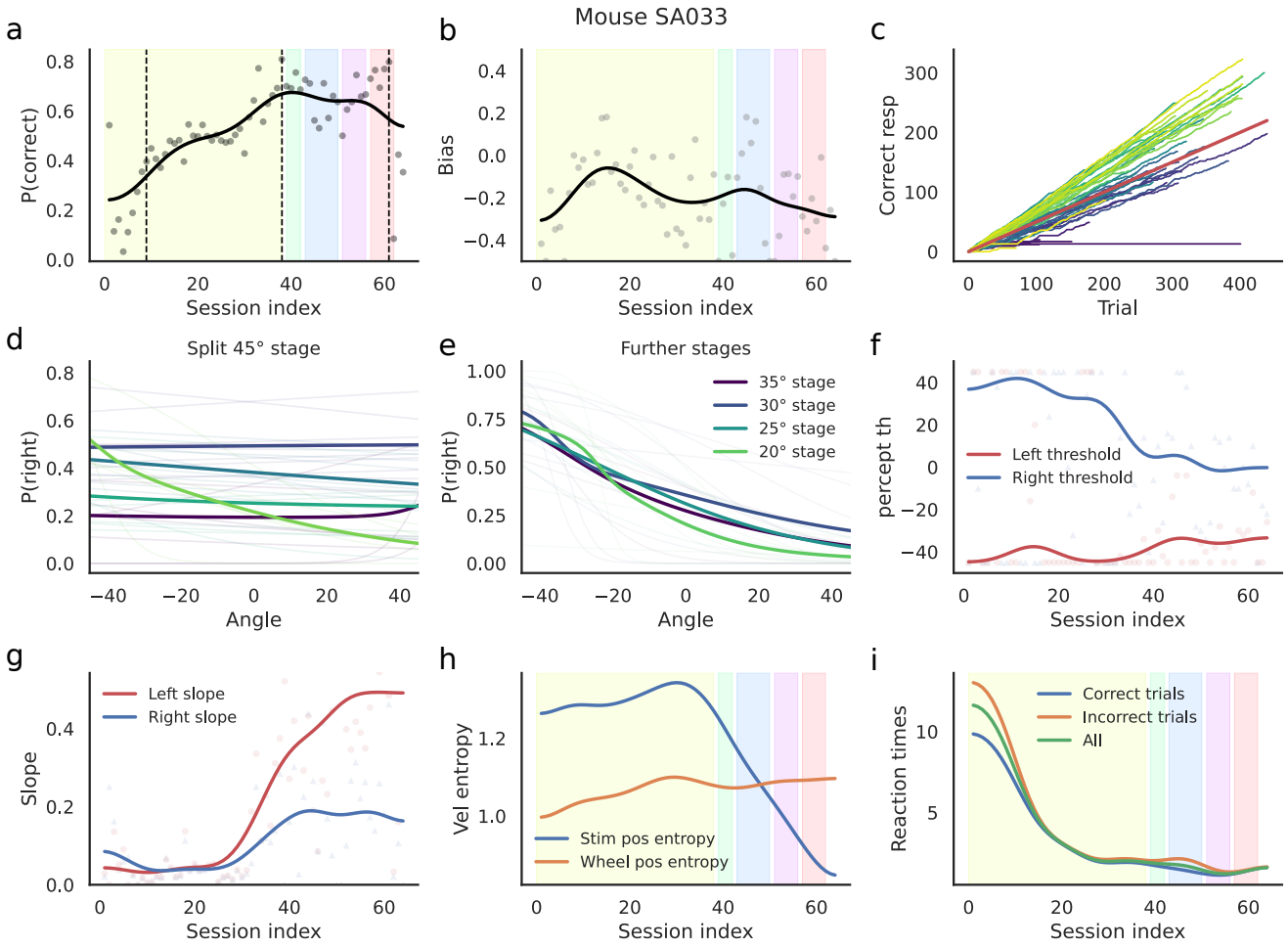



**Figure 3.3:** Behavioural data summary across sessions for one example mouse (SA033). (**a**) Proportion correct non-repeat trials over the course of training. Dashed vertical lines indicate the training sessions shown in figure 3.2. (**b**) bias (probability of choosing CW at 0° minus 0.5) over the course of training (**c**) cumulative correct trials (or reward amount) for individual sessions (lines). Dark blue to light green line colour signifies increasing session number. The red line is unity. (**d**) psychometric curves for 5 groups of sessions (quintiled) whilst mouse was naive. Dark blue to light green indicates earlier to later sessions up to 45° acquisition. (**e**) psychometric curves for further training stages, after the introduction of sparse probe trials at non-modal cue stimulus angles to improve curve estimation. (**f**) perceptual orientation threshold determined from interpolation of psychometric curve at 70% correct value, computed separately for left and right sides of the curve.(**g**) psychometric curve slopes, computed separately for left and right sides of the curve, between 0 and 45°. (**h**) average Shannon entropy of the stimulus orientation (stim pos entropy), or wheel position (wheel pos entropy) over sessions. (**i**) average reaction time (RT) across sessions. RT is the time between stimulus appearance (start of closed loop) and choice being made (stimulus orientation reaching 0° or $-\theta_{trial}$° degrees))

After precision changed, we were surprised to observe that mice generalized rapidly to the novel cue angles. Some mice exhibited generalization in the first session, maintaining their

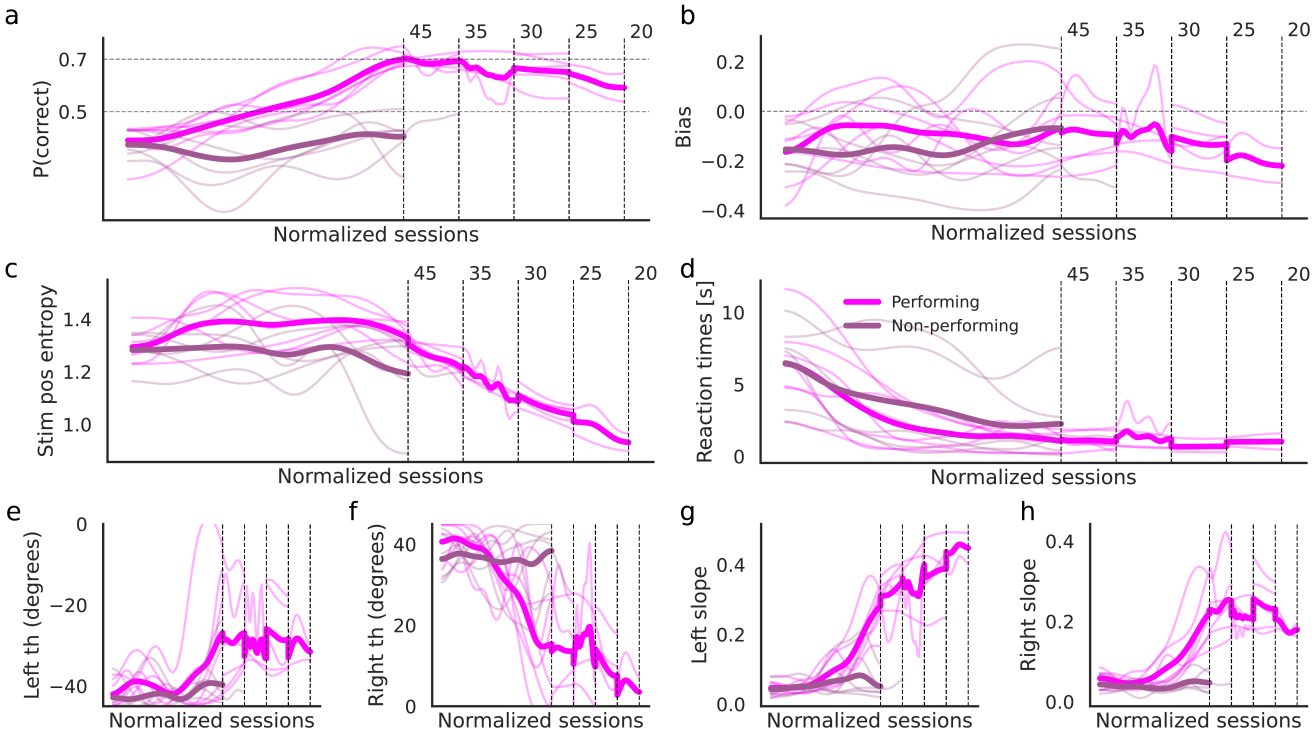



**Figure 3.4:** Average learning dynamics of behavioural metrics over the course of task training. For each subplot, thick lines denote the average across mice split by learning status; performing or non-performing. Thin lines correspond to individual mouse traces. Sessions were normalized to allow comparison over mice, accounting for the different number of sessions up to each training stage.(**a**) Proportion correct non-repeat trials. (**b**) side bias, defined as probability of choosing CW at 0° - 0.5. (**c**) Entropy of the stimulus position (orientation) array. (**d**)(**e**) perceptual threshold (orientation (°)) determined from interpolation of psychometric curve at 70% correct value for left / CCW side of the curve or (**f**) for right/CW side of the curve (**g**) psychometric curve slope taken on the left/CCW side or (**h**) the right/CW side.

performance (P(correct)) entirely, whilst for others, there was a transient, almost negligible drop in P(correct)). Mice reached the proficiency criterion for the 35° in 3-5 days on average, with 3 being the floor level according to our definition of proficiency. This is in stark contrast to the training taken to acquire the initial 45° stage. Once we had discovered this in an early training cohort, for some mice trained later, we skipped the 35° training stage to allow more time for training in higher precision stages. This explains gaps in the data plotted in figure 3.4 for 35°. We could not, however, entirely change the curriculum without restarting the study, losing precious data. We decided against this due to time constraints. Performance did drop for mice at the 25° stage, and two mice that reached the 20° stage.

The entropy of the seen stimulus orientation (stimulus position entropy) decreased substantially

at the time mice learned the task, and continued to drop with further training (figure 3.4c). Overall, RTs decreased early in training for all mice on average (figure 3.4d). 2 of the 6 non-performing mice did however not display the typical rapid reduction in RTs. This could indicate that these non-performing mice struggled to form the initial sensory-motor-reward association that was necessary for further task progress.

Next, we examined average perceptual thresholds for left and right sides of the psychometric curves (figure 3.4e and f). We also examined left and right psychometric curve slopes, which measure perceptual sensitivity (figure 3.4g and h). These metrics offered clues to the behavioural strategy, and to whether perceptual learning was occurring over the course of further training. There was often asymmetry in left and right thresholds for performing mice at the time of initial learning, reflecting uneven psychometric functions around 0°. One-sided strategies and side-bias complicated the ability to use psychometric threshold as a singular measure of learning as is common in human and primate psychophysics. At the end of all training for the best mice, average left and right thresholds were -30 and under 10 degrees respectively. Psychometric slopes for non-performing stayed close to 0 on average (figure 3.4g and h). During the process of task acquisition, when P(correct) was first increasing above 0.5, left and right psychometric slopes both became steeper for learning mice (0.3 left, 0.22 right at proficiency, taken between 0 and -45° or 45° for each stage shown). For further training stages, on average the left slope continued to increase. This demonstrates that perceptual sensitivity improved with further training.

We next asked if it was possible to better understand how the components of task learning develop over time. We explored whether it was possible to dissociate different elements of learning from the temporal trajectories of the behavioural metrics discussed. We approached this by simply comparing the time-series of p-values for each metric to the metrics recorded at the first training time-point (naive, figure 3.5a,b). We also compared temporal trajectories between learning versus non-performing mice (figure 3.5c,d) and for non-performing mice vs. naive (figure 3.5e). Sessions were normalized to allow comparison over mice, accounting for the different number of sessions taken by each mouse to learn the task.

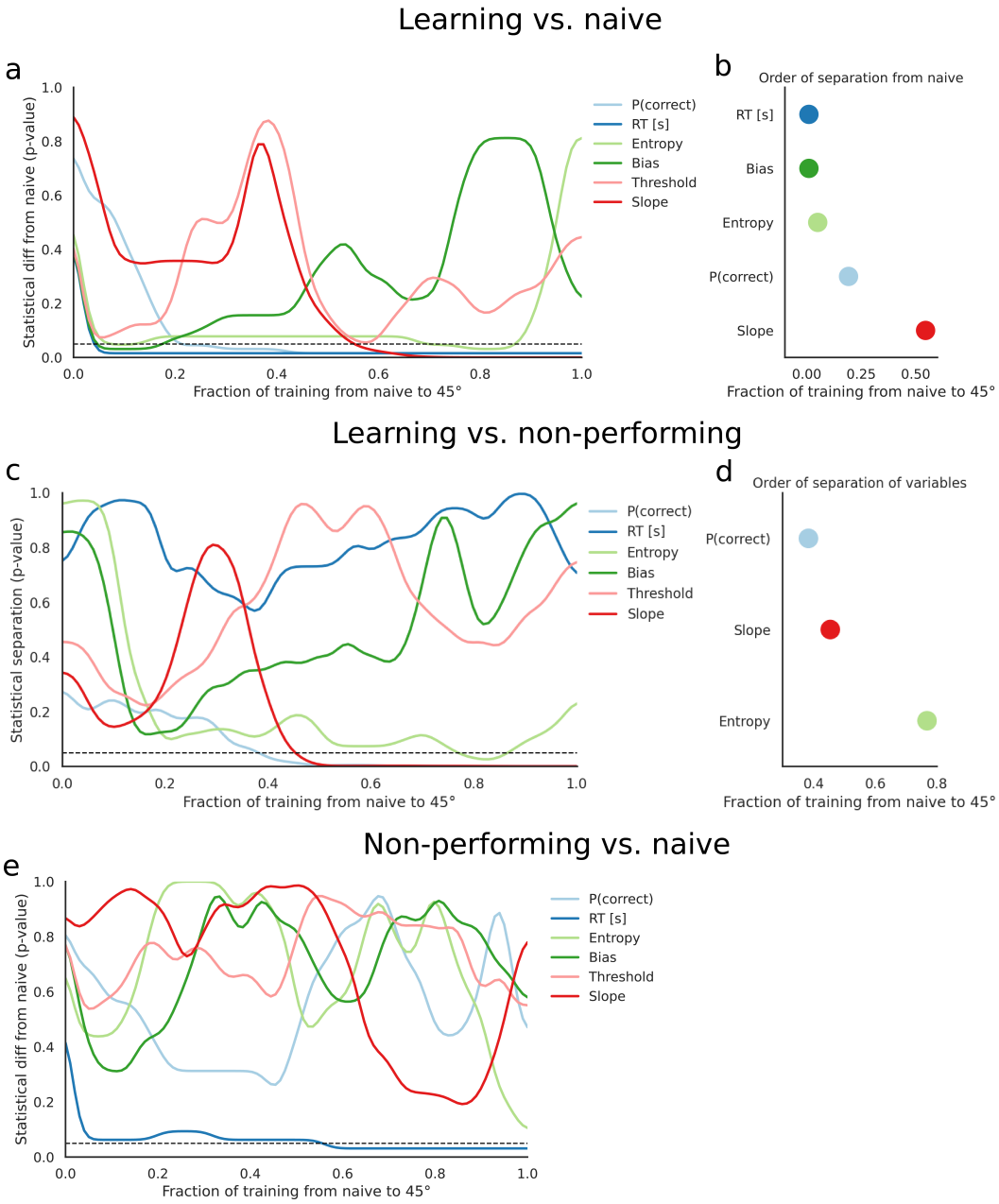



**Figure 3.5:** Statistical separation of behavioural trajectories, comparing (**a&b**) learning mice vs. naive, (**c&d**) learning vs. non-performing mice through training, and (**e**) non-performing mice vs. naive. 'Naive' refers to the behavioural metric values at the first training time-point (static). For the left column, the p-value from a Mann-Whitney test comparing the two described groups is plotted for each of the 6 behavioural metrics (see legend). The horizontal (time) axis is the fraction of the first learning period up to when 45° was acquired (0 to 1). For non-performing mice it is the fraction of all sessions. The dashed horizontal line marks p=0.05. The right column compares the temporal onset of statistical significance (P<0.05) for each metric. Here, psychometric slopes and thresholds were combined across left and right sides.

For all mice, RTs rapidly separated from the naive level (decreased) at the start of training (figure 3.5a and e). Of all the variables, RTs separated the least between learning status (figure 3.5c). RTs did therefore not depend on task learning, instead likely reflecting sensorimotor association. Soon after, learning mice but not non-performing then displayed separation of stimulus position entropy, reflecting wheel skill. Next, P(correct) separated from naive values for learning mice 20% into first stage training, and from non-performing mouse values 40% into training. Slope of the psychometric curve finally changed last, just after 50% of the first learning stage. P(correct), slope and entropy were the only metrics that statistically seperated between learning and non-learning mice. Notably, the stimulus position entropy did not statistically separate from naive levels for non-performing mice. Overall, this analysis indicates different components of task learning develop in order. First, a sensorimotor association is made that lowers RTs. Wheel control skill then follows. Only then does task performance develop, requiring both CW and CCW wheel movements, recognition of CW and CCW stimuli and knowledge of task contingencies. Once competent in these skills, perceptual sensitivity improves finally.

Figure 3.6 shows the stimulus angle exposure time distribution seen by each mouse by session, from early to late training (dark to light colours). For learning mice (top 2 subplot rows, labelled), angles inside $\pm\theta_{trial}$ were seen for proportionally more time as training progressed. For some mice, this was balanced across CW and CCW sides. For other mice, exposure time for angles between $-\theta_{trial}$ and 0° was greater than for angles between 0 to $-\theta_{trial}$. These mice were slower to move the stimulus to goal for CW correct compared to CCW correct trials. This reflects differences in wheel skill or wheel strategy.

### 3.3.2 Localisation of mouse cortical visual areas in vivo using widefield calcium imaging

In order to target 2-photon recordings to areas V1, LM, LI and AL, widefield calcium imaging was used to determine the borders of visual areas. These maps were later used to assign each recorded neuron to a visual area. To achieve this, we recorded a bulk fluorescence signal from GCaMP6s across the region of dorsal cortex in the cranial window whilst bars with a flashing checkerboard

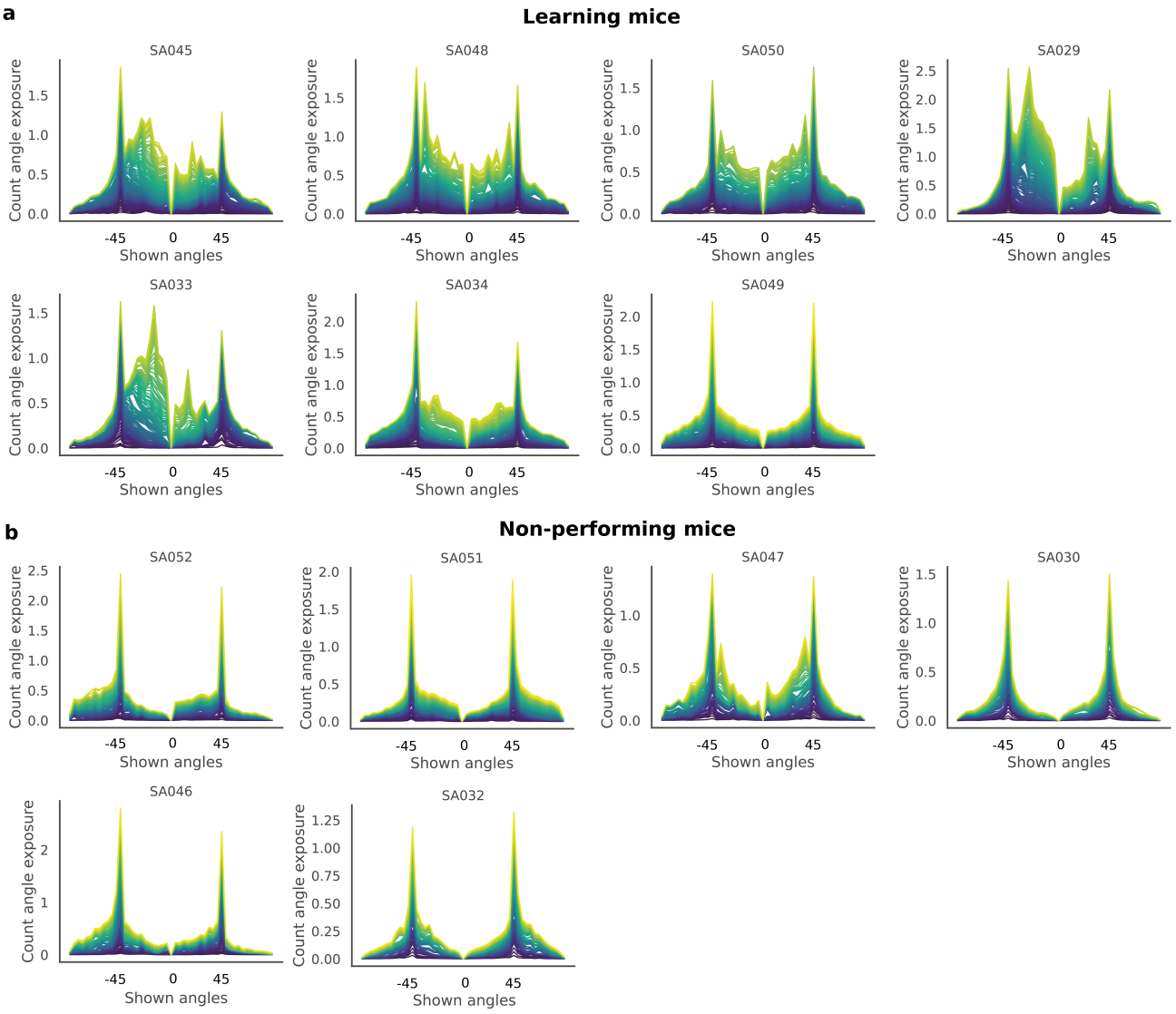



**Figure 3.6:** Cumulative stimulus angle exposure time distribution in closed-loop period per training session. Dark to light colours indicate early to late training sessions. Learning mice area shown on the top 2 subplot rows, non-performing mice on the bottom two.

pattern were swept across the screen in the 4 cardinal directions. This was repeated 12 times per direction, before stimulus-triggered average $\Delta F/F$ movies for each direction were created.

Widefield fluorescence recordings were used to create a map of the different visual cortical areas for each mouse, overlaid on an image of the cerebral vasculature. Briefly, this involved computing a Fourier transform of the signal in each pixel of stimulus triggered averaged videos, isolating the signal at the frequency of the visual stimulation. As each visual area contains a

retinotopic representation of visual space, the phase of this signal in each pixel maps on to the position of the stimulus along the altitude or azimuth axis depending on the cardinal direction of the stimulus. Plotting the phase of each pixel reveals a continuous gradient within a visual area, with altitude and azimuth gradients mutually perpendicular. The gradients were used to create a field sign map. A field sign map provides information about the orientation of retinotopic representations in the cortex, specifically whether a given visual area's representation of space is mirrored (inverted) or non-mirrored. Neighboring visual areas typically exhibit opposite field signs, meaning that gradient reversals can be used to delineate the borders between areas. Borders of areas V1, LM, AL and LI were determined using an automated analysis of field sign maps, described in methods 3.4.3.3.

Overall, V1 and LM were imaged in all 13 mice included in the study results. AL was imaged in 11 of 13 mice, including 6 learning mice. It was possible to reliably image area LI in the cranial windows of just 7 of the 13 mice included in the study results, of which 4 were learning mice. LI was typically closer to the right edge of the cranial window which made it more difficult to image with 2-photon microscopy; off-center movement of the 16x microscope objective was limited due to the deep imaging well made of the head-plate, cement and gasket. In addition, due to particular curvature of the brain where LI lies, the brain was less 'flattened' here with respect to the window making it harder to find a good layer 2/3 imaging plane.

### 3.3.3 Estimating neuronal orientation tuning curves in 4 visual areas with in vivo 2-photon calcium imaging

We measured orientation responses in populations of neurons with single-cell resolution across 4 visual cortical areas. This was performed before learning when mice were 'naive', after the initial acquisition of the behavioural task (early learning) and at up to 4 time points following further training (late learning). 3 non-performing (NP) mice were imaged a second time after undergoing training for a similar amount of time as 'learning' mice. Naive recordings were performed whilst mice were water deprived and mouse body weights were in the target range

of 80 to 85%, in case hydration state affected neuronal responses. Recordings made after the initial naive stage were initiated within the first 30 minutes after a behaviour session ended.

To measure orientation tuning in single cells, in vivo 2-photon (2P) laser scanning microscopy was used to image fluorescence emitted from the calcium indicator GCaMP6s in neocortical layer 2/3 excitatory cells (figure 3.7a). Specifically, recordings were made within the depth range of 160 to 220 microns from the pial surface. Spike responses detected with slow variants of GCaMP have been shown to correlate highly with ground-truth spike trains measured with electrophysiology.

#### 3.3.3.1 Visual stimuli

To evoke neural orientation tuning responses, full-field, spherically-corrected drifting sinusoidal gratings were displayed with a spatial frequency of 0.045 cycles per degree, temporal frequency of 0.69 cycles per second and fixed phase. This meant that grating stimuli used for behaviour training and sampling neural orientation responses had the same spatial frequency, meaning visual cells potentially involved in behavioural responses were more likely to be sampled. Orientations were sampled randomly, without replacement from a set of values ranging from 0 to 360 degrees, in 1-degree increments. Further details can be found in 3.4.5.2.

#### 3.3.3.2 Imaging settings

Imaging settings involve a trade-off between quantity of neurons recorded and quality of single-neuron data. With the goal of reconstructing high-resolution orientation tuning functions in single neurons, we chose to image each 2P imaging field of view (FOV) in an individual experiment. We used resonant scanning to capture 1024 x 1024 pixel images of an 800 x 800 µm FOV, with soma diameters of approximately 16 pixels. Images were acquired at a rate of 15 frames per second (FPS). Although one can technically capture the slow-decaying GCaMP response with a slower frame rate according to the Shannon-Nyquist samping theorem, which would then allow additional FOVs to be sampled simultaneously, we also considered that averaging a higher number of frames would improve the signal to noise ratio (SNR) of stimulus

response estimates. For example, in a 2-second post-stimulus response window, one would capture 30 frames vs. 6 frames for 15 vs. 3 FPS sampling of one imaging plane. As dwell time per pixel would be equal in both settings, and assuming noise across frames is uncorrelated, SNR will be improved. In addition, any noise source whose Nyquist frequency is higher than the frame rate is less likely to alias into the frequency range of the GCaMP signal, thus preserving SNR.

Positioning of the 2P imaging FOV was guided by comparing the morphology of large blood vessels to those visible in the widefield visual area map for each mouse, aiming for the same 5 locations each time (figure 3.7b). These FOVs were: the V1/LM border, in the central-left visual field; medial V1, representing the slightly more lateral left visual field; lateral LM; plus areas LI and AL. The area in the cranial windows corresponding to LI and AL was smaller, meaning that fewer neurons overall were sampled from those areas.

Although it is possible that some neurons matched up between recordings, we did not explicitly attempt to record the same neurons across time-points. This option was trialled in a pilot study and was deemed to be too ambitious given the number of recordings that were necessary. This is because the angle of the head-plate differed slightly across mice; mice had to be removed from the stereotactic ear-bars during surgery to attach the head-plate at a 40 degree angle, reducing consistency. The head-fixation apparatus positioning in several axes was adjusted for each mouse because of the sensitivity of the behaviour performance to positioning. Extra time would have been required to find the exact plane in 3D space for matching cells between each recording, meaning mice would have been head-fixed for too long. In addition, we were not aware of suitable pre-existing analysis tools for tracking cells over recording sessions, meaning the analysis would be very time consuming.

#### 3.3.3.3 Dataset characterisation

Following pre-processing procedures (materials and methods 3.4.5), we captured a median of 690 (IQR 369) putative neuronal ROIs - hereafter referred to as neurons or cells - per FOV. An example FOV with corresponding ROIs is shown in figure 3.7c. Split by area, imaging FOVs

from areas V1 and LM had a median of 739 (IQR 312) detected neurons, area LI had 469 (IQR 213) neurons, and AL had 612 (IQR 216) neurons. The average fluorescence traces from 20 neurons from figure 3.7c are shown in figure 3.7d. Fluorescence traces were deconvolved to get a non-negative inferred spike rate-like 'activity' variable (figure 3.7e, 3.4.5.3). The mean activity in the 1-second post-stimulus period was used to fit orientation tuning curves (OTCs).

7 of 163 recordings (4.3%) were entirely excluded from the dataset on quality control grounds (3.4.5.3) which was mainly due to Z-drift. 17 of the used recordings (11%) with late-onset or very minor gradual Z-drift were truncated to leave just the orientation tuning experiment, and re-processed to pass quality control checks, meaning a subset neurons have OTCs but no spatial receptive field (RF) data. For the recordings that were truncated, 4779 neurons were from the naive stage, 4181 neurons were from proficient stages, and none were from non-performing mice. These neurons were distributed across V1, LM, LI and AL (2486, 4446, 1078 and 695 respectively). We confirmed that the main results of the study are robust to excluding these neurons entirely.

figure 3.7f shows the total number of neurons in the dataset split by area and training stage. Across training stages, the naive stage contained most neurons, from 13 mice, because collecting this data was not contingent on acquisition of the behavioural task. The number of neurons sampled then reduces over subsequent training stages, due to several mice not progressing in the task. Most neurons were sampled from areas V1 and LM. Fewer neurons were captured from area LI, due to issues mentioned in 3.3.2 A past study indicated that a lower proportion of cells are responsive to oriented grating stimuli in LI and AL compared to V1 and LM (Marshel, Garrett, *et al.* 2011). This predicts that fewer neurons per FOV would be detected by the Suite2P classifier whose performance depends on neuronal activity.

#### 3.3.3.4 Fitting orientation tuning curves

Neuronal responses were aligned to stimuli and split into trials. Examples of responses to trials sorted by orientation are shown in figure 3.8b for the neurons in figure 3.8a. For each orientation tuning trial (section 3.4.5.2), the average response in the 1-second post-stimulus period was

calculated. The responses were then averaged across direction, meaning that responses to angles 0 to 180 degrees were averaged with responses for 180 to 360 degrees (the same angle but with opposite grating drift direction). This produced a $180x1$ response vector for each cell (red points in figure 3.8c), which was used to fit an orientation tuning curve.

Gaussian process regression (GPR) was employed to fit orientation tuning curves (3.4.5.5). Three example tuning curves are shown in figure 3.8c. GPR is a non-parametric Bayesian approach that provides a flexible way to model data without assuming a specific functional form. It works by defining a prior distribution over possible functions and updating this distribution with observed data to obtain a posterior. In the context of tuning curves, GPR allows the data to dictate the smoothness and overall shape of the curve for each neuron individually. We chose this method because it imposes minimal assumptions on the shape of tuning functions, resulting in a less biased estimate of the true underlying tuning curves. This is in contrast to parametric methods like Gaussian or polynomial fitting, which impose more rigid assumptions. Fitting a Gaussian function - a traditional approach - would impose symmetry of the OTCs around their peak for example. We ensured that the fitted curves respected the circularity of the data – the 0 and $179^{th}$ elements of the vector represent responses to orientations that are 1 degree apart, so the function should have similar values at those points. The downside of this method is that it is computationally expensive, requiring several days on a high performance computing cluster to fit the dataset.

Orientation values between 0-180 are expressed relative to 0 degrees (vertical) on a semicircle, with negative values denoting counterclockwise orientations i.e. those between 90-180. For example, 135 degrees is mapped to '-45 degrees', signifying a 45 degree counterclockwise rotation relative to 0. Positive values between 0 and 89 denote clockwise angles relative to 0 degrees and are unchanged. This allows an easy mapping between the behavioural task and results.

#### 3.3.3.5 Estimating spatial receptive fields

Immediately following the orientation tuning experiment, a 'sparse noise' experiment was performed to map the spatial receptive fields of the neurons in the same imaging FOV. This would allow us to ask whether any changes in orientation tuning were restricted to the location of the stimulus in the behaviour task, which we call the trained location. In this experiment, a sequence of trials was shown, where each trial contained a small set of different probe stimuli - white or black squares measuring 7 x 7 visual degrees (figure 3.7a). Probes were presented at different locations on a 15 x 16 grid of space in an uncorrelated sequence that sampled each location 10 times (3.4.5.7). For each neuron, trials for each probe location were averaged (figure 3.9a) to produce a receptive field (figure 3.9b). Each receptive field was assigned a Z-score value (methods 3.4.5.7) which reflected the strength of the receptive field centre. The distribution of RF Z-scores over the dataset is shown in 3.9c). This was a right-skewed distribution, with a median Z-score of 4.3. RF size was estimated based on variance of the responses over space in altitude and azimuth directions (methods 3.4.5.7, figure 3.9d). The weighted average receptive field center location was estimated for each neuron, and the center coordinate plus RF radius was used to determine whether a neuron was in the trained location. The detailed procedures can be found in methods 3.4.5.7. A map of RF centre locations is shown in 3.9f. To validate the RF center locations, we plotted neurons on a subset of V1/LM border FOVs and coloured each neuron coloured by its azimuth location (figure 3.9g) or altitude location (figure 3.9h). Azimuth is known to be mirrored over the V1/LM border, so we confirmed that the RFs approximately followed a gradient of increasing azimuth perpendicular to the border, with the central visual field running along the border (Zhuang *et al*. 2017). Some hetereogeneity was present, which was expected based on prior results. In addition, an altitude gradient should run perpendicular to the azimuth gradient, from the top left (lower visual field) to bottom right (upper visual field) in 3.9h.

#### 3.3.3.6 Neuron inclusion criteria

Signal to noise ratio and recording quality could conceivably co-vary with changes in training level, because over the months following the surgery, the cranial window can gradually lose clarity, for example due to re-growth of dura or bone. Expression of GCaMP could also change. Studies using electrophysiological methods like single-unit recordings typically select neurons at the recording stage. For example, neurons are only 'visible' if they display stimulus responsiveness (A. Schoups *et al.* 2001) making it likely that all recorded data points are really neurons, with good signal to noise, possibly even with a bias towards neurons with higher selectivity. However, with 2P calcium imaging, typically a large area of cortex is recorded simultaneously and activity-dependent classifiers are used to detect ROIs. Despite manual curation of ROIs based on visualising images, we considered that some proportion of the dataset could be non-neuronal and/or noisy and introduced some filtering steps to prevent this from biasing our population-based measurements.

A higher proportion of non-neuronal or highly noisy ROIs in the data could manifest as apparent changes in population orientation tuning parameters like selectivity or bandwidth, in a direction that depends on the parameters of the method used to fit OTCs, in particular here the length scale bounds of the of the Gaussian process regression. Because the dataset contained over 100,000 neurons, it was not possible to manually check the quality of each orientation tuning curve fit. To account for the above considerations, we filtered neurons conservatively based on inclusion criteria. These were: OTC SNR greater than 2.5, and for analyses that include spatial RF data, RF Z-score greater than 3. SNR calculation is described in 3.4.5.6 and Z-score in 3.4.5.7. In addition, OTC bandwidths were cut off below 9 degrees which excluded over-fitted noisy ROIs. As these procedures are not standardised, we made best efforts to choose these parameters in a principled way based on inspecting random subsets of fitted tuning curves from naive-stage data, before results were known.

Figure 3.10a provides a general characterisation of the fitted OTCs grouped by visual area, and the outcome of filtering neurons based on the criteria stated above. Cells from area RL are

included in this figure only, for interest, as there were insufficient cells here for analysis split by training level and mouse. The total cells detected likely depends on several factors such as recording length, recording quality, plus parameters of the ROI detection algorithm. Filtering cells based on their OTC SNR removed approximately half of the dataset for areas V1, LM and AL, leaving 46%, 53% and 52% of cells remaining, respectively. LI had a higher proportion of SNR < 2.5 cells, leaving only 38% of cells after filtering. Cells with suprathreshold OTCs and spatial receptive fields comprised 30-31% of total cells for V1, LM and AL, and only 20% for LI. Most neurons with supra-threshold OTC SNR still had low orientation selectivity (< 0.3). Medium to high selectivity cells (selectivity > 0.3) comprised 49% of cells that had suprathreshold OTC SNR in V1 and LM, 45% in AL and only 29 % in LI. Distributions of orientation selectivity by area are shown in 3.10b, demonstrating right-skew distributions for each area, with slightly lower selectivity on average in LI. Figure 3.10c demonstrates that there were relatively similar right-skewed bandwidth distributions in each area, with medians between 24 and 25 degrees.

Figure 3.10d and e show histograms of modal and mean orientation preferences, indicating a clear bias towards the cardinal orientations (horizonal and vertical gratings). Horizontal orientation preferences were particularly over-represented in the LM and V1 datasets. This is consistent with previous findings (Kreile *et al.* 2011; Fahey *et al.* 2019). Differences in mean from modal preferences may be accounted for by any deviations from symmetric, single-peaked tuning functions (Failor *et al.* 2025).

### 3.3.4 Changes in neuronal tuning across stages of task precision and visual areas

In this section, we examine the learning dynamics of orientation tuning variables, asking whether they change from naive through to proficient stages of the behaviour task, and whether this differs according to visual area. This allows us to test the hypotheses that were formed based on the neural network models of perceptual learning, as well as further data-driven exploratory analysis. All analyses performed are based on repeated observations i.e. quantifying the change ($\Delta$) within each mouse using the naive stage as a baseline. As such, each mouse is considered

to be one statistical sample, meaning that the results for all neurons belonging to one mouse were averaged together before comparing across mice.

#### 3.3.4.1 Slope changes were not specific to the 0° orientation

We quantified the slope of OTCs at the target orientation in the behavioural task, which was 0 degrees (figure 3.11, column 1). The model predicted changes would be specifically targeted to this orientation. For comparison, we also examined the slope at several other behaviourally relevant reference orientations, such as -45 and 45 degrees (averaged), which were the cue stimuli seen at the start of trials in the initial phase of training, and 90 degrees, which was seen at the end of all incorrect trials (figure 3.11, columns 2 and 3). We note that unlike in previous studies that used a static orientation task, it is not straightforward to define a 'task-irrelevant' orientation.

Absolute slope values were measured over the interval of 1 sample of the normalized fitted curve, which corresponded to 0.36 degrees. The measured slopes were expressed as function of the distance of each neuron's modal preferred orientation (PO) from the chosen reference orientation (RO; for example, 0 or 90 degrees). This is demonstrated in figure 3.11. This distance determines the steepness of the OTC slope; neurons whose PO is exactly at the RO will have a slope of 0, whilst surrounding neurons (in orientation space) with a PO just offset from the RO will have a relatively steep slope. The gradient of the slopes this PO-RO axis depends on OTC bandwidth.

The neurons with the steepest slopes at the trained orientation correspond to the 'most informative' neurons. In previous studies, changes in OTC slope occurred in this specific steep slope interval of PO-trained orientation space (A. Schoups *et al.* 2001). We therefore quantified slopes only in neurons whose PO was a defined distance of the reference orientation, specifically between 3 and 18 degrees. This interval is visualised with vertical dotted lines in each subplot in figure 3.11. All further analysis of OTC slopes is performed only for neurons preferring orientations from within ±3 to ±18 degrees of the reference orientation. This excluded neurons with zero slope at the reference orientation, as well as neurons with relatively low slope. Because the total sample size of filtered LI and AL neurons was already limited before filtering according

to PO as described, the estimate of average slope within this interval was highly variable, meaning results were inconclusive (3.11d and e). The same visualisation is shown for non-performing mice in figure 3.12.

OTC slopes measured after task training were subtracted from slopes measured at the naive level. These slope changes (Δ slopes) are plotted in figure 3.13. Raw slopes for each training stage are shown in the appendix (figure 14). Because LI was not imaged in every mouse, combined with the multiple data filtering steps relating to OTC SNR and preferred orientation, the sample size of mice for each training stage for area LI was ultimately only 1 or 2 mice (a minimum of 10 cells at each stage were required for a given data point to be included). Results from area LI are unclear due to these sampling issues.

Figure 3.14a visualises the results for the 3-task relevant reference orientations according to visual area and training stage. These results displayed substantial variability across mice, making a clear interpretation difficult. In order reduce to variability arising from sampling fewer mice at later training stages, we pooled results from all 'late learning' training stages together and compared this to 'early learning' (figure 3.14b). The 'late learning' stage consisted of all neurons from task stages 35°, 30°, 25° and 20° for each mouse. This allowed us to compare early learning (where mice had just acquired the initial 45° stage) to later learning stages where the precision of the orientation discrimination task was increasing.

We expanded the range of reference orientations analysed to further explore the dataset, asking whether there was some pattern in the relationship between slope change and reference orientation. Slope change for all mice recorded post-task acquisition as a function of reference orientation is plotted in figure 3.15. All task-proficient stages were pooled for learning mice.

This visualisation reveals that traces for individual mice displayed moderate variability across different reference orientations. For non-performing mice, the results displayed more substantial variability, due to the lower sample size of 3 mice. Despite this, a pattern was apparent in the average across mice, grouped by learning status.

Firstly, for learning mice, on average positive slope change was seen across the whole range of orientations. For non-performing mice, average slope change was non-positive across most reference orientations. The slope at 0° orientation on average increased by 0.003 in learning mice and by 0.001 in non-performing mice, but this was not statistically significant in either case (Wilcoxon statistic=7.0, p=0.3 learning, statistic=1.0, p=0.5 non-performing). Slope change at 0° also did not statistically differ across learning status (Mann Whitney U statistic=15.0, p=0.38). This result was inconsistent with our theory-driven hypothesis. We next performed some data-driven exploratory analysis.

Figure 3.15 revealed orientations on OTCs where positive slope change was both larger in magnitude and more consistent across all learning mice, in particular at ±20 degrees. Average slope changes taken at -20 or 20 degrees orientation were 0.006 to 0.009 units (response/degree) for learning mice (statistically different from 0, Wilcoxon statistic=0.0, p(adjusted)=0.031). Conversely, slopes were -0.003 for non-performing mice (NS). These slopes were significantly different across learning status (Mann Whitney U=21.0, p(adjusted)=0.05). Slope change compared at 0 and 90 degrees did not differ significantly (Wilcoxon statistic = 14.0, p=1.0 for learning mice, statistic=1.0, p=0.5 for non-performing). Given that the data, pooled across proficient training stages and areas displayed high variability across reference orientations, we did not analyse the results further by area and individual training level (these are visualised in the appendix, figures 15 and 16).

#### 3.3.4.2 Correlation of slope change with behavioural performance

We next asked whether the performance of each mouse in the behavioural task correlated with the magnitude of slope change (figure 3.17). A behaviour score was assigned to each mouse, which was fraction total correct trials out of all trials, across all sessions. For this analysis, neurons were pooled from all proficient stages for each mouse. This resulted in a single summary delta slope value per mouse. The delta slope results for 20 and -20° orientations were used here. There was no significant correlation for any area. Although coefficients were positive for areas V1 and LM, this is best explained by the grouping of slope changes by

learning status. We also confirmed that the slopes measured at the naive stage alone did not predict behaviour score (appendix, figure 17).

#### 3.3.4.3 Slope change was not specific to neurons with spatial receptive fields in the trained location

Finally, we asked whether slope change was a function of distance from the trained location centre (figure 3.18). We analysed neurons that had preferred orientations near 0 degrees, had receptive field data and suprathreshold receptive field Z-score values. Here, we compared naive with proficient slope distributions, with all proficient stages pooled for learning mice. Sample sizes of neurons with suprathreshold OTC SNR and RF Z-scores with valid preferred orientation split by RF spatial locations were not sufficient for a more fine grained analysis over each visual area or training stage.

We firstly performed a simple analysis, asking whether there was a difference in average slope change for neurons inside versus outside the trained location in learning mice. The mean slope change was positive for both trained and untrained spatial locations (mean slope change was 0.001 ± 0.0005 and 0.002 ± 0.0018 degrees within trained and non-trained locations, respectively). Within each mouse, there was no difference between slopes at the trained and non-trained locations (Difference = 0.0006 ± 0.002, Wilcoxon statistic=16.0, p=0.82).

We also explored whether Euclidean distance from the trained location center was associated with degree of slope change. Neurons were grouped into bins according to the Euclidean distance of their receptive field centre coordinate from the trained location centre, and the slope change within each distance bin was calculated relative to the naive level for each mouse (figure 3.18a). Across mice, this visualisation did not reveal a clear pattern in slope changes across Euclidean distance. The Euclidean distance is direction invariant. In figure 3.18b, we also calculated slope change for neurons within squares on a 2D spatial grid, preserving information about direction in altitude and azimuth axes. Again, slope changes did not appear to be localised to neurons preferring the trained location, which is marked with a cross in figure 3.18b. Only spatial locations that were sampled adequately were included in these analyses. Specifically,

there had to be at least 10 plus 10 filtered neurons for naive plus proficient stages per mouse per spatial location, for a data point to be included. Hence, due to insufficient sampling of more distal locations in visual space, the locations analysed were restricted to a relatively central portion of the left visual field (no further than 50 degrees in azimuth).

#### 3.3.4.4 Summary of slope changes

Overall, the slope results suggest that slopes increased across a broad range of orientations, and this was not specific to 0°, the trained orientation. Slopes increases were instead most consistent at ± 20° orientations. This change was dependent on task performance, not just task experience. Slope change across broad ranges of orientations could point to a general sharpening of OTCs. To confirm this, we next examined OTC bandwidth changes.

#### 3.3.4.5 Change in orientation tuning curve bandwidth

We examined changes in the OTC bandwidth distributions before and after task training. We first compare naive and pooled post-training stages, then go on to provide a more detailed analysis by specific training stage, spatial location and preferred orientation. All OTCs were included in this analysis, after filtering for OTC SNR as described earlier. Because we compared distributions, for every analysis a minimum of 10 filtered neurons were required at both the naive stage and the training stage being compared. Interestingly, the median OTC bandwidth at the naive stage was lower for non-performing mice than for mice that learned the task (figure 3.19). Combining data from all visual areas, the median bandwidth for non-performing mice was lower by an average of 3 degrees, and this difference was statistically significant (Mann Whitney U: 1.0, p= 0.033).

The median bandwidth across all 4 areas combined decreased after training within mice that learned the task by an average of 2.67 degrees (comparing naive and pooled proficient stages of any preferred orientation). Conversely, median bandwidth in nonperforming mice increased by an average of 0.22 degrees. The difference in median bandwidth change distributions between learning and non-performing mice was statistically significant according to a Mann Whitney-U

test (U=0.0; p=0.0167; two-sided test). Figure 3.19 shows the break-down of naive-proficient and naive-non-performing bandwidth changes split by visual area.

For a more holistic analysis of bandwidth distribution changes beyond the median values, figure 3.20 presents the differences in kernel density estimates (Δ KDEs) between the naive and post-training stages, split by learning status (proficient or non-performing). KDE change reflects the change in the probability density of observing bandwidths at specific values, providing insight into the local changes within the distribution. Positive values indicate an increase in the probability of observing specific bandwidths after training, while negative values represent a decrease.

In learning mice, the Δ KDE curves show a consistent leftward shift in the bandwidth distribution across all 4 visual areas at the proficient stage, indicating that neurons tend to have narrower bandwidths after training. In contrast, non-performing mice exhibited a rightward shift, suggesting broader bandwidths, except for in area AL. Changes in the cumulative distribution functions are plotted in appendix figure 18.

Median bandwidth changes split by training level and visual area are displayed in figure 3.21. This shows that bandwidth decreases were maintained across levels of training for each learning mice. With filtered data split by both level and area, sample sizes of neurons contributing to each data point were small. As a result, there was no clear pattern of further change after task acquisition that was consistent across mice. To improve sample sizes, we again pooled neurons from all 'late learning' stages.

We next asked whether a mouse's overall performance score in the behavioural task had a linear relationship with the magnitude of median bandwidth change, pooled across all post-training time-points. As all areas displayed bandwidth decreases on average in learning mice, we initially combined data across all 4 areas. All learning and non-performing mice were included in this analysis. A Pearson correlation analysis revealed strong, statistically significant negative correlation coefficient (R=-0.65, P=0.044, figure 3.22). We then split the data by area to ask whether this was true for all areas. For V1, LM and LI, R values were of similar magnitude:

-0.64, -0.62 and -0.70 respectively. However, these correlation coefficients were not statistically significant after multiple comparisons correction, using a false discovery rate of 5% (P(adjusted)= 0.11, 0.11 and 0.25 respectively). For area AL, the correlation coefficient was weaker (R=-0.34, P (adjusted) = 0.40). Both the neural and behaviour score variables measured at each training stage individually are plotted in appendix figure 19.

Overall, it is not clear that the score/bandwidth correlation exists when considering data from either learning status individually. Indeed, in the 2D space of behavior score versus delta bandwidth (Figure 3.22), the two clusters of data points, grouped by learning status, can be linearly separated by a 1D line along either axis. This raises the question of whether the relationship is linear (where the magnitude of the behavior score corresponds to the magnitude of delta bandwidth) or binary (where the learning status is associated with a categorical outcome: whether or not there is a decrease in delta bandwidth). However, the sample size of mice from each learning status group is too small to answer this confidently.

We next asked whether the change in a mouse's age, i.e. the number of days elapsed since the naive imaging time-point and subsequent imaging time-points could account for bandwidth changes, or the differences observed between learning and non-performing mice. The purpose of this was to ensure that an age difference did not account for the results, in case OTC properties change with age or passive sensory experience whilst in their cages. The bandwidth distributions for learning and NP mice did tend to converge towards a mutually similar median value for the results shown in figure 3.19. The average age change between naive and non-performing imaging time-points was 73 days for non-performing mice (80 for SA051, 82 for SA052, 58 for SA047). For learning mice, the mean age change up to proficiency for stages 45, 35, 30, 25 and 20 degrees was 63, 70, 73, 85 and 88 days, respectively. Non-performing mice therefore aged similarly to learning mice over the course of training. In figure 3.24, at the 45 degree training stage, where sample size of mice was the largest, the age differences of learning mice were quite variable, ranging from 37 to 89 days. However, most mice already displayed bandwidth decreases. There did not appear to be a linear relationship between age change and delta median bandwidth here, and no correlation coefficient for any stage approached statistical significance

(R and P values shown on subplots in 3.24). Overall, the passage of time or aging alone does not seem to predict changes in the OTC bandwidth distribution.

We performed some further analyses to ask whether any other variable that could change over time might also relate to change in median bandwidth, explaining away any potential dependence on task learning. For example, with months elapsed between imaging time-points when studying long-term learning, the clarity of the cranial window and thus the recording quality can degrade.

Firstly, we found that the correlation of figure 3.22 was robust to band-pass filtering the bandwidth distribution strongly, so that more extreme high and low bandwidth values were excluded from the calculation. For example, selecting only cells with a bandwidth between 15 and 30 removed the tails of the bandwidth distribution entirely, yet the correlation was still strongly negative (R=-0.76,P=0.011). This is consistent with the kernel density changes plotted in figure 3.20, which shows changes in the main body of the distributions. A Steiger's Z test for differences in dependent correlation coefficients showed the result for the original versus filtered bandwidth distributions did not differ significantly (T=0.48,p=0.64). This made it less likely that the result was due to an increase in noisy, non-neuronal or non-selective ROIs that may have remained in the dataset after filtering by OTC SNR and imposing a minimum bandwidth constraint. This would present in the dataset as neurons in the tails of the bandwidth distribution, because fitting the Gaussian process regression to highly noisy ROIs results in either over-fit (bandwidth at lower-bound) or highly smooth (high bandwidth) fitted tuning curves, depending on the length scale bounds used.

We also confirmed that the correlation of figure 3.22 was robust to fitting tuning curves with different parameters, namely higher length scale bounds for the optimizer (20 to 40), which forced each neuron to have a smoother tuning curve. Here, the correlation coefficient was -0.81 (stronger by -0.16) with P=0.005. A Steiger's Z test for differences in dependent correlation coefficients showed the result for the two sets of length scale bounds did not differ significantly (T=0.74,p=0.48).

We next asked whether correlation of other variables with bandwidth change was present (figure 3.25). Firstly, the average neuronal response to the blank stimulus did not correlate with behaviour score, excluding the possibility that a reduction in background activity could explain the bandwidth decrease (figure 3.25a, R=0.37, P(adjusted)=0.70). In addition, OTC SNR did not correlate with bandwidth change (figure 3.25b, R=0.26, P(adjusted)=0.70). Finally, the average amplitude of OTCs before normalization did not correlate with bandwidth change (figure 3.25c, R = 0.14, P(adjusted) = 0.71).

We next explored the distribution of bandwidth changes in retinotopic space, for neurons that had receptive field data and suprathreshold receptive field Z-score values. Here, we compared naive with proficient bandwidth distributions, with all proficient stages pooled for learning mice. Sample sizes of neurons with suprathreshold OTC SNR and RF Z-scores split by RF spatial locations were not sufficient for a more fine grained analysis over each training stage.

We firstly performed a simple analysis, asking whether there was a difference in average bandwidth change for neurons inside versus outside the trained location in learning mice. Bandwidth change was negative on average in all spatial locations: the mean bandwidth change for the trained location was -3.9 degrees and was -4.6 degrees for non-trained locations. Within mice, on average the difference between non-trained minus trained location was -0.93 ± 0.37 degrees. This was not statistically significant (Wilcoxon statistic=2.0, pvalue=0.094, N=6).

We also explored whether Euclidean distance from the trained location center was associated with degree of bandwidth change. Neurons were grouped into bins according to the Euclidean distance of their receptive field centre coordinate from the trained location centre, and the bandwidth change within each distance bin was calculated relative to the naive level for each mouse (figure 3.26a). On average, bandwidth decrease occurred in all distance intervals, and there was no difference between the different intervals for any mouse. The euclidean distance is direction invariant. In figure figure 3.26b, we also calculated bandwidth change for neurons within squares on a 2D spatial grid, preserving information about direction in altitude and azimuth axes. Again, changes did not appear to be localised to neurons preferring the trained location, which is marked with a cross in figure 3.26b. Figures 3.27 and 3.28 show the euclidean

distance and spatial grid analyses of bandwidth change grouped by visual area. Again, changes were not localised to neurons preferring the trained location for any one visual area.

Only spatial locations that were sampled adequately were included in these analyses. Specifically, there had to be at least 10 plus 10 filtered neurons for naive plus proficient stages per mouse per spatial location, for a data point to be included. Hence, due to insufficient sampling, the locations analysed were restricted to a relatively central portion of the visual field (no further than 60 degrees in azimuth).

We finally compared the behaviour score/ bandwidth change Pearson correlation coefficients between the bandwidth dataset taken for neurons with spatial receptive fields inside and outside of the trained location. The correlation coefficients for both locations were almost identical (R=-0.7482 and R=-0.7485). A Steiger Z-test (Steiger 1980) naturally then indicated that the differences in correlation coefficients were not statistically significant (T=-0.001,P=0.999).

Overall, there was no evidence from the analysis of our dataset that indicated changes in bandwidth were specific to neurons whose receptive field centers were inside, or overlapping with, the trained location. Changes appeared to be diffuse across the spatial locations in the central-left visual field that were sampled adequately.

Finally, we asked whether changes in bandwidth were a function of preferred orientation (figure 3.29). This allowed used to ask whether changes related to the trained orientation or other task-relevant orientations. We took a sliding-window approach to grouping cells by their preferred orientation: a window size of ±16 degrees around each preferred orientation was used to compute bandwidth changes. This ensured that an adequate sample size of neurons contributed to each point on the preferred orientation axis for a given mouse. The analysis suggested that bandwidth change was a function of preferred orientation, and the direction of change differed according to learning status. Specifically, in learning mice, bandwidth decreases were greatest for neurons preferring orientations that were seen on correct trials only (inside the range of -45 to 45 degrees). This was most prominent for neurons preferring angles close to the trained orientation (0°) where the median bandwidth change was -4.98° (significantly different from 0 change, Wilcoxon statistic

= 0.0, p(adjusted)=0.022). This bandwidth change for PO near 0° differed statistically within-subject from neurons with PO near 90° (Wilcoxon statistic=0.0, p(adjusted)=0.022). Conversely, for non-performing mice, the function over PO was a similar bell shape around 0° but inverted in direction - bandwidth change magnitude seemed to subtly increase around the trained orientation but this change was not statistically significant from 0 within-mouse (Wilcoxon statistic = 0.0, p=0.25). However, there was a statistically significant difference of 8.8 degrees between bandwidth change values at preferred orientation = 0 degrees compared between learning and non-performing mice (Mann Whitney U=0.0, P(adjusted) = 0.022).

At preferred orientation = 90 degrees, the difference between learning and non-performing mice was much smaller (-1.84 degrees) but was still statistically significant (Mann Whitney U=0.0, P(adjusted) =0.02). In addition, bandwidth was significantly lower at POs near 90° for learning mice (Wilcoxon statistic =0.0, p(adjusted=0.02). The average bandwidth at PO=0 degrees appeared to decrease by 3 degrees when early learning (the 45 degree stage alone) and subsequent late learning stages were considered separately (appendix figure 21).

To visualise these changes, normalized OTCs were averaged for mice with preferred orientations around 0 degrees (figure 3.30). This was consistent with the measured small but significant increases in bandwidth for proficient neurons in learning mice, and subtle bandwidth increase for non-performing mice. In addition, the region of the OTCs with slopes around ±20 degrees orientation (grey shading in figure 3.30) appeared to be the major area responsible for bandwidth changes, consistent with the previous analysis of slope changes.

The data for each mouse individually is plotted and split by visual area in figure 3.31 (all proficient stages were pooled here) and training stage in figure 3.32 (all areas were pooled here). Overall, this shows that changes were most consistent across mice in data from areas V1 and LM. A pattern was not obvious for LI data taken alone, and was relatively inconsistent in area AL alone, but there were smaller samples of neurons for each preferred orientation bin for each mouse. Figure 3.32 suggests that the pattern was consistent when considering individual training stages.

### 3.3.5 Selectivity-dependent changes in orientation tuning curves

Our dataset enabled us to reproduce the analysis of a recent study that compared area V1 orientation tuning in mice before and after a static orientation discrimination task (Failor *et al.* 2025). Static refers to the fact that grating stimulus orientation was fixed within a trial. In the aforementioned study, after task acquisition, OTCs were shown to be suppressed at the two cue orientations but predominantly in neurons with low selectivity. Here we removed our OTC SNR constraint in order to allow neurons with very low selectivity to be included, in line with the original study. After repeating the analysis, we did not observe a similar suppression of average OTCs for learning mice in any visual area at the 45 degrees or other cue orientations in area V1, LM or AL. Averaged OTCs by selectivity, training stage and visual area are shown in figures 3.33, 3.34, and 3.35.

## 3.4 Methods

### 3.4.1 Animal use

All procedures were carried out under license from the UK Home Office in accordance with the Animal (Scientific Procedures) Act, 1986.

Transgenic B6;DBA-Tg(tetOGCaMP6s)2Niell/J mice expressing the fluorescent calcium indicator GCaMP6s in excitatory cortical neurons were used for all experiments (Wekselblatt *et al.* 2016). This mouse line was generated from crossing tetracycline operator (tetO)-GCaMP6s mice with mice expressing tetracycline-controlled transactivator (tTa) driven by the calcium/calmodulin-dependent protein kinase type II alpha chain (CaMK2A) promoter. Mice were aged 5-8 weeks when surgery was performed. Small mice weighing under 18g were not used. Mice included in the results were male, as their generally larger size seemed to enhance performance in the motor aspect of the behavioural task. Mice had access to food ad libitum, and were housed in groups of up to 4 in ventilated cages with a 12-hour light/dark cycle. Mice were not used for any other study.

### 3.4.2 Surgery for cranial window implantation

Large, chronic cranial windows were implanted over visual cortex of the right cerebral hemisphere with aseptic stereotactic surgery. Materials, instruments and saline were purchased as sterile if single-use, or were otherwise autoclaved before each procedure.

#### 3.4.2.1 Drugs and anaesthesia

Dexamethasone (1mg/kg) was administered subcutaneously at least 2 hours before surgery, which successfully reduced brain swelling following the craniotomy. Analgesics Buprenorphine (0.1 mg/kg Vetergesic, Ceva Animal Health Ltd) and Meloxicam (5mg/kg Metacam solution for injection, Boehringer Ingelheim) were injected subcutaneously before anaesthesia. Local anaesthetic Bupivacaine Hydrochloride (0.05 ml of 5.28 mg/ml Marcain Polyamp, Aspen) was injected subcutaneously to the scalp, before it was sterilised with chlorhexidine gluconate and isopropyl alcohol (ChloraPrep). Mice were kept on an integrated heat mat throughout surgery with the head secured in a stereotactic frame. Inhalation anaesthesia was induced using 3% isofluorane (TEVA UK), and following loss of the hindlimb pedal withdrawal reflex, anaesthesia was maintained with 1.5 % isoflurane. Following surgery, the mouse was kept in a recovery cage placed on a heat mat for up to 2 hours. Mice were provided with strawberry jelly medicated with Meloxicam for 3 days (Metacam Oral Suspension, Boehringer Ingelheim). Mice were weighed and monitored daily whilst recovering for at least 7 days after surgery.

#### 3.4.2.2 Craniotomy

To perform the craniotomy, a longitudinal incision was made from 1mm posterior of the eyes to between the ears. Temporal muscle anterior to the right ear and occipital muscle overlying the interparietal bone was excised to expose the underlying skull. The skull surface was then cleaned with a bone scraper (Fine Science Tools). The center coordinates for an optimal window

placement over lateral V1, LM, LI and AL were +0.1mm anterior and +3.8mm lateral relative to lambda - placing the window as far posterior as possible whilst narrowly avoiding the transverse sinus, which lies approximately underneath the lambdoid sutures. Custom-machined aluminium headplates were secured to the skull with dental cement (Super-Bond K058E, Sun Medical). The headplate was angled at approximately 40 degrees in the 'roll' axis, so that the window, when placed, and thus the cortical surface, would be close to flat with respect to the headplate. A 4mm circular craniotomy was drilled with a dental drill at 20,000 rpm. The area was washed in sterile saline for 3 minutes and the skull fragment was removed with forceps. Next, the dura was carefully removed. Durotomy improved 2-photon imaging clarity and importantly seemed to preserve the clarity of the cranial window for months after the surgery, reducing bone re-growth under the window glass – this was crucial for imaging after long-term learning. A dual tiered circular glass window was then pressed down onto the exposed cortex using a cocktail stick and secured with cyanoacrylate tissue adhesive (Vetbond, 3M 1469) and dental cement. The window was made from a 5mm and 4mm round cover glass (Multi Channel Systems CS-4R 640724 / CS-5R 640731), glued together with UV adhesive and cured under a UV lamp.

#### 3.4.2.3 Light blocking

Refinements to the surgery were made to improve light blocking, reducing light from the stimulus screen passing through the skull intro the microscope objective during imaging. Opaque dental cement was mixed with a small amount of black pigment powder before mixing. In addition, a black rubber gasket (RS components 749-604) was attached to the headplate with this dental cement, providing an opaque interface with the microscope objective for 2-photon imaging and also creating a well to assist water immersion.

### 3.4.3 Mapping visual areas

#### 3.4.3.1 Widefield calcium imaging

For widefield calcium imaging, the cortical surface was illuminated with a blue LED light to excite fluorescence. Emitted fluorescence from GCamP6s was collected with 4x objective lens, and captured with a sensitive CMOS camera (ORCA-Fusion, Hamamatsu C14440-20UP) following an EGFP filter cube. Frames were captured at 10 frames per second (100ms exposure); over 10 times greater than the frequency of the visual stimulation. The plane of focus was moved to de-focus just below the vasculature during recordings. This ensured a spatially continuous functional signal across the window was acquired, otherwise the segmentation of the different visual areas would be impaired due to dark regions.

Light from the stimulus screen was blocked from entering the objective using a 3D-printed sleeve and blackout material (Thorlabs BK5).

Throughout all imaging mice were head-fixed and were oriented upright with the head and eyes in a natural position in roll and pitch axes. Because the headplates were at an approximately 40 degree angle, the nosepiece of the microscope was tilted so that the 2-dimensional imaging plane was parallel with the cranial window, i.e. the window was flattened.

#### 3.4.3.2 Visual stimuli

A screen (High pixel density IPS LCD, LG LP097QX1, refresh rate 60 Hz) was placed in view of the mouse's left eye at a 30 degree angle, covering 127 degrees visual angle (azimuth). A sequence of spherically corrected visual stimuli were presented to make retinotopic maps, following the procedures described in Zhuang *et al.* 2017, using the corresponding code respository (WarpedVisualStim). Against a grey background, bars containing a checkerboard pattern, with a width of 7 visual degrees, were swept across the screen in the 4 cardinal directions, at a frequency of 0.05 Hz for the sweeps across the horizontal axis and 0.06 Hz for the vertical axis. This was repeated 12 times per direction. A photo-detector (Thorlabs) was used to record

photocurrents from a 150 pixel sync-square region in one corner of the screen alongside camera frame timings in order to synchronise recordings with stimuli.

#### 3.4.3.3 Map creation

Widefield retinotopic maps were processed using Python as follows. First, the photo-detector signal was filtered with low-pass and Haar wavelet filters and then thresholded to detect the timing of visual stimuli. Stimulus triggered average (STA) $\frac{\Delta F}{F}$ movies were then created for each of the 4 directions of visual stimulation. The activity in the 2 seconds before each stimulus was used as $F_{baseline}$. Next, a Fourier transform was applied to the time-series signal from each pixel. The signal in each pixel at the frequency of visual stimulation for the relevant stimulus direction was isolated, described by a single Fourier coefficent. Next, phase maps for each direction were produced by taking the angle between (arc tangent of) the real and imaginary parts of the Fourier coefficient. Opposite direction maps were averaged together to produce altitude and azimuth maps.

Following this preprocessing the NeuroAnalysisTools repository was used to for generating visual area borders. Briefly, information about the screen position in the mouse's visual space was combined with phase maps to produce position maps. Field sign maps were then calculated, where field sign reflects whether the representation of visual space is mirrored or nonmirrored. The field sign maps were then Gaussian filtered and binarised to create a set of patches. Patches were merged or split depending on whether they contained a single map of visual space, amongst other criteria described in Zhuang *et al.* 2017. Patches were manually labelled as V1, LM, AL or LI based on anatomical maps (Zhuang *et al.* 2017).

#### 3.4.3.4 Assigning neurons to visual areas

Each neuron from a 2-photon recording was assigned to a visual area using a macro written in ImageJ/FIJI software. For this process, a minimum of 5 vasculature markers that were present in both 2P and widefield images were manually selected and then aligned with a least squares

procedure. This transformation was used to move the widefield visual area map in to the space of the 2P image, creating a visual area map that overlapped with 2P image pixels. Overlap between the pixel coordinate for the centre of each neuronal ROI and masks encompassing the area of each visual area was then used to assign cells.

### 3.4.4 Behaviour

#### 3.4.4.1 Task trial structure sequence

The specific sequence of events in a single trial of the OD task was as follows. The screen background was set to grey throughout the session during both trials and inter-trial intervals. First, there was a quiescence period of 0.7 seconds during which the mouse had to hold the wheel still. The quiescence period was reset if the wheel movement exceeded a threshold of 10 degrees. After this, a Gaussian-windowed sinusoidal grating was presented at a fixed location on the screen. The standard deviation of the Gaussian window was 12 degrees, and the spatial frequency was 0.045 cycles per degree. The grating appeared oriented either clockwise (CW) or counterclockwise (CCW) by an angle of $\theta_{trial}$ degrees relative to the target orientation, which was 0 degrees (vertical). The value of $\theta_{trial}$ determined the precision or difficulty of the trial, where lower values correspond to higher precision and difficulty.

After a delay of 200 ms during which wheel movement was decoupled from the grating orientation, the interactive or 'closed loop' part of the trial was initiated. This was indicated by a 5 kHz, 100 ms auditory tone. During the closed-loop, wheel movements were coupled to grating orientation as described by equation 3.4.4.1. During the closed-loop window, mice had to indicate whether the grating was oriented clockwise or counterclockwise by rotating a steering wheel, which had angle $\theta_{wheel}(t)$, with their forelimbs. This was coupled to the orientation of the grating stimulus $\theta(t)$ via a wheel gain factor $\beta$.

$$\theta(t) = \theta_{trial} + \theta_{wheel}(t) \cdot \beta \tag{3.1}$$

If the grating was rotated in the correct direction i.e. by at least $-\theta_{trial}$ degrees, a water reward of 2µl was given. Conversely, if the grating was rotated in the incorrect direction by at least $+\theta_{trial}$ degrees, the trial ended, a white noise auditory stimulus was played for 500 ms, and an additional time-out period was added to the inter-trial interval. The grating orientation was fixed for 1 second once it reached $\pm\theta_{trial}$ before disappearing. The trial would end after 60 seconds if no response was made and be classified as a 'time-out' response.

Incorrect trials were repeated so that one-sided performance was not rewarded. If performance on repeat trials started to increase relative to that of non-repeat trials, incorrect trials were repeated with probability 0.6-0.8 for that mouse. This was to deter the potential policy of responding the opposite way if repeating a trial after choosing incorrectly.

Each session was ended automatically if the total trials reached 500. Alternatively, the session was ended manually if the session went beyond 30-40 minutes. Sessions early on in training were made shorter, to aid habituation.

#### 3.4.4.2 Behaviour apparatus

The hardware design and parts list for the behaviour task can be found in Burgess *et al.* 2017. In our study, only one screen was used and Fresnel lenses were not applied. The behaviour apparatus and stimulus screen were built on top of breadboards (Thorlabs), and connectors were soldered to all wires so that the apparatus could be transferred between behaviour boxes and 2-photon microscope with ease. Spherical correction was applied to all stimuli to account for the angle of the screen.

#### 3.4.4.3 Software

The open-source Rigbox toolbox for MATLAB was used to program and run the experiments and record data streams Bhagat *et al.* 2019; github.com/cortex-lab/Rigbox.

#### 3.4.4.4 Water restriction

Mice were water restricted beginning no earlier than 7 days after surgery for a maximum of 14 weeks. Mice were weighed daily during this period, to maintain a target of 85% body weight. During the first week of water restriction, mice were hand-fed their water allowance daily using a 1ml syringe. This helped train the mice to lick the water reward spout in the behaviour task. It also helped habituate them to human contact, eventually being encouraged to climb onto the hands to receive water. The water reward delivery system was regularly calibrated to ensure the correct volume of water was given during training.

#### 3.4.4.5 Metrics

To calculate the entropy of the stimulus or wheel position for each trial, the following method was used. The time-series array of seen orientations or wheel positions in degrees was taken for non-repeat trials. The array was binned into equal-width intervals, and elements in each bin were counted.

To calculate Shannon entropy, the following equation was used:

$$H = -\sum_{i=1}^{n} p_i \log(p_i) \tag{3.2}$$

Where $n$ is the number of bins discretizing the data, and $p_i$ is the normalized probability of the data falling into bin $i$.

The per-session entropy value was calculated by averaging across entropy values computed for each trial individually.

### 3.4.5 2-photon imaging of neuronal activity

#### 3.4.5.1 2-photon imaging

2-photon (2P) laser scanning microscopy (Denk *et al.* 1990) was used to record fluorescence emitted by the calcium sensing protein GCaMP6s (Chen *et al.* 2013) in populations of cortical layer 2/3 excitatory neurons with single cell resolution. A Bruker Ultima 2Pplus microscope was used, with a femtosecond-pulsed Coherent Chameleon Vision-S Ti:Sapphire laser tuned to 920nm and raster-scanned across the sample using a resonant scanning galvanometer. A 16x/0.8-NA Nikon water immersion objective was used for 2P imaging. Because water gradually evaporated throughout long recordings, a 50:50 mixture of double distilled water and ultrasound gel (Anagel), spun to remove bubbles, was used for lens immersion.

5 fields of view were imaged per imaging session: V1-LM border, V1 medial, LM lateral, LI, and AL. The order of areas was chosen using a random number generating algorithm each time to mitigate any potential effect of time elapsed since behavioural task performance. An orientation tuning experiment was done first followed by retinotopic mapping with sparse noise stimuli in the same field of view. Across imaging sessions, I aimed for approximately the same imaging field of view and cortical depth.

For each recording, a single 800 x 800µm field of view (FOV) was imaged with 1024 x 1024 pixel resolution acquired at 15 frames per second. These settings were chosen to provide high-resolution in space and time with a large number of neurons. Prairieview software (Bruker) was used to control the microscope image acquisition, whilst open-source PackIO (Watson *et al.* 2016) software was used to record frame timings alongside other synchronisation signals via a National Instruments BNC 2090A board and DAQ. GCaMP6s was excited with 50mW power on sample, to avoid heating the tissue. The pockels cell (Conoptics) attenuation necessary to achieve 50 mW imaging power on sample was determined daily before imaging sessions using an optical power meter (ThorLabs PM100D).
Throughout imaging, mice were head-fixed and the custom designed head-fix apparatus was rotated so that mice were oriented upright with the head and eyes in a natural position in roll

and pitch axes. The microscope nosepiece was rotated (Bruker rotating nosepiece module) so that the 2-dimensional imaging plane was parallel with the cranial window, i.e. the window was flattened to ensure the same cortical layer could be imaged uniformly across the imaging FOV.

#### 3.4.5.2 Visual stimuli for recording orientation tuning curves

Full-field drifting gratings with temporal frequency 0.69 cycles per second and spatial frequency 0.045 cycles per degree were used and spherically corrected as in section 3.4.3.2. Each drifting grating was shown for 700ms and is referred to as one 'trial'. A grey screen with the same mean luminance as the gratings was shown in between trials, and this inter-trial interval (ITI) was sampled from a uniform distribution between 2100 and 2500 milliseconds. To align visual stimuli with 2-photon imaging frames, trial and ITI onsets were recorded with a photodetector and sync square as in section 3.4.3.2, sampled at 20 kHz. A total of 360 trials were presented for each recording, along with one 'blank' trial where no grating was shown. In each trial, the angle of the grating was randomly selected from a range of 0 to 360 degrees, in 1-degree increments. Each angle was used only once and no angles were repeated.

#### 3.4.5.3 Cell detection

Frames of each 2P recording were registered (motion corrected) and segmented to detect ROIs corresponding to putative neuronal somata using open-source Suite2P software (Pachitariu, Stringer, Dipoppa, *et al.* 2017); github.com/mouseland/suite2p. ROIs detected by the Suite2P classifier for each recording were manually curated to remove neuropil, parts of dendrites or any other pixels deemed to be non-neuronal.

At this stage, using tools provided by Suite2P, each recording was checked for registration quality, excessive motion and Z-drift.

Deconvolution of calcium fluorescence traces was performed using Suite2P. The Suite2P algorithm is based on the OASIS algorithm but with regularization steps omitted (Friedrich *et al.*

2017; Pachitariu, Stringer, Dipoppa, *et al.* 2017), discussed in Pachitariu, Stringer & Harris 2018. First, 70% neuropil subtraction was performed for the time-series fluorescence trace from each neuronal ROI according to equation 3.3, then baselining using the rolling maximum of the rolling minimum in a 60 second window was performed before spike deconvolution. The kernel for deconvolution had a timescale of 1.26s.

$$F(t) = F_{soma}(t) - F_{neuropil}(t) \cdot 0.7 \tag{3.3}$$

Where $F$ is neuropil subtracted fluorescence, $F_{soma}$ is fluorescence at the soma, $F_{neuropil}$ is fluorescence from the neuropil surrounding the soma, and 0.7 is the neuropil coefficient.

#### 3.4.5.4 Aligning imaging frames with trial timings

For each recording, deconvolved fluorescence traces were split into 361 orientation tuning trials and a variable number of retinotopic mapping trials, using trial timings that were extracted from the photodetector trace. Photodetector traces were noisy, for example due to the 60 Hz refresh rate of the screen, so had to be filtered. Filtering steps were chosen with the goal of extracting the known target number of stimuli whilst minimising any disruption to the detected onset time of stimuli. A low-pass filter was used followed by a Haar wavelet filter, which is useful for square-wave type signals. A sweep of different thresholds was then used to find the correct number of trial onset times, increasing in small increments. If no suitable threshold was found, parameters for the filters were optimised, starting from the most conservative or least amount of filtering and increasing in small increments.

For quality control of the trial timing detection algorithm, the trial timings were plotted on the raw trace for each recording and manually inspected. In addition, heatmaps of the neural responses aligned to stimulus onset times were plotted for the top 3 neurons in each recording and inspected. Finally, the distribution of the number of orientation tuned cells detected in each recording was plotted, and the left tail of this distribution was investigated to identify any issues.

#### 3.4.5.5 Fitting orientation tuning curves with Gaussian process regression

Orientation tuning curves were fit with Gaussian process regression (GPR). A non-periodic radial basis function (RBF, or squared exponential) kernel (covariance function) was used. The RBF kernel is described by equation 3.4, and is parameterised by a length scale (smoothness) parameter, $l$.

$$k(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right) \tag{3.4}$$

The length scale $l$ was optimised for each neuron using the L-BFGS-B algorithm, with the values bound between 2 and 10. To ensure the function matched up at the edges (circularity), we extended the data around its edges by 2 times the length scale upper bound used before fitting. A heteroscedastic noise kernel was used to estimate uncertainty. 500 samples were taken from the resultant function to form an orientation tuning curve (1 sample per 0.36 degrees). The Python Scikit learn sklearn.gaussianprocess library was used with the noise kernel acquired from github.com/jmetzen/gp$_e$*xtras*.

#### 3.4.5.6 Orientation tuning metrics

Orientation tuning curves were normalized by dividing each fitted GPR curve by its maximum value. Modal preferred orientation was calculated through finding the orientation value corresponding to the tuning curve peak.

Mean preferred orientation is the circular mean of a neuron's responses to different stimulus orientations (Failor *et al.* 2025). This represents the "average direction" of the neuron's tuning across all orientations and is defined as the argument of:

$$z = \frac{\sum_\theta r_\theta e^{2i\theta}}{\sum_\theta r_\theta} \tag{3.5}$$

where $r_\theta$ is the neuron's response to a given orientation $\theta$.

Orientation selectivity was defined as 1-circular variance for each cell. This is the modulus of $z$ in the expression above i.e. the length of the vector in the complex plane. This ranges from

0 to 1, and is a measure of tuning sharpness in orientation space: a selectivity value close to 1 indicates strong selectivity for a specific orientation, while a value close to 0 reflects weak or broad tuning, with the neuron responding similarly to multiple orientations.

Absolute slope values were measured over the interval of 1 sample of the normalized fitted curve, which corresponded to an orientation difference of 0.36 degrees.

The bandwidth of an orientation tuning curve was determined using the full width at half maximum, calculated with the Scipy function scipy.signal.peakwidths.

#### 3.4.5.7 Spatial receptive fields

For recording spatial receptive fields, 'sparse noise' stimuli were used. The screen was split up into a 15 x 16 grid of squares of size 7 x 7 visual degrees. For each trial, a set of probes was displayed on a grey background for 2 seconds with no ITI. Each probe corresponds to a black or white square - black squares are called 'OFF' probes whilst white squares are called 'ON' probes. Probes centers were separated by a minimum euclidean distance of 30 visual degrees. All grid locations were sampled once per sequence in ON and OFF states, and 10 different sequences were shown per experiment meaning there were 10 repeats for each grid location.

Recordings were divided into trials, each corresponding to the presentation of a single 2-second stimulus frame. In each trial, a random subset of probes, positioned at different locations on a spatial grid, was shown. For each grid location, the neuronal response was averaged across all trials where a probe was present at that location after excluding outlier trials. Outliers likely signified cases where a probe outside the receptive field was present, but a preferred probe appeared in one or more of the trials. Specifically, trials with responses more than 1.5 times the interquartile range above the third quartile (75th percentile) were defined as outliers and excluded. This resulted in a spatial grid of responses for each neuron. A grid was computed for for ON and OFF probes separately, and those maps were then summed together.

To get the RF center location coordinate, first a circular mask of radius 28 degrees was drawn around the grid square containing the maximum response (i.e. the argmax location). This was based on the assumption that there is a single valid RF center and therefore was implemented to avoid any distant noise from biasing the estimate of the center location coordinate. The responses inside the mask were then thresholded, removing responses below 33% of the maximum. To get the RF centre location, the coordinates of the suprathreshold responses were then averaged together, weighted by the corresponding response magnitude.

To estimate the size (radius in visual degrees) of the RFs, we calculated the covariance matrix ($\Sigma$) of the centered spatial coordinates within the mask, weighted by response magnitudes. This is given by:

$$\Sigma = \frac{1}{\sum_{i=1}^{N} w_i} \begin{bmatrix} \sum_{i=1}^{N} w_i (x_i - \bar{x})^2 & \sum_{i=1}^{N} w_i (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^{N} w_i (y_i - \bar{y})(x_i - \bar{x}) & \sum_{i=1}^{N} w_i (y_i - \bar{y})^2 \end{bmatrix}$$

where the weighted mean azimuth coordinate $\bar{x}$ is:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

and the weighted mean altitude coordinate $\bar{y}$ is:

$$\bar{y} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}$$

The RF radius was then estimated as:

$$\text{RF radius} = \left( \frac{\sigma_x + \sigma_y}{2} \right) \times 1.177$$

where $\sigma_x$ and $\sigma_y$ are the standard deviations in altitude and azimuth directions taken from the weighted covariance matrix.

This approach is equivalent to fitting a 2-dimensional Gaussian to the response-weighted coordinates and calculating the average half-width at half-maximum (HWHM) of the Gaussian in the altitude and azimuth directions.

The RF Z-score of each neuron $Z_{RF}$ is the Z-score at the RF center and was calculated as follows:

$$Z_{RF} = \frac{\max(RF) - \mu_{RF}}{\sigma_{RF}}$$

Where: $\max(RF)$ denotes the maximum value in the spatial grid, $\mu_{RF}$ is the mean of all trial-averaged responses in the spatial grid, and $\sigma_{RF}$ is the standard deviation over the values in the spatial grid.

#### 3.4.5.8 Statistics

1 mouse was considered as 1 statistical unit for all analysis in this study. This ensured imbalances in sample sizes of neurons did not bias results towards 1 mouse. It also meant changes had to be consistent across several mice in order to be statistically significant. Error bars in plots represent ±1SEM aross mice.

Where multiple related statistical tests were performed, each p-value was adjusted using the Benjamini-Hochberg procedure in order to control the false discovery rate at 0.05. Only adjusted p-values (p(adjusted)) are reported.

### 3.4.6 Code availability

All data was analysed and visualised with Python 3.9, and the libraries Numpy, Matplotlib, Scipy, Scikit learn, Pandas, Seaborn and iPython. Code was developed using Sublime Text and Jupyter Lab. Code is available at the Vismap repository hosted on Github. All figures were made in Python and are entirely reproducible with the code provided.

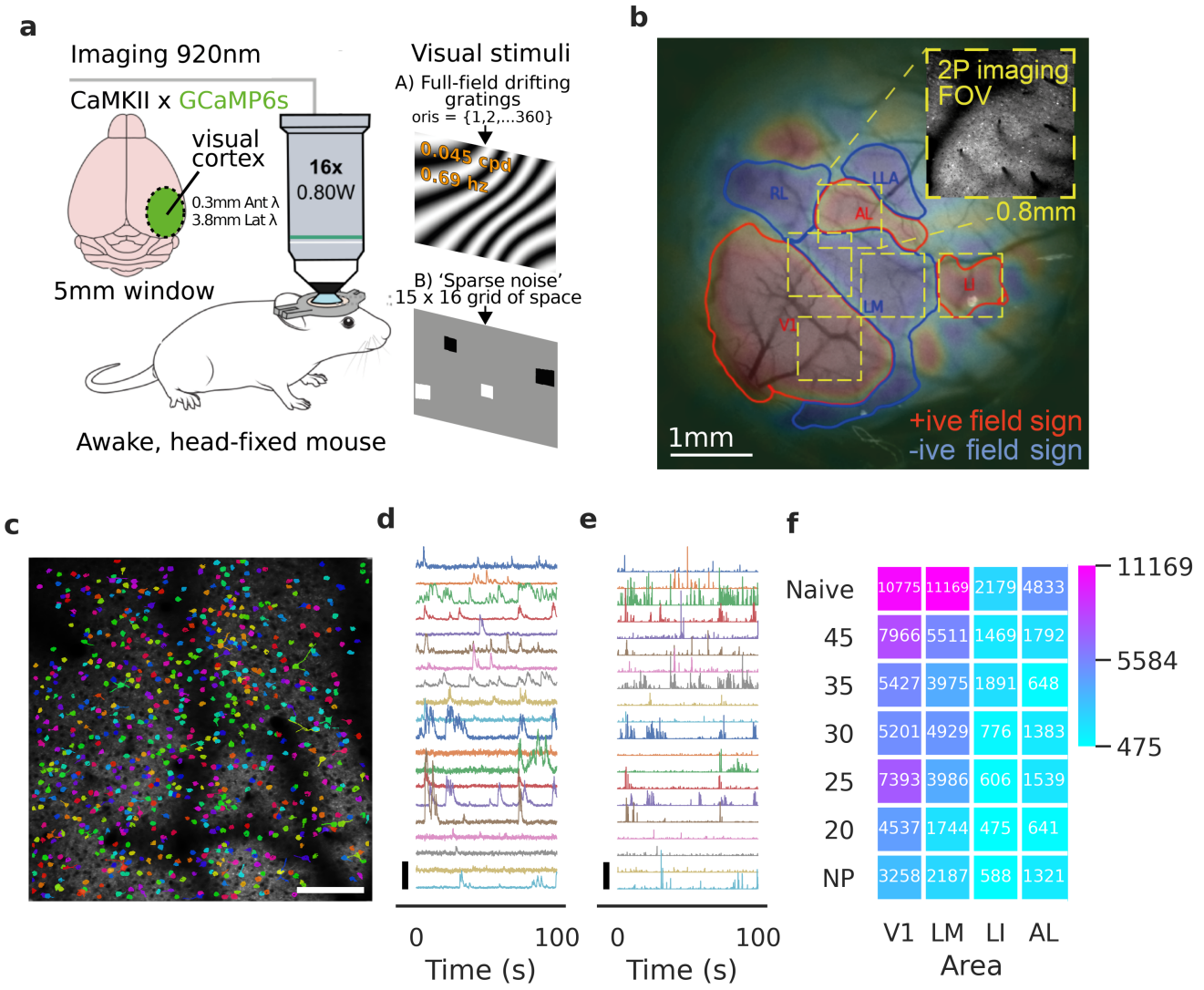



**Figure 3.7:** (**a**) Schematic of the 2-photon imaging set-up. Left: the location of the cranial window placement on the right hemisphere of a transgenic mouse brain expressing GCaMP6s. Imaging was performed during visual stimulation in awake, head-fixed mice with a 920 nm tuned laser and a 16x objective. Right: example frames from the two sets of spherically-corrected stimuli that were used - oriented full-field drifting gratings for orientation tuning experiments, or sparse noise probe stimuli for mapping spatial receptive fields in single cells. (**b**) Example widefield 1-photon image of a cranial window, showing cerebral vasculature overlaid with the field sign map derived from a 1-photon retinotopic mapping experiment. Blue colour indicates a negative field sign whilst red colour indicates positive field sign. Area borders derived from automated processing of the field sign map are displayed along with manually assigned area labels. The yellow-dashed squares indicate the approximate positioning of each 800 x 800 µm 2-photon imaging field of view used for single neuron mapping experiments. The inset shows an example 2-photon imaging field of view (128-frame average). Scale bar = 1mm. (**c**) Example mean image from one 2-photon recording, overlaid with the 770 ROIs that were detected with the Suite2P classifier. ROIs are coloured randomly. Scale bar = 200 pixels/156 µm. (**d**) Example fluorescence (DF/F) traces from 20 ROIs in c, showing a 100 second slice of the recording. Each trace is the neuropil-subtracted time-series averaged across pixels belonging to 1 ROI. Each row/colour represents a different ROI (putative neuron). Scale bar = 5 DF/F units. (**e**) Inferred (deconvolved) spiking activity for the same ROIs as in (c). Scale bar = 700 units. (**f**) Summary of the pre-processed dataset consisting of 111076 ROIs, split by training stage and visual area.

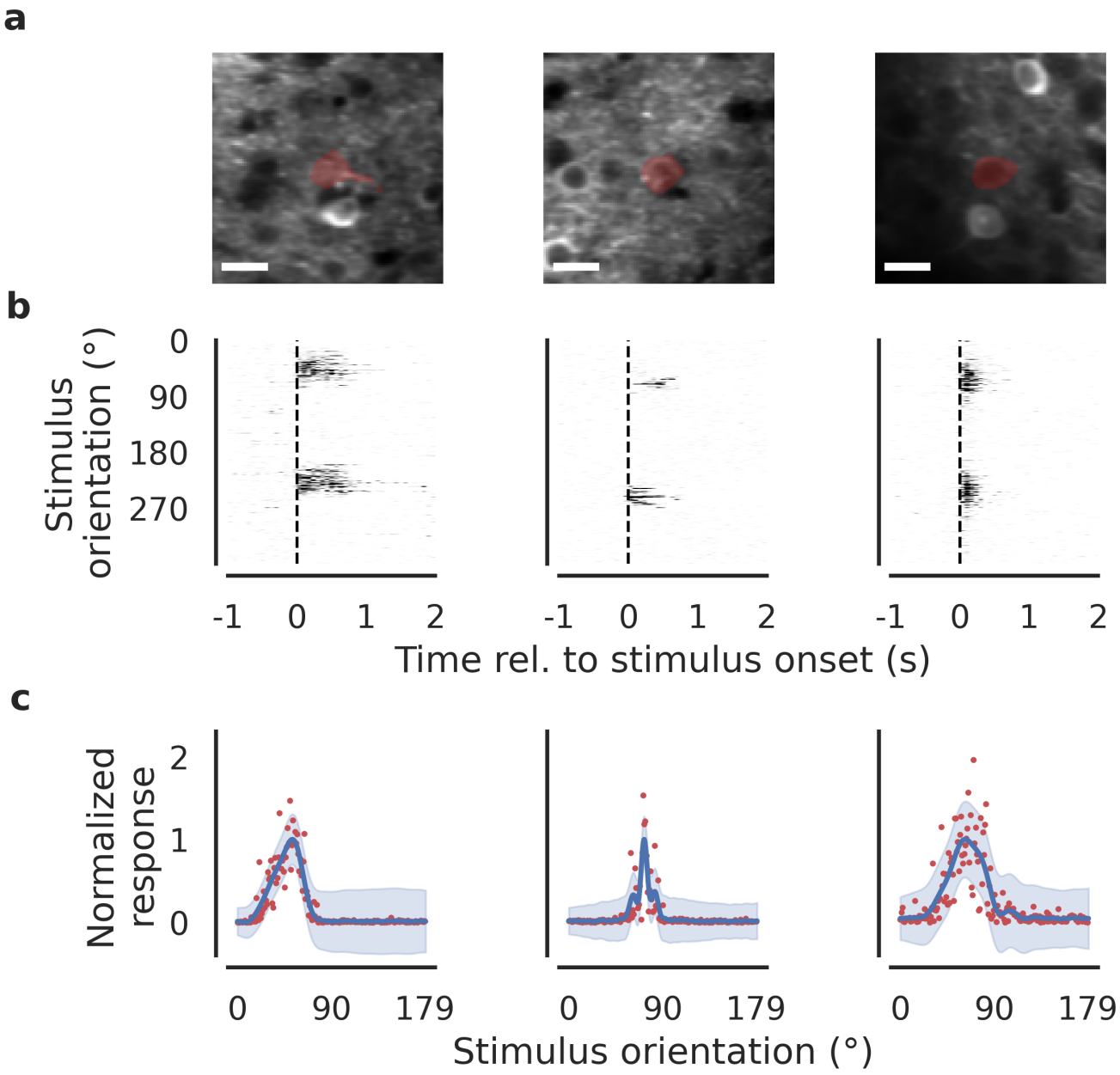



**Figure 3.8:** Example orientation tuning curve data from 3 cells. (**a**) shows an average image of each cell in greyscale, with ROI pixels detected by Suite2P overlaid in red. Scale bar = 15.6 µm (20 pixels). The left column shows a cell from V1, the middle column from LM, and the right column from LI. (**b**) displays deconvolved spike responses for each cell, split into trials aligned to stimulus onset (dashed vertical line). Trials are sorted by stimulus orientation before averaging over direction. (**c**) presents the fitted tuning curves based on the data in (**b**). The blue curve represents the function predicted by Gaussian process regression, sampled every 0.36 degrees of orientation and normalized to its maximum. The light-blue shading indicates the uncertainty estimate across the input domain - greater uncertainty occurs in flat regions despite low observational noise due to the lack of sufficient data to reliably constrain the underlying function. Red dots represent the mean post-stimulus response, direction-averaged and normalized by the same factor as the tuning curve.

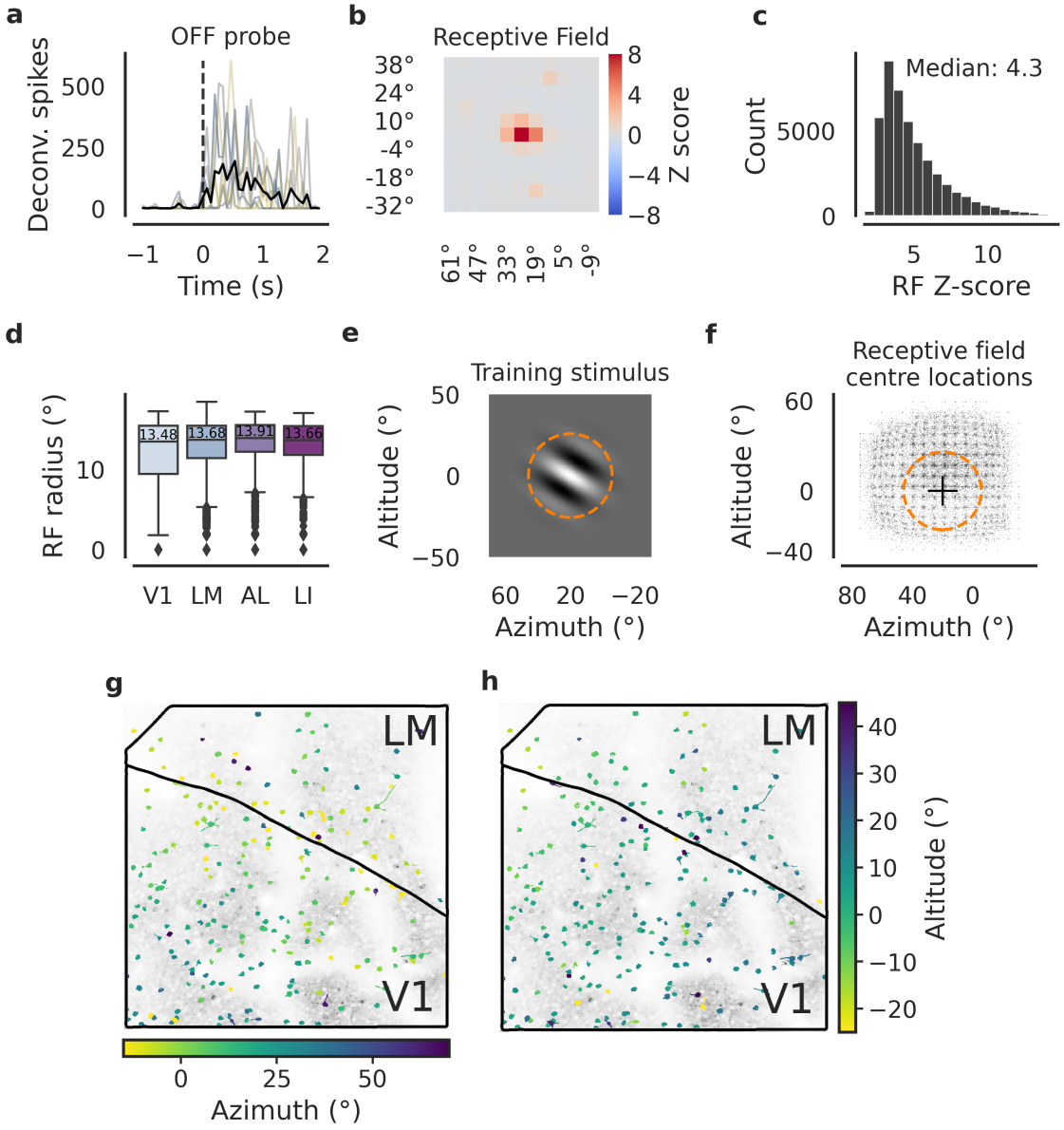



**Figure 3.9:** Spatial receptive field data. **(a)** Example deconvolved spikes traces for a single neuron during 'off' probe trials that contained the cell's preferred location (3.1° altitude, 25.7° azimuth). Thick black line = mean over trials. Dashed vertical line = stimulus onset. **(b)** a heatmap showing the spatial receptive field for the cell in (a). Colour represents the Z-scored average response at each spatial location. Each square measures 7 x 7 visual degrees. **(c)** shows the right-skew distribution of receptive field Z-score values, which is the Z-scored response at each cell's preferred grid location. **(d)** is a box plot indicating the average receptive field radius, defined as the average half-width at half-max in altitude/azimuth directions of a 2D Gaussian over the responses in (b). **(e)** shows an example training stimulus at -45° orientation. The dashed circle indicates the region we define as the trained location, which is where the Gaussian window is at 10% of its maximum. **(f)** is a map of the receptive field centre location of every neuron. The cross denotes the stimulus centre location, and the dashed orange circle outlines the trained location as defined in (e) **(g,h)** shows an example 2-photon imaging field of view that was centered over the V1/LM border. The black line denotes the estimated border separating V1 and LM, derived from the widefield retinotopic map aligned to the 2P FOV. Cells in **(g)** are coloured by their azimuth location whilst cells in **h** are coloured by their altitude location.

**a**

| Area | Total cells | OTC SNR > 2.5 (% of total) | Selectivity > 0.3 (% of OTC SNR > 2.5) | OTC SNR > 2.5 & RF Z > 3 (% of total) | RF trained location overlap | Experiments |
|---|---|---|---|---|---|---|
| V1 | 44557 | 20370 (46%) | 10018 (49%) | 13717 (31%) | 9223 (67%) | 140 |
| LM | 33501 | 17840 (53%) | 8804 (49%) | 9910 (30%) | 5718 (58%) | 133 |
| AL | 12157 | 6319 (52%) | 2852 (45%) | 3743 (31%) | 1998 (53%) | 55 |
| LI | 7984 | 3032 (38%) | 875 (29%) | 1602 (20%) | 820 (51%) | 60 |
| RL | 2289 | 974 (43%) | 393 (40%) | 578 (25%) | 327 (57%) | 31 |

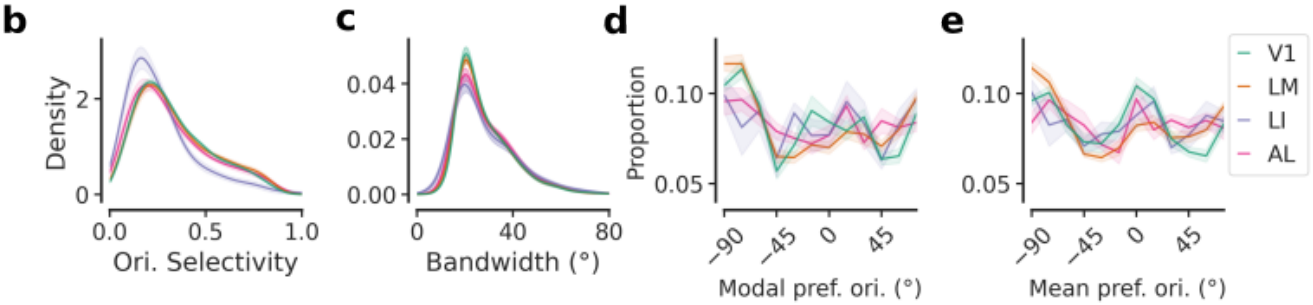



**Figure 3.10:** (**a**) A table characterising the overall orientation tuning data across visual areas, with neurons from all training stages pooled. Total cells is the number of neuronal ROIs after pre-processing and removal of OTC bandwidths below 9 degrees. OTC SNR = orientation tuning curve signal to noise ratio (3.4.5.6). [Orientation] selectivity refers to 1-circular variance (3.4.5.6). RF Z-score = spatial receptive field center Z-score (3.4.5.7). For cells with trained location overlap, the percentage of overlapping cells is expressed out of filtered cells i.e. those with supra-threshold OTC SNR and RF Z-score values.'Experiments' refers to the number of unique individual recordings that contributed to the statistics for the corresponding visual area. (**b**) Kernel density plot of the average distribution of of orientation selectivity by area, across mice (shading is ± 1 SEM). All training stages are pooled. (**c**) Kernel density plot of the average distribution of of OTC bandwidths (full width at half max) by visual area, across mice (shading is ± 1 SEM). All training stages are pooled. (**d**) Histogram plot of the average distribution of modal orientation preferences, averaged across mice. Histogram bin centres fall upon the cardinal orientations and bin centers are connected with straight lines (shading is ± 1 SEM). All training stages are pooled. (**e**) Histogram of the average distribution of mean orientation preferences, across mice (shading is ± 1 SEM). All training stages are pooled.

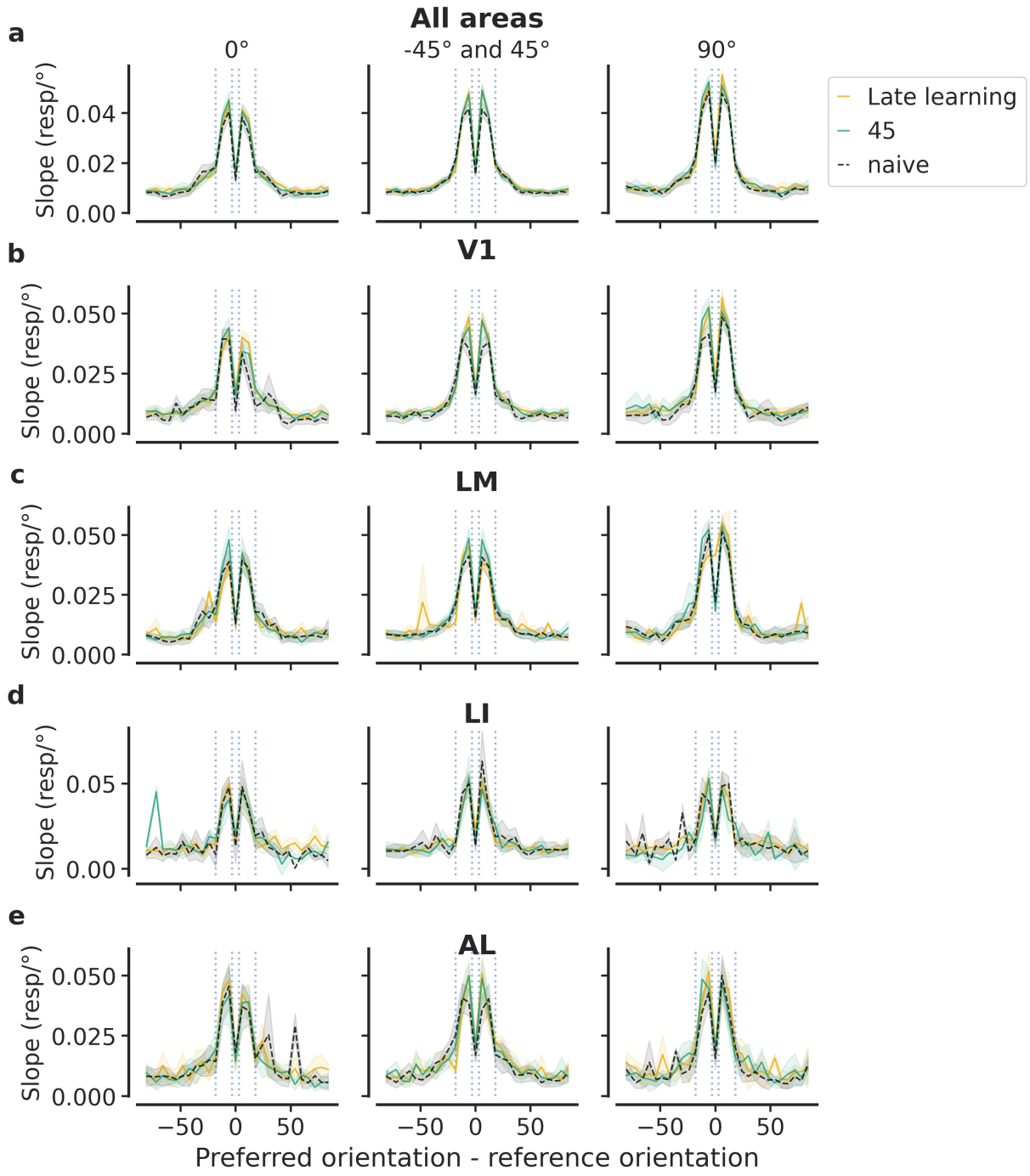



**Figure 3.11:** Visualisation of orientation tuning curve slopes at different stages of learning. For visual clarity, results from behaviour training stages 35, 30, 25 and 20 degrees were pooled together and labelled as 'late learning'. Only learning mice were included in this plot. All spatial receptive field locations were included. Slopes were taken at 3 reference orientations (columns): 0 degrees, -45 and 45 degrees combined, and 90 degrees. Slopes were binned in intervals of 6 degrees and the mean absolute slope within each bin is plotted, averaged across mice. Shading indicates ±SEM across mice. For the combined condition, two slopes per neuron were calculated then averaged. Neurons from all areas are pooled together in row (**a**), whereas the results for each visual area individually are plotted on subsequent rows (**b** to **e**).

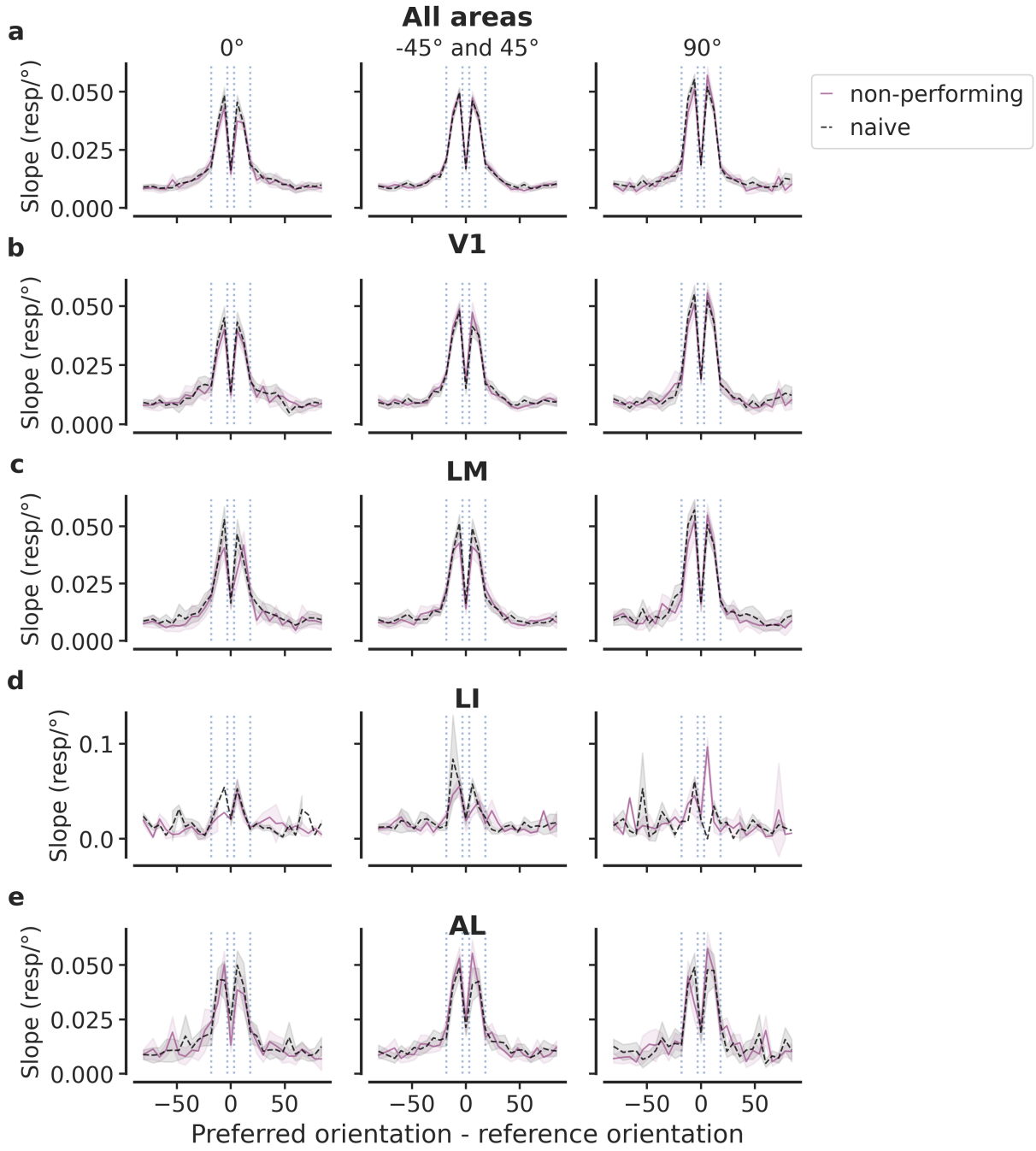



**Figure 3.12:** Visualisation of orientation tuning curve slopes at naive and post-training time-points, for non-performing mice only. 'Non-performing' indicate the the post-training imaging time-point for non-performing mice. All spatial receptive field locations were included. Slopes were taken at 3 task-relevant reference orientations (columns): 0 degrees, -45 and 45 degrees combined, and 90 degrees. Slopes were binned in intervals of 6 degrees and the mean absolute slope within each bin is plotted, averaged across mice. Shading indicates ± SEM across mice. For the combined condition, two slopes per neuron were calculated then averaged. Neurons from all areas are pooled together in row (**a**), whereas the results for each visual area individually are plotted on subsequent rows (**b** to **e**).

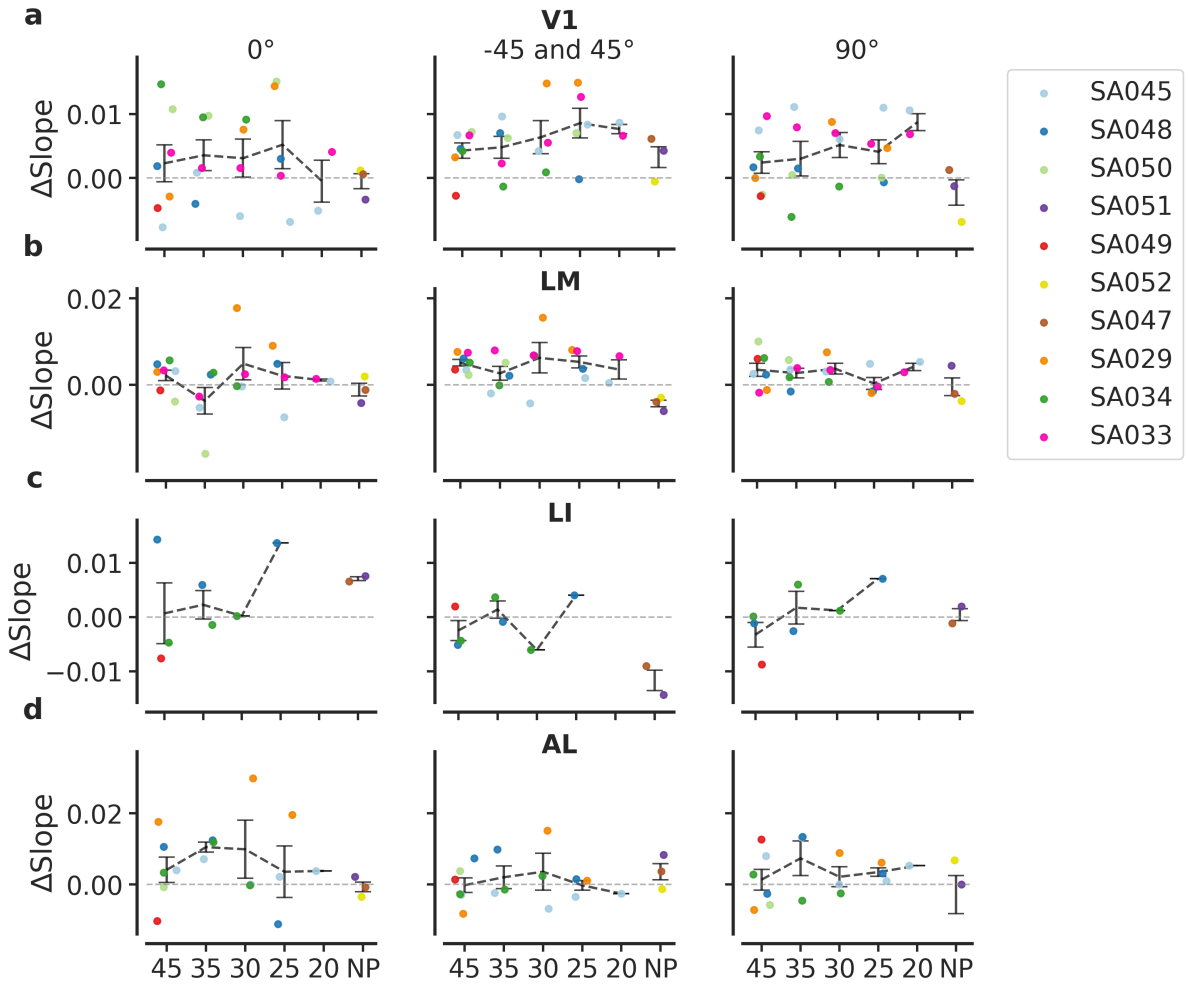



**Figure 3.13:** Average Δ slope within-each mouse, split by training level (horizontal axis of each subplot) and visual area (subplot rows **a** to **d**). Different columns represent Δ slope measured at a different orientation on each orientation tuning curve: 0°, -45 and 45°, plus 90°. -45 and 45° are combined as they are the two cue orientations in stage 1 of the behaviour task. Each coloured data point corresponds to one mouse (mouse ID is in the legend). A minimum of 10 cells were required for each training stage / area / mouse combination. Note that data in different columns comprises different neuronal populations, because slopes were only taken for neurons whose PO was between ±3 and ±18 of the given reference orientation. NP = non-performing. Error bars denote ± 1 SEM across mice)

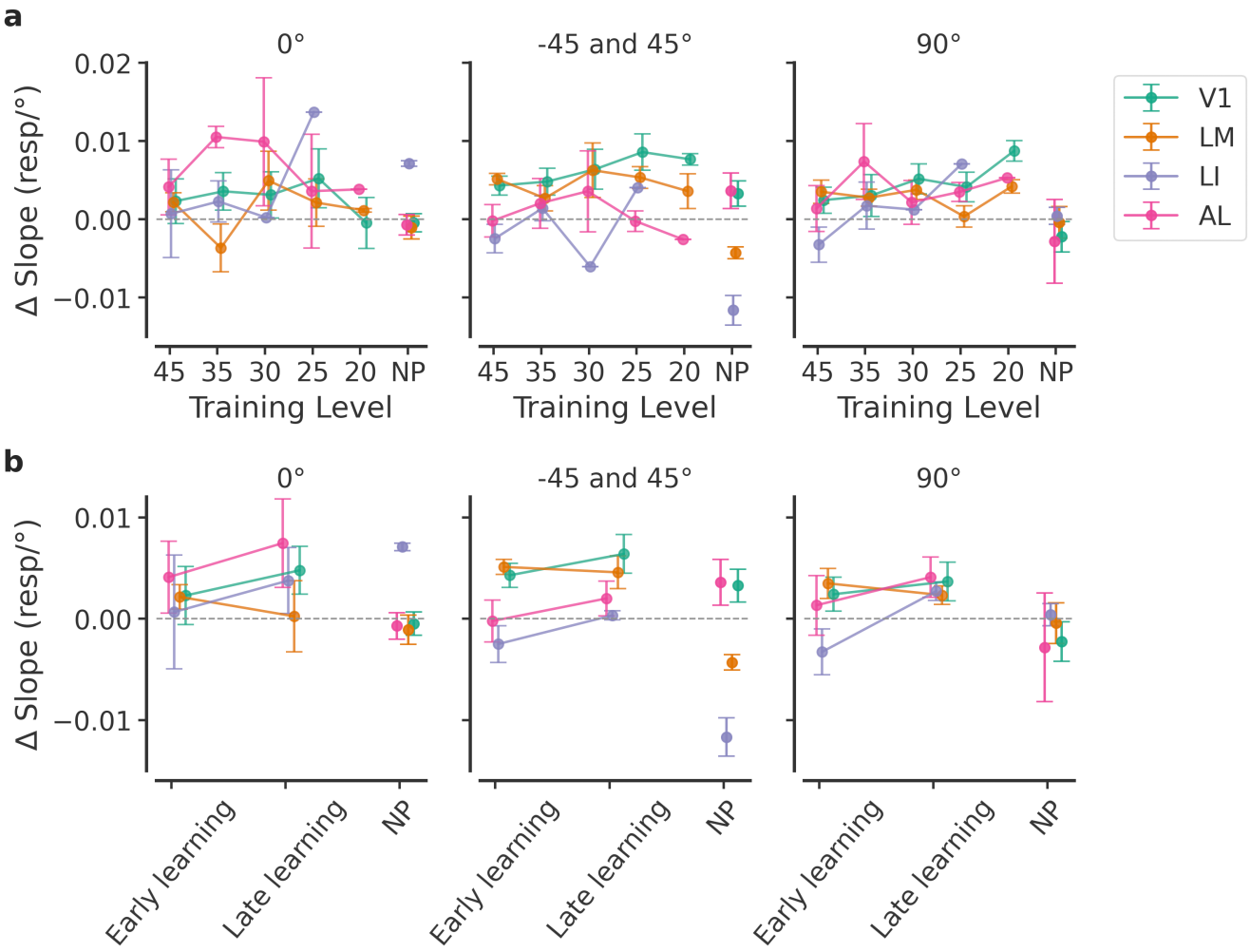



**Figure 3.14:** Summary of average Δ slopes across mice (**a**) split by individual training level (horizontal axis of each subplot) and visual area (coloured lines, see legend). In (**b**), neurons from 'late learning' stages were pooled together. 'Early learning' represents the 45° training stage alone. Different columns represent Δ slope measured at a different task-relevant orientation on each orientation tuning curve: 0°, -45 and 45°, plus 90°. -45 and 45° are combined as they are the two cue orientations in stage 1 of the behaviour task. A minimum of 10 cells were required for each training stage / area / mouse combination. Note that data in different columns comprises different neuronal populations in a given mouse, because slopes were only taken for neurons whose PO was between ±3 and ±18 of the given reference orientation. NP=non-performing. Error bars denote ± 1 SEM across mice.)

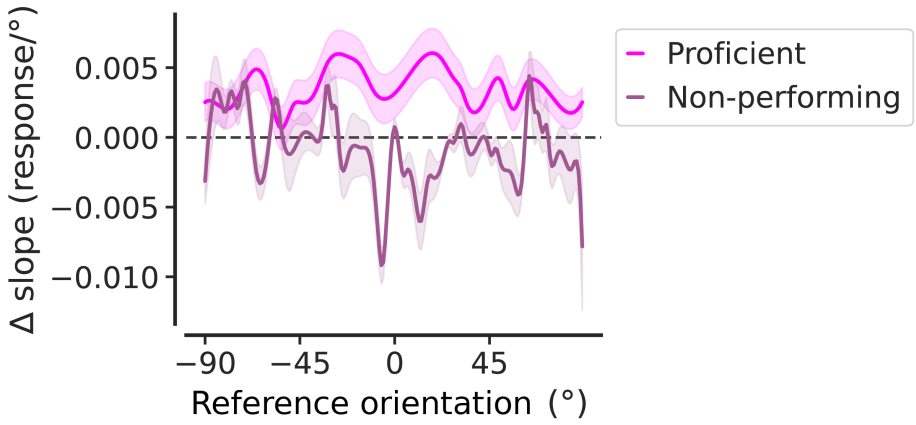



**Figure 3.15:** Average Δ slopes at different reference orientations across learning mice (all proficient stages pooled; magenta) and non-performing mice. Slopes were only taken for neurons whose PO was between ±3 and ±18 of the given reference orientation. Shading represents ±1 SEM across mice.

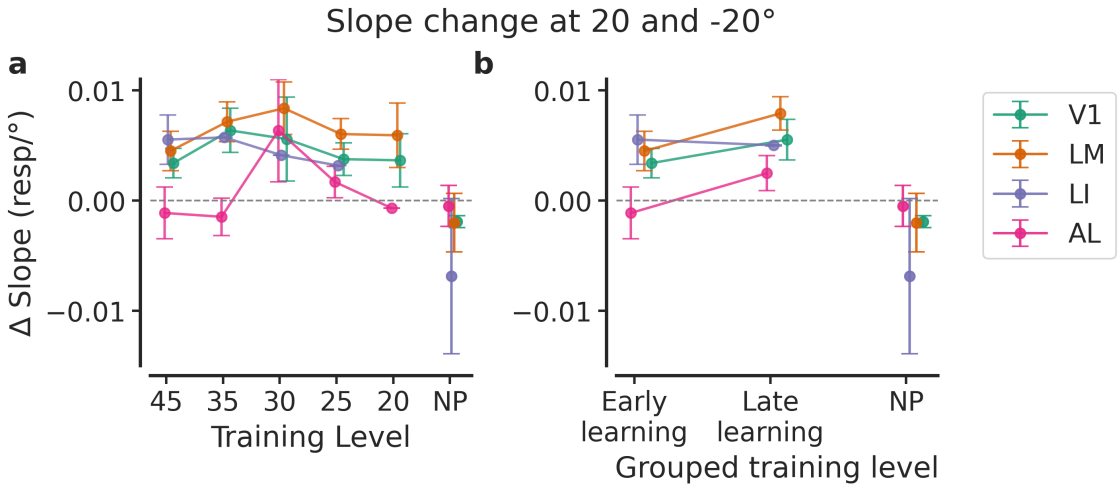



**Figure 3.16:** Average Δ slope measured at -20° and 20° degrees reference orientations on each orientation tuning curve. Slopes at V1, LM and LI consistently increased from their naive baseline, for learning mice only. (**a**) is split by individual training level and visual area (coloured lines, see legend). In (**b**), neurons from 'late learning' stages were pooled together. 'Early learning' represents the 45° training stage alone. A minimum of 10 cells were required for each training stage / area / mouse combination. PO was between ±3 and ±18 of the given reference orientation. NP=non-performing. Error bars denote ± 1 SEM across mice.

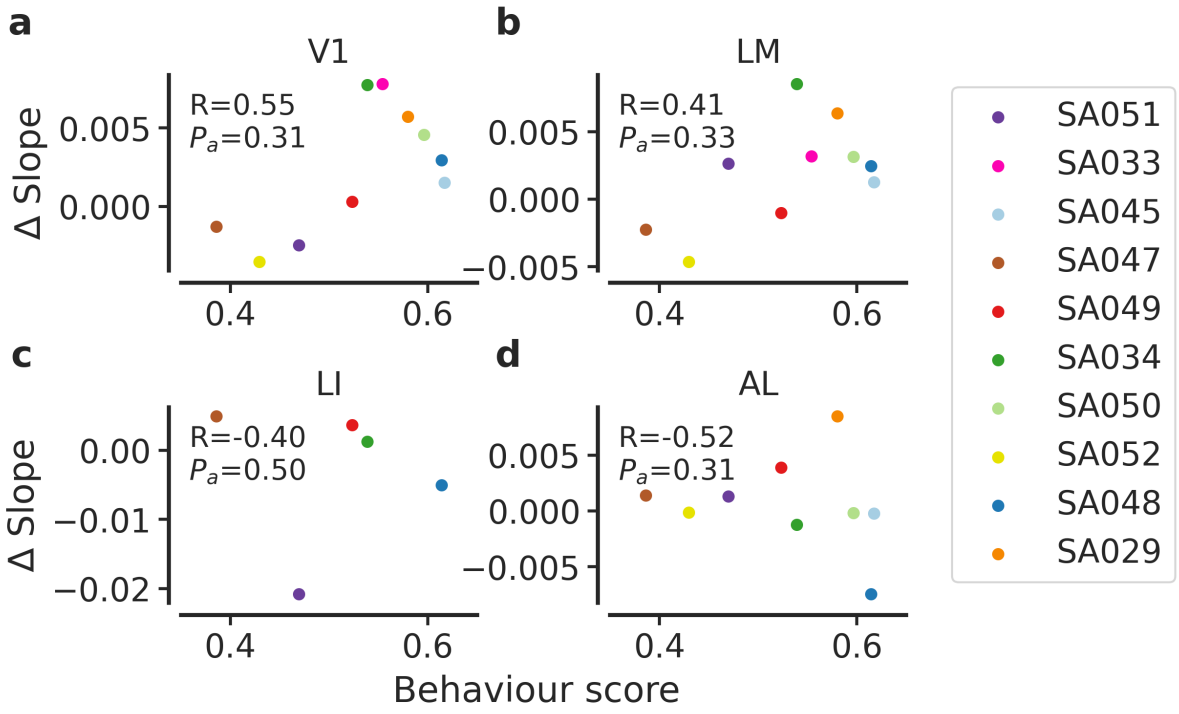



**Figure 3.17:** Correlation of Δ slope at ±20° and and behaviour score for each mouse. Behaviour score was the fraction of correct (rewarded) non-repeat trials across all training. Each coloured data-point corresponds to one mouse. The legend shows the mouse ID. R and $P_a$ values are from correlation analysis. $P_a$ signifies multiple-comparisons corrected p-values.

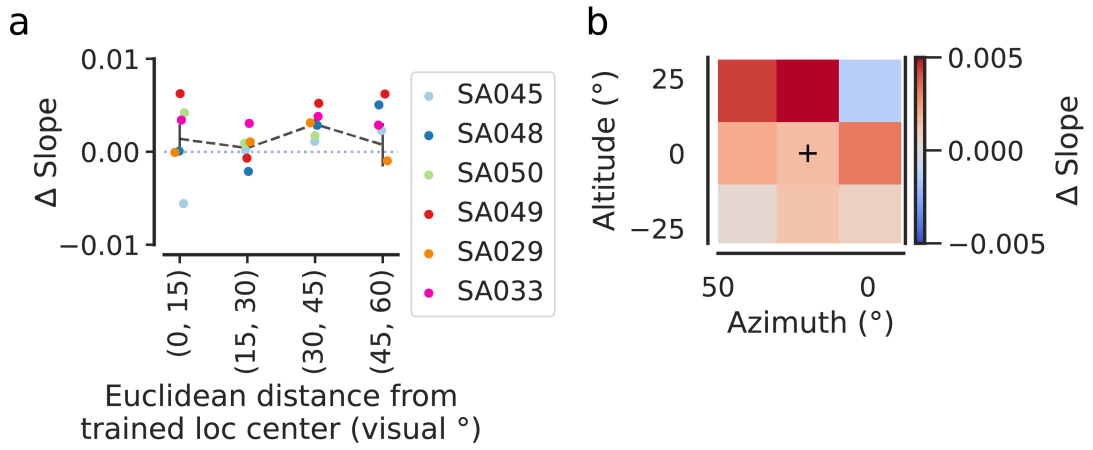



**Figure 3.18:** Median slope change at ±20 degrees orientation as a function of receptive field centre distance from the trained location. Only learning mice are included here, comparing naive and proficient (pooled) training stages, with all visual areas combined. Only neurons whose PO was between ±3 and ±18 degrees were included. **a** Slope change for neurons in intervals of Euclidean distance of their receptive field centre coordinate from the trained location centre. Slope change within each distance interval was calculated relative to the naive level for each mouse in the same distance interval. Black dashed line shows the average change across mice. Error bars denote ± 1 SEM.**b** Slope change for neurons within squares on a 2D grid of visual space. Each square represents the mean value across learning mice. For both subplots, a minimum of 10 cells at naive and proficient stages was required before including a data point for 1 mouse.

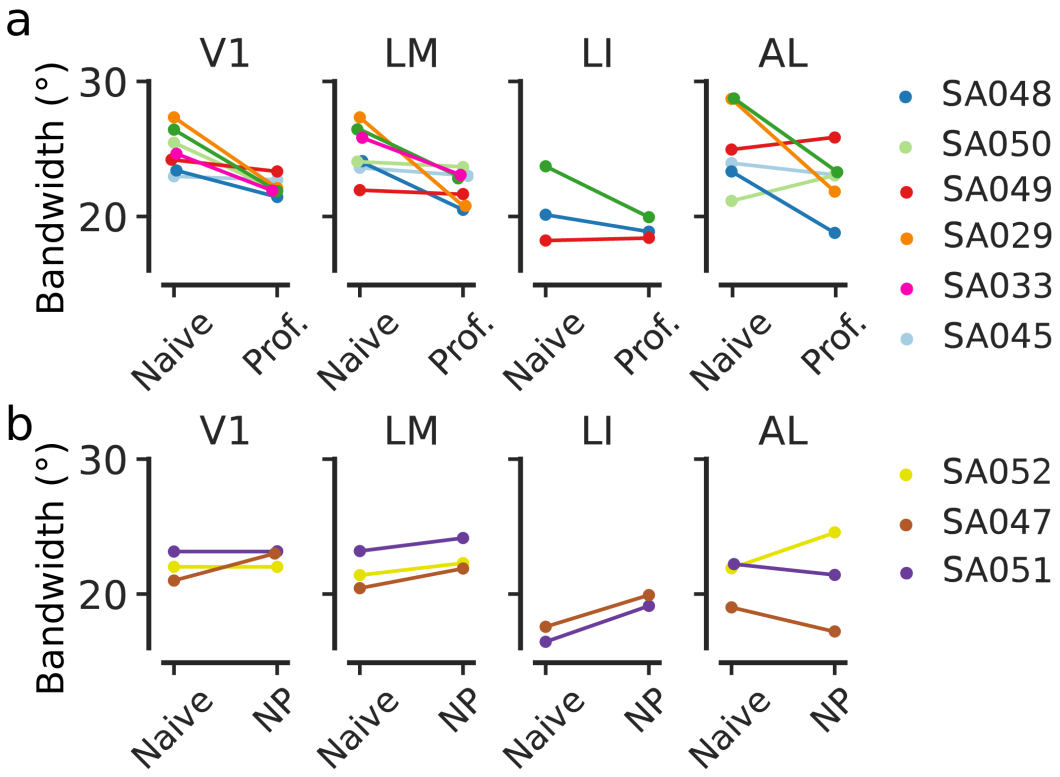



**Figure 3.19:** Orientation tuning curve (OTC) median bandwidth change from the naive stage to subsequent stages of training (**a**) for learning mice, (**b**) for non-performing mice (NP). Prof. = proficient (neurons measured at all post-learning stages are pooled together). The stage marked as NP corresponds to the second imaging time-point after task training for non-performing mice.

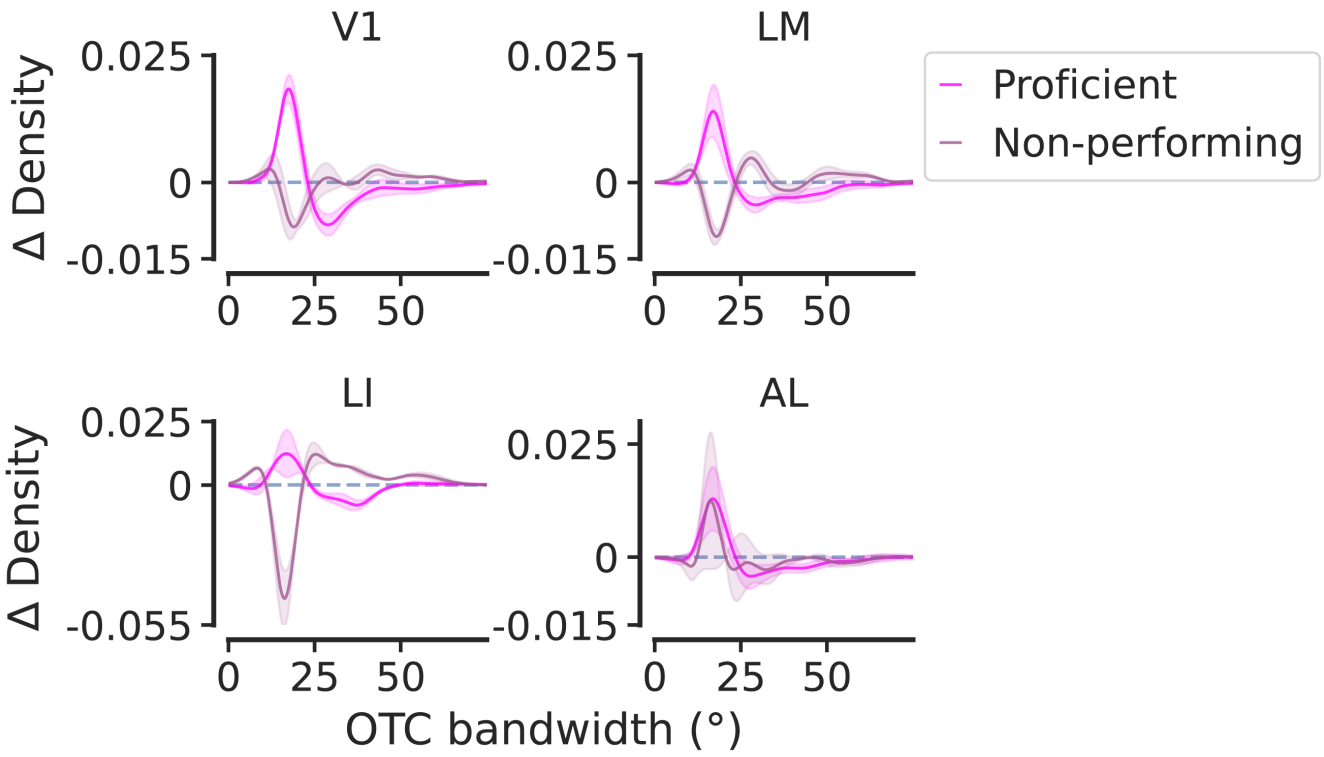



**Figure 3.20:** Difference in kernel density estimates (Δ density) between the naive stage and subsequent training stages, split by visual area (V1, LM, LI and AL). Results are shown for proficient stages in learning mice (magenta lines) or for non-performing mice (purple lines). Each line represents the average Δ density across mice, shading is ± 1 SEM across mice.

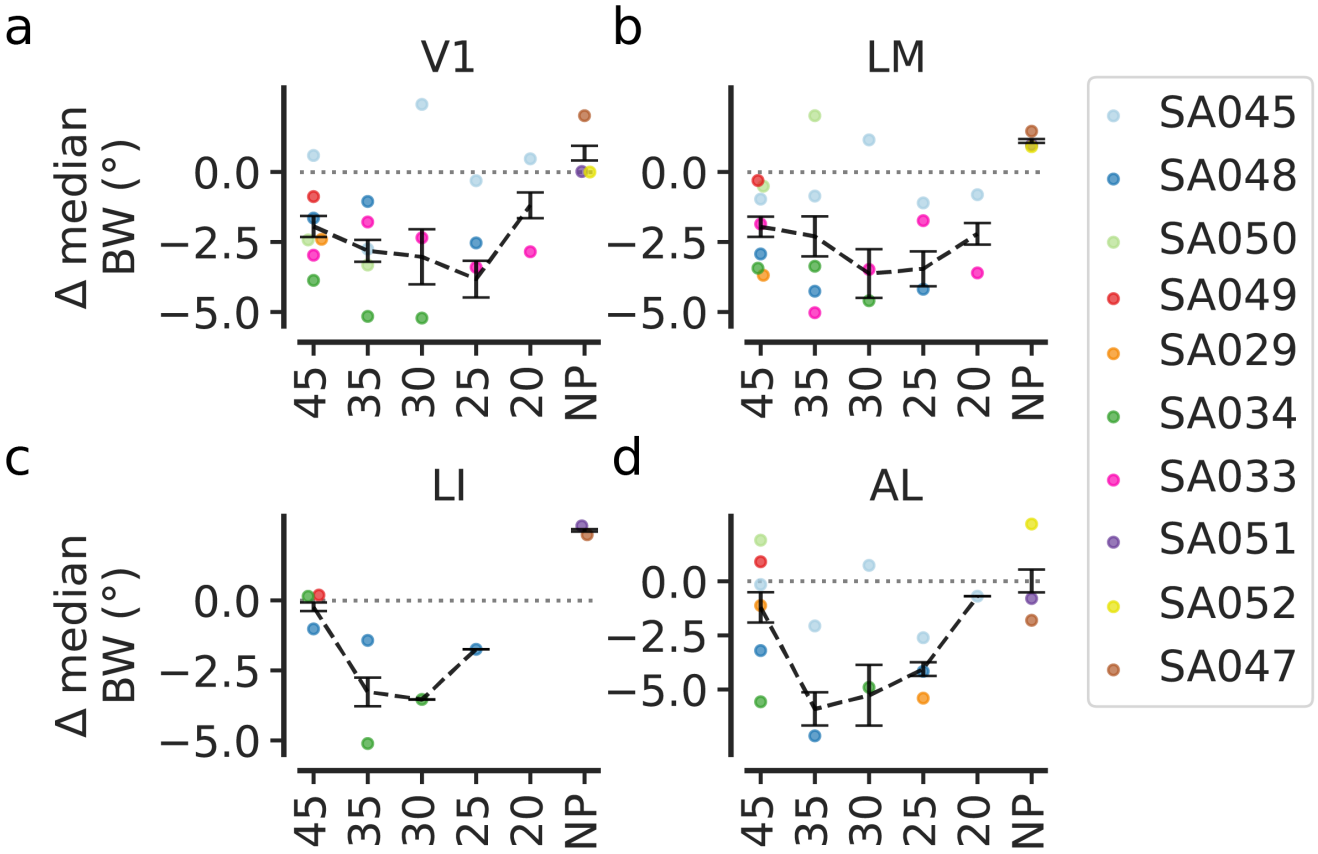



**Figure 3.21:** Change in median bandwidth (BW) split by visual area (subplots; V1, LM, LI, AL) and training stage (horizontal axis). Training stages 45 - 20 area for learning mice whilst NP = non-performing stage for non-performing mice. The black dashed line shows the mean change across mice. Error bars represent ± 1 SEM across mice.

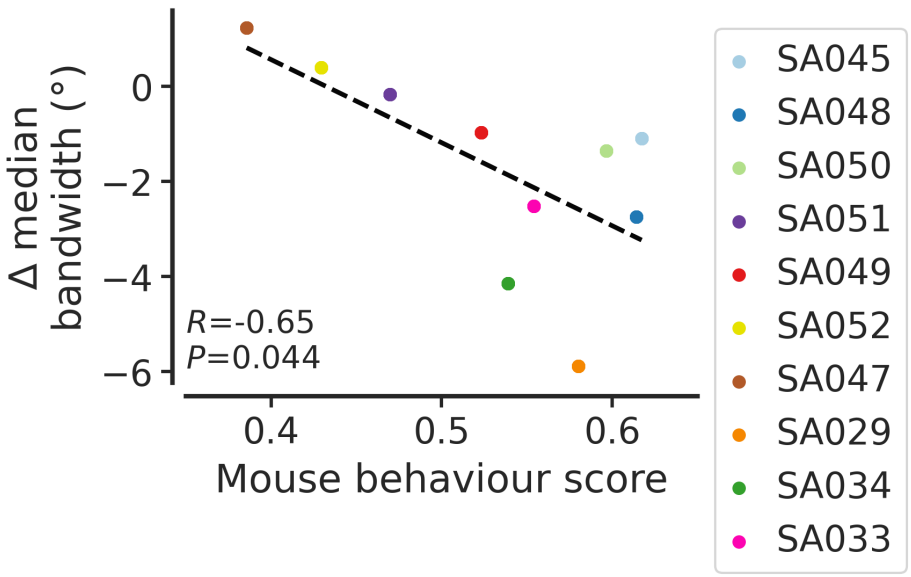



**Figure 3.22:** Mouse behaviour score (proportion correct trials across all of training) versus median bandwidth change for each mouse. R and P values shown are from Pearson correlation. Dashed line = linear regression line ($y = \text{slope} \cdot \text{behaviour score} + \text{intercept}$).

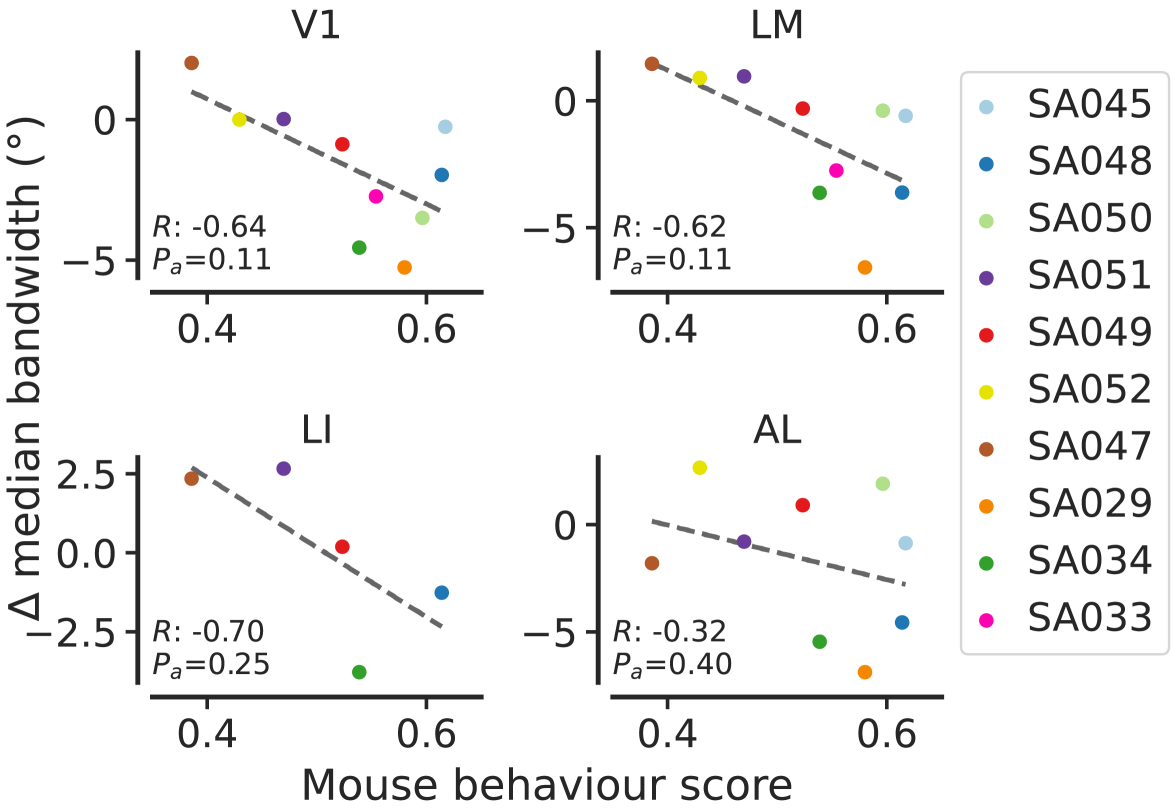



**Figure 3.23:** Mouse behaviour score (proportion correct trials across all of training) versus median bandwidth change for each mouse, split by visual area (subplots; V1, LM, LI and AL). R and P values shown are from Pearson correlation. $P_a$ denotes adjusted p-values following multiple-comparison correction. Dashed line = linear regression line ($y$ = slope · behaviour score + intercept).

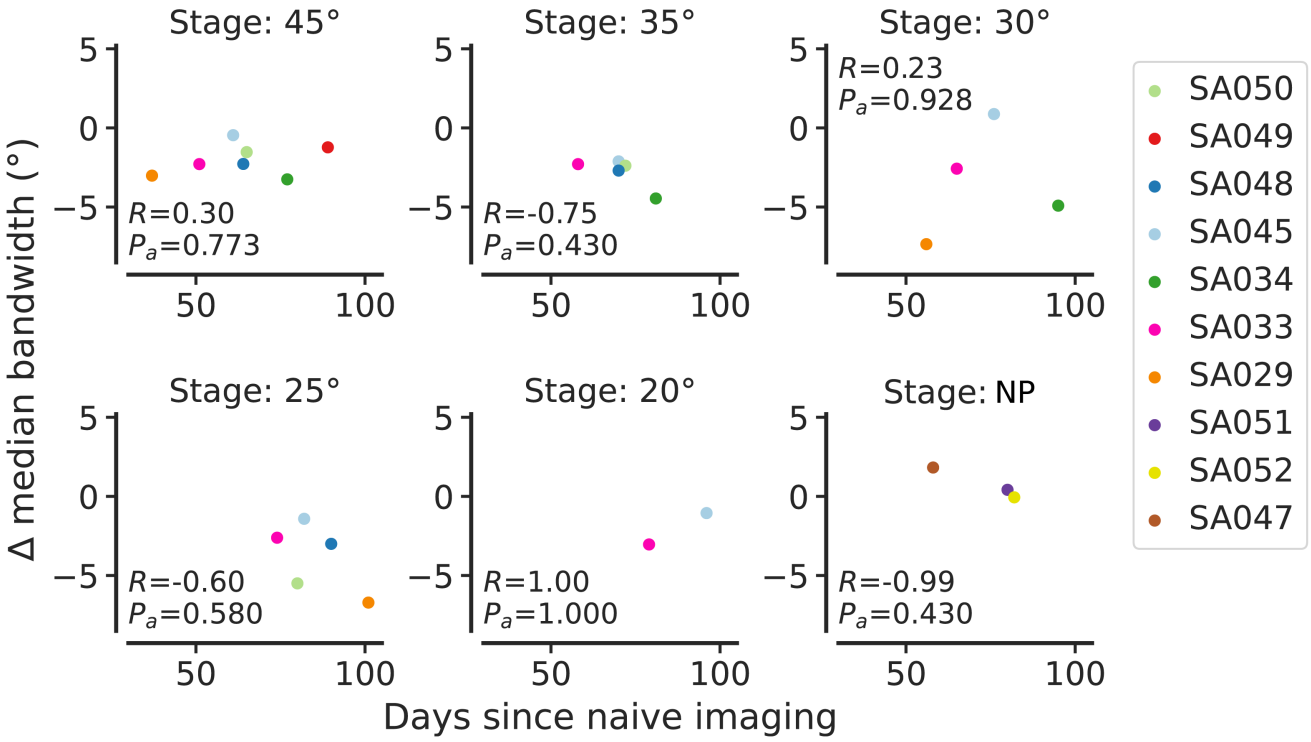



**Figure 3.24:** Days elapsed since naive imaging stage versus median bandwidth change measured at each training level (subplots). R and P values shown are from Pearson correlation. $P_a$ denotes adjusted p-values following multiple-comparison correction. NP = non-performing.

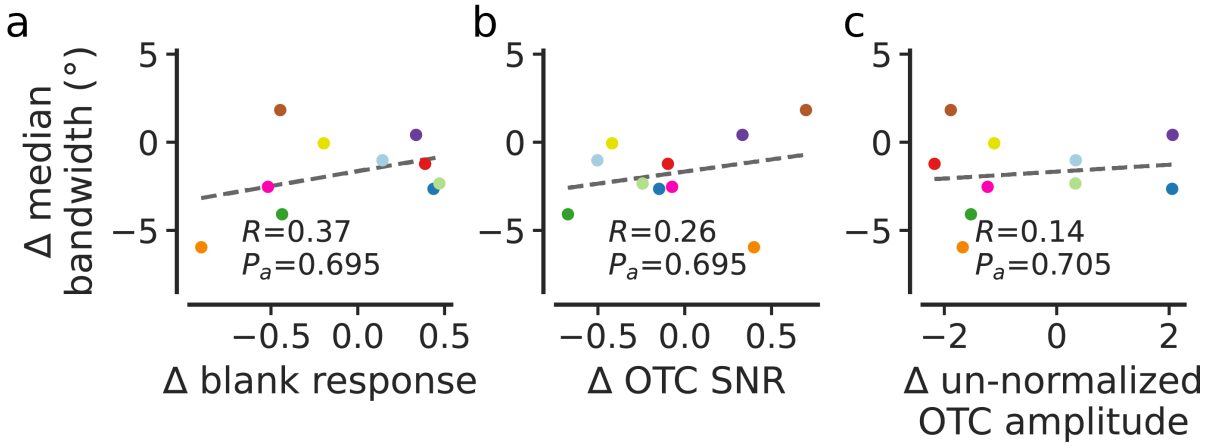



**Figure 3.25:** Changes in control variables versus median bandwidth change between naive and proficient (pooled) or non-performing stages. (**a**) Average change in neuronal response (deconvolved spikes) to the blank stimulus. (**b**) Change in orientation tuning curve signal to noise ratio (OTC SNR). (**c**) Change in average amplitude (maximum minus minimum) of OTCs before normalization. The black dashed line is from linear regression ($y$ = slope · behaviour score + intercept).

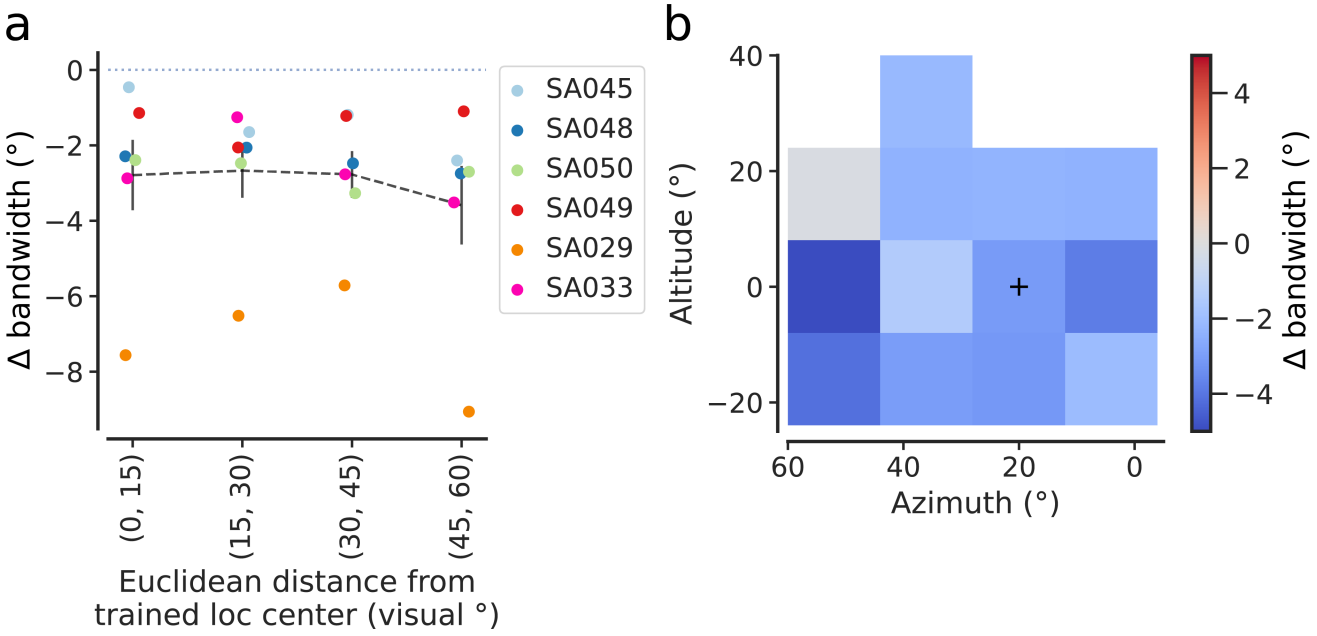



**Figure 3.26:** Median bandwidth change for learning mice, comparing naive and proficient (pooled) training stages, all visual areas combined. **a** Bandwidth change for neurons in intervals of Euclidean distance of their receptive field centre coordinate from the trained location centre. Bandwidth change within each distance interval was calculated relative to the naive level for each mouse in the same distance interval. Black dashed line shows the average change across mice. Error bars denote ± 1 SEM.**b** Bandwidth change for neurons within squares on a 2D grid of visual space. Each square represents the mean value across learning mice. For both subplots, a minimum of 10 cells at naive and proficient stages was required before including a data point for 1 mouse.

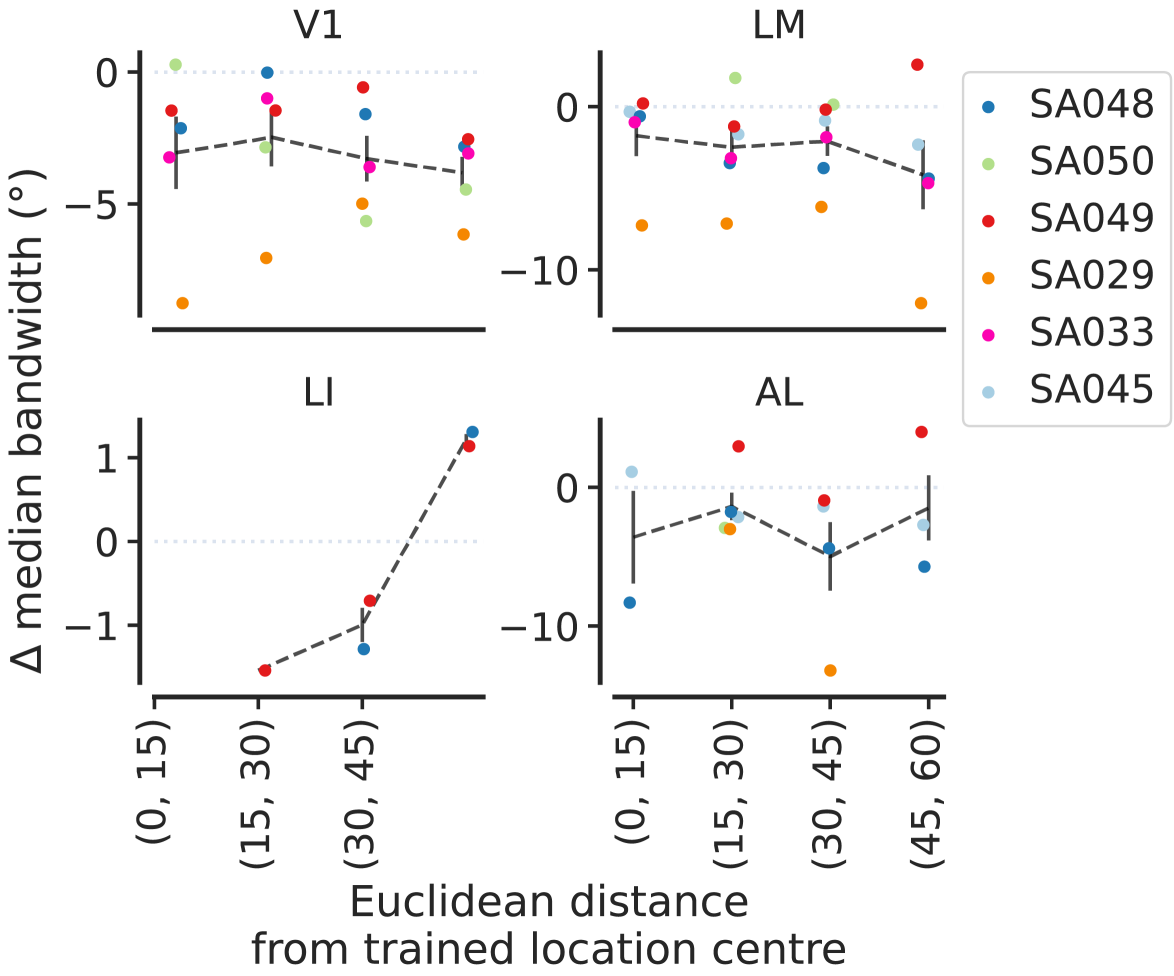



**Figure 3.27:** Median bandwidth change as function of Euclidean distance from trained location centre, split by visual area (subplots; V1, LM, LI and AL). Only learning mice were included here, comparing naive and proficient (pooled) training stages. Bandwidth change within each distance interval was calculated relative to the naive level for each mouse in the same distance interval. Black dashed line shows the average change across mice. Error bars denote ± 1 SEM. A minimum of 10 cells at naive and proficient stages was required before including a data point for 1 mouse.

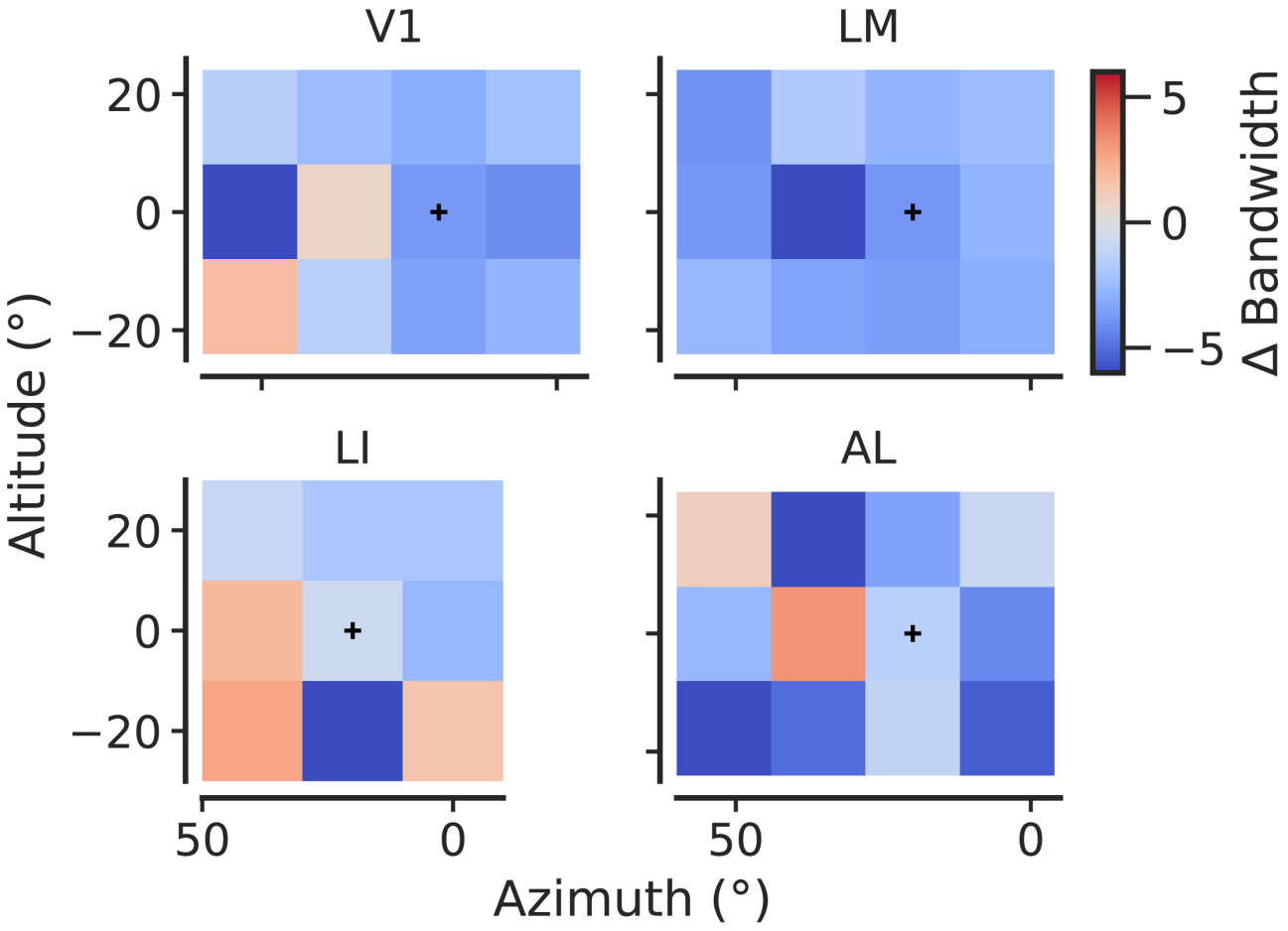



**Figure 3.28:** Change in median bandwidth for neurons within squares on a 2D grid of visual space, split by visual area. Each square represents the mean value across learning mice for neurons with receptive field centers inside the square. A minimum of 10 cells at both naive and proficient stages were required before including a data point for 1 mouse in each square.

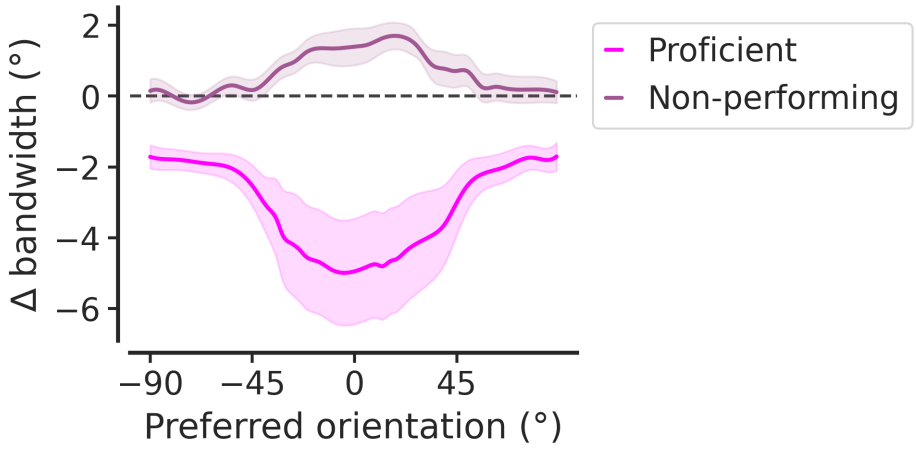



**Figure 3.29:** Change in median bandwidth as a function of preferred orientation, with all visual areas combined. For each preferred orientation, neurons with preferred orientation ± 16° were included in the calculation. Data for learning mice is pooled across proficient learning stages ([45,35,30,25,20], magenta). The mean change across mice is displayed, around which shading denotes ± 1 SEM.

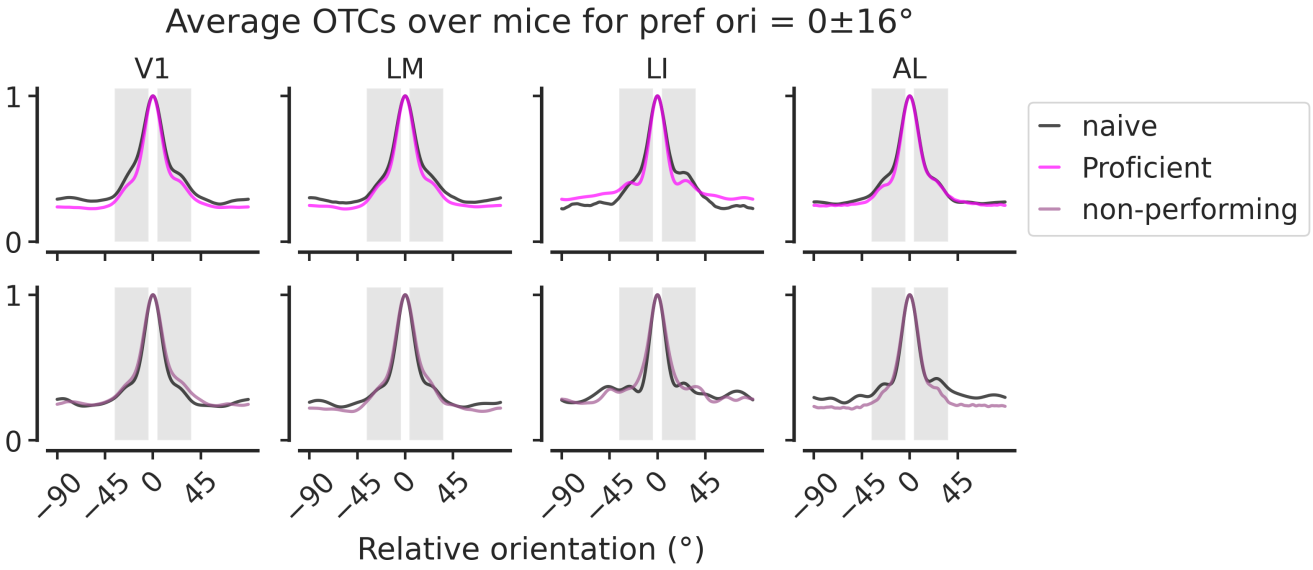



**Figure 3.30:** Average OTCs for learning and non-performing mice to illustrate bandwidth changes. Neurons with preferred orientation = 0±16° were centered then averaged together. Grey shading indicates the OTC region corresponding to ± 20°.

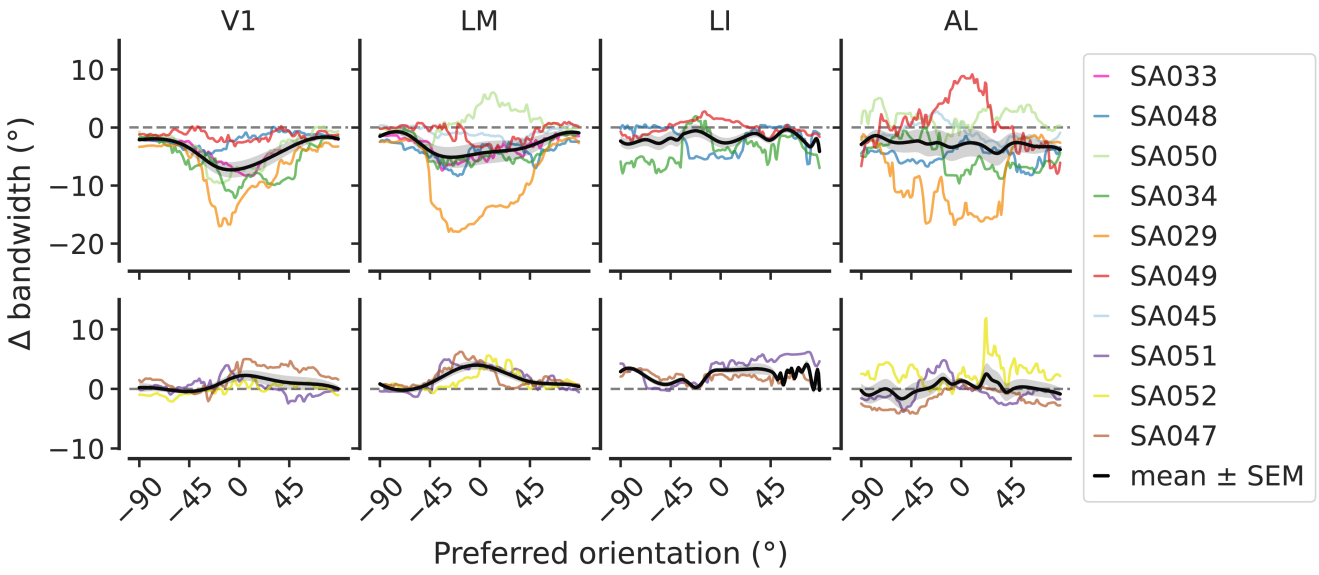



**Figure 3.31:** Change in median bandwidth as a function of preferred orientation, split across visual area (columns) and learning status (rows). For each preferred orientation, neurons with preferred orientation ± 16° were included in the calculation. Data shown compares the naive stage with all proficient learning stages for learning mice (top row of subplots) or naive to non-performing stage data from non-performing mice (bottom row of subplots). Results for individual mice are shown coloured by mouse ID, whilst the mean change across mice is displayed in black, around which shading denotes ± 1 SEM.

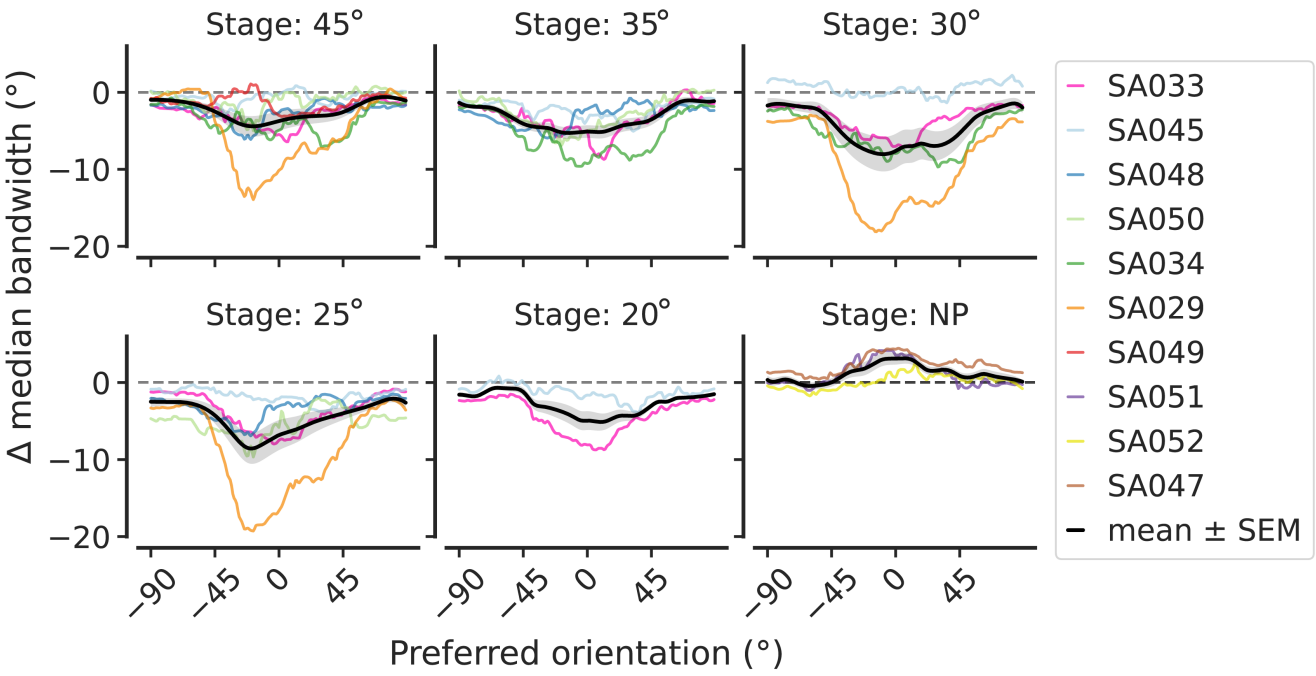



**Figure 3.32:** Change in median bandwidth as a function of preferred orientation, split across individual training stages (subplots). For each preferred orientation, neurons with preferred orientation ± 16° were included in the calculation. Data shown compares the naive stage each individual training stage. NP = non-performing. Results for individual mice are shown coloured by mouse ID, whilst the mean change across mice is displayed in black, around which shading denotes ± 1 SEM.

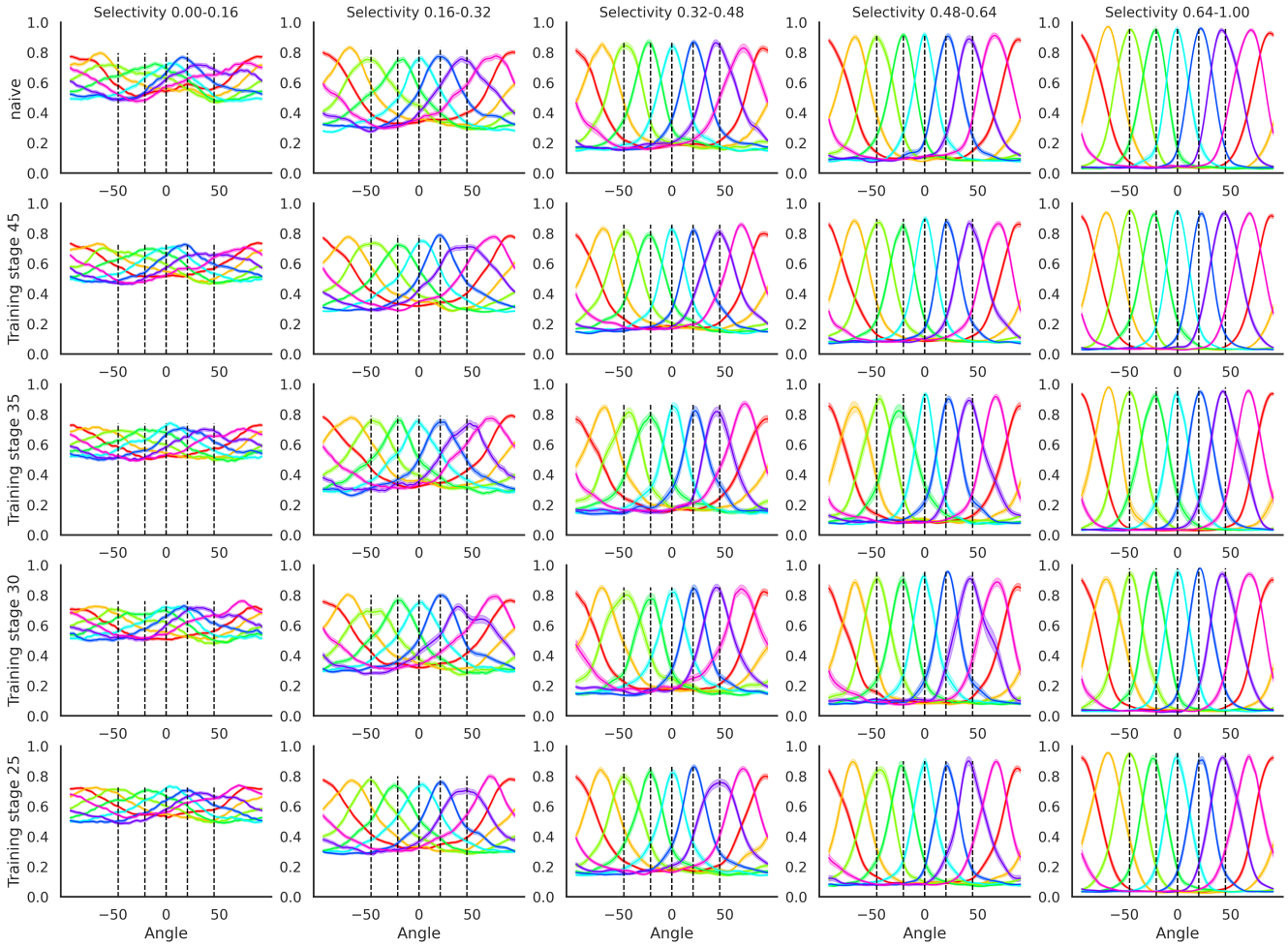



**Figure 3.33:** Area V1 average orientation tuning curves (OTCs) as a function of orientation selectivity (1-circular variance, columns) and training stage (rows). To allow averaging, neurons were placed in to 1 of 8 bins according to their mean preferred orientation and snapped to the nearest bin.

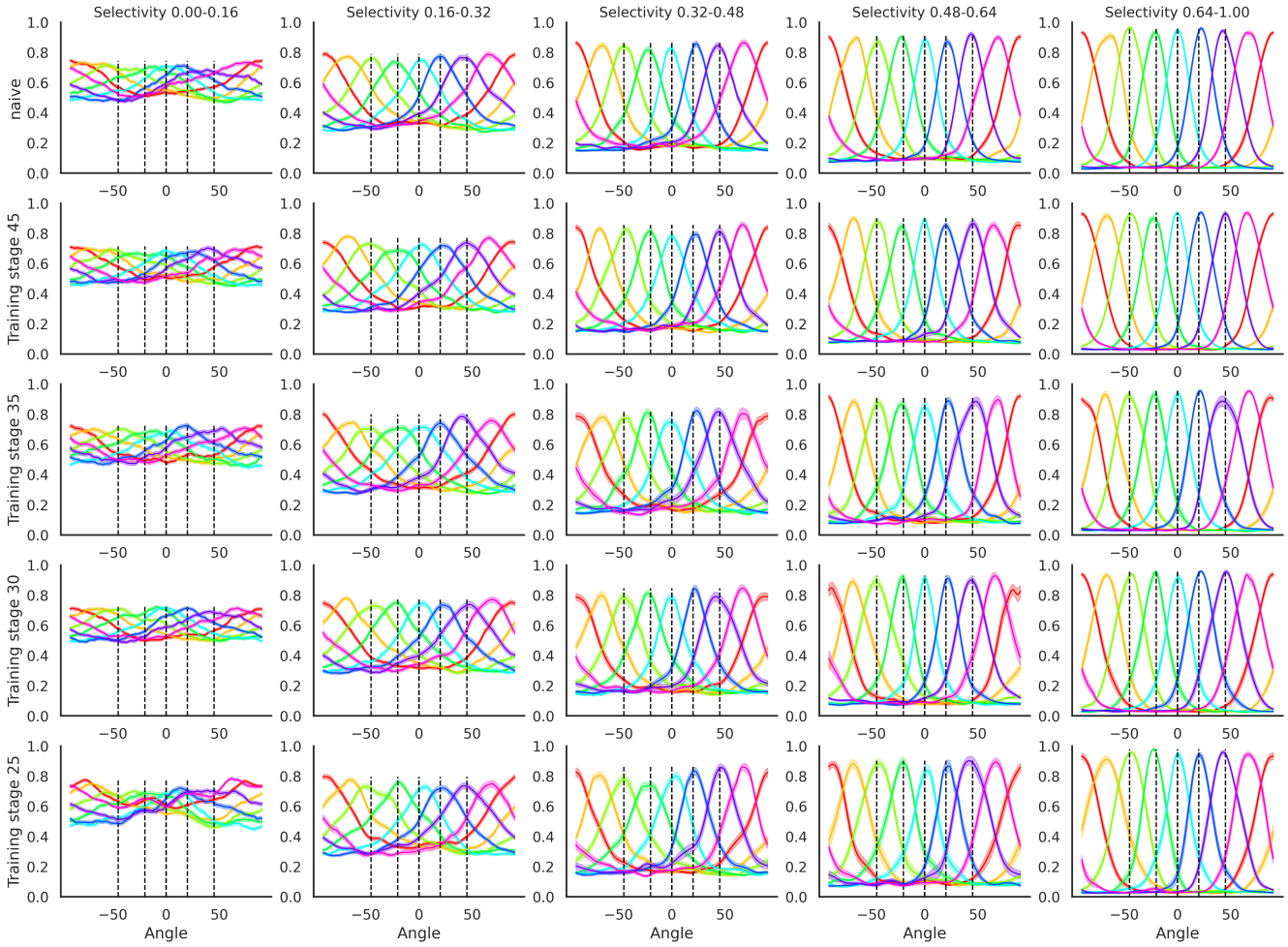



**Figure 3.34:** Area LM average orientation tuning curves (OTCs) as a function of orientation selectivity (1-circular variance, columns) and training stage (rows). To allow averaging, neurons were placed in to 1 of 8 bins according to their mean preferred orientation and snapped to the nearest bin.

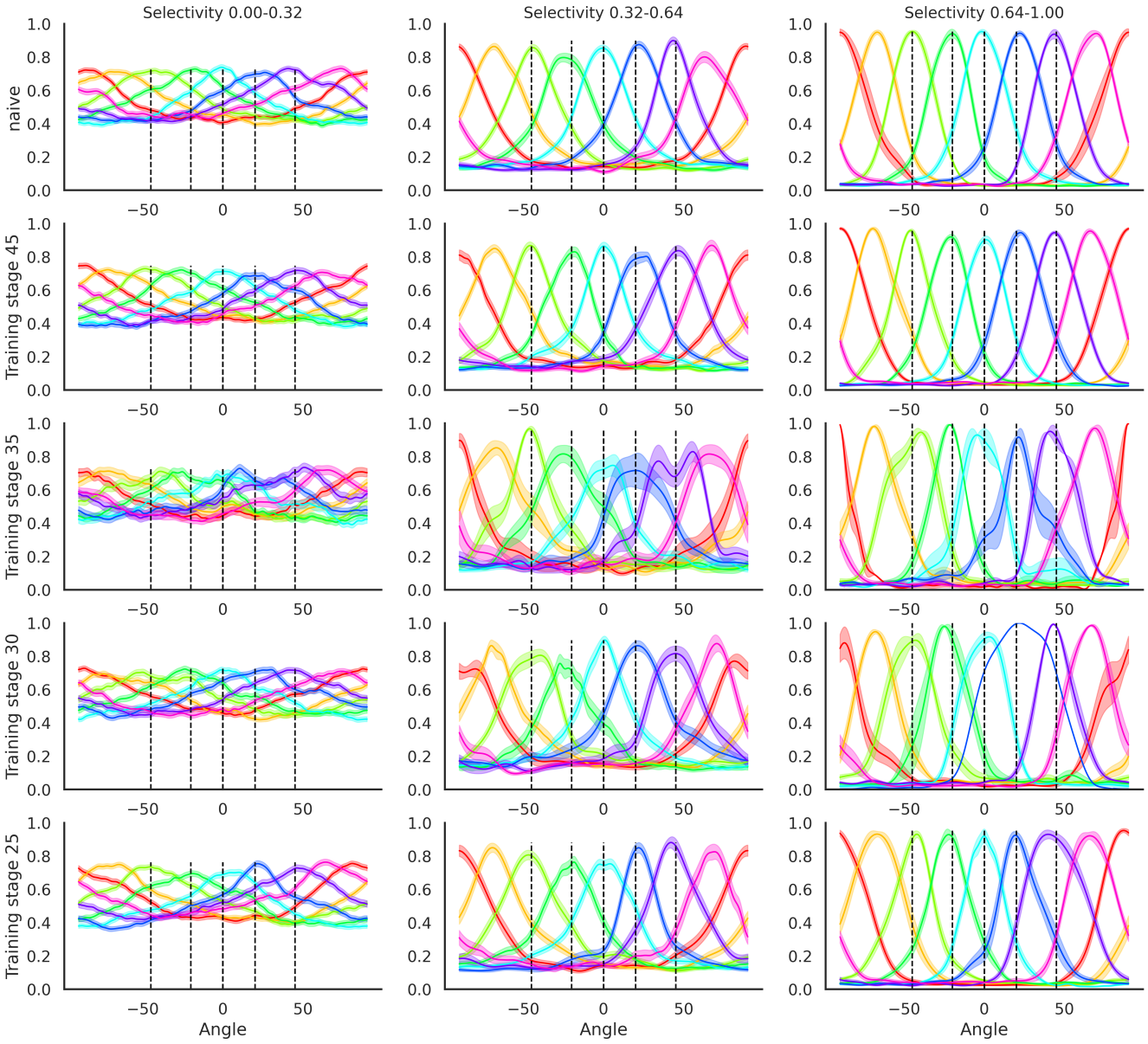



**Figure 3.35:** Area AL average orientation tuning curves (OTCs) as a function of orientation selectivity (1-circular variance, columns) and training stage (rows). To allow averaging, neurons were placed in to 1 of 8 bins according to their mean preferred orientation and snapped to the nearest bin.

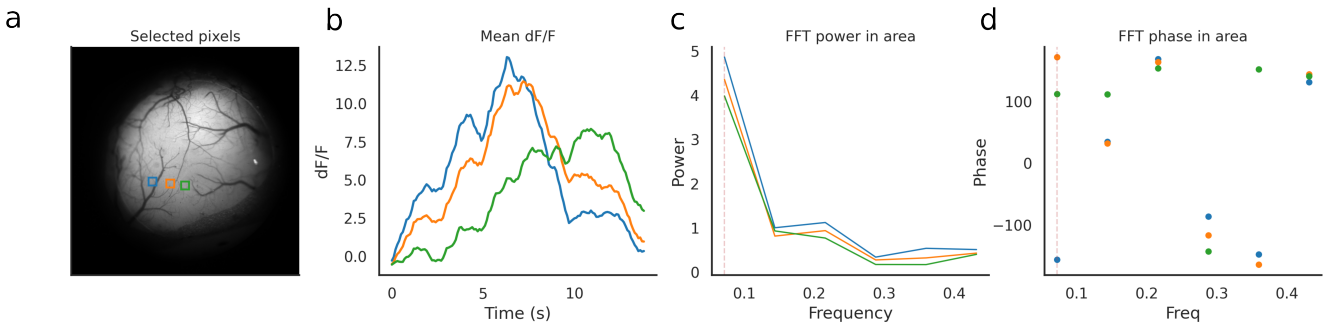



**Figure 3.36:** (**a**) an example image of the cranial window vasculature of mouse SA029 taken with a 4x objective before widefield retinotopic mapping. The pixels in the spatial areas indicated by the coloured boxes (inside area V1) were analysed in (b) (c) and (d). (**b**) the mean fluorescence signal across pixels in each box during one cycle of visual stimulation in the azimuth direction. (**c**) following a Fourier transform, this plot shows the power in each frequency, with the frequency of stimulation indicated by the dashed vertical line. The power is the highest at this frequency. (**d**) shows the phase of each frequency, which demonstrates that for the frequency of stimulation (dashed vertical line), the phase is ordered in space, reflecting the stimulus moving in the azimuth direction from lateral V1 left (blue box) to right (green box).

# 4

# Discussion

## 4.1 Single neuron activity perturbations

We developed a way to design and interpret experiments involving single-neuron activity perturbations, based on a circulant assumption that allows us to infer recurrent tuning curves from measuring the effect of a perturbation on a neural population.

In order to be able to make useful inferences from perturbations, we began with a simplifying assumption about the tuning of sensory neurons, namely that weight matrices are circulant. We then demonstrate analytically that there exists a space of different network architectures, employing different connectivity motifs, that produce identical firing rate responses to any sensory stimulus. Next we ask whether single neuron perturbations would distinguish these networks. We derive 'influence functions' that describe how the network response to a single-neuron perturbation depends on tuning coefficients in each network, and find that perturbation responses differ amongst networks that respond identically to a sensory stimulus, reflecting the underlying recurrent connectivity motifs, which we show can be reconstructed by reversing the influence function.

We next demonstrated applications of these general results, showing how both normative and mechanistic principles could be tested using perturbations. We described a test of efficient coding principles and extended this to the nonlinear case of sparse efficient coding with threshold-linear units, finding that the temporal dynamics of the response become crucial. We also predict the effect of perturbing 'error' and 'state' neurons in a predictive coding network, demonstrating how these cell types must be dynamically coupled to optimize the input prediction objective.

The circulant assumption is a strong assumption, yet there are places where it seems to be consistent with neural data and circulant tuning has been used in previous theory work. Circulant tuning is standard in models of visual orientation tuned neurons (Ben-Yishai *et al.* 1995; Rubin *et al.* 2015). Measurements of orientation tuning in visual cortical neuron populations seem to be approximately consistent with this assumption (Rossi *et al.* 2020), and neuronal stimulus tuning is often analysed as an average curve centered around the preferred orientation which implicitly assumes repeated tuning. Slightly more general, Toeplitz matrix tuning is commonly used to model the tuning of visual neurons to space for example, e.g. the spatial model stabilized supralinear network (SSN) has Gaussian Toeplitz tuning (Rubin *et al.* 2015). It has been shown that eigenvalues and other properties of Toeplitz matrices and circulant matrices can be asymptotically equivalent, in the limit of large N (Gray 2006; Zhu & Wakin 2017). This says that if networks are sufficiently large, our analysis in circulant networks may approximate some features of the more general case of Toeplitz networks, where tuning functions are still repeated but do not 'wrap around the edges'. We do not analyse this case here, but it could offer a way to extend this framework.

We only considered symmetric circulant matrices here, corresponding to even tuning curves with real frequency coefficients. Asymmetric weight matrices give rise to more complex dynamics due to complex valued frequency coefficients which could cause oscillatory behaviour in the dynamics. We have not considered these dynamics in this work, but it could be extended to this case if relevant.

Circulant matrices are also normal, however in a 2-population network such as an inhibition stabilized network with all circulant weights, the overall dynamics matrix can still be non-

normal meaning that the system eigenvectors are non-orthogonal. This means that the transient part of the neuronal response can contain more complex dynamics like balanced amplification (Hennequin *et al.* 2012). At steady-state, the responses would still depend on weight matrix frequency coefficients in the linear case. Transient neuronal responses following stimulus onset are features of both prediction error encoding neurons and neurons in networks with non-normal dynamics, yet arise from different mechanisms in both models. This points to a limitation of our analysis.

Furthermore, must acknowledge that inferences about network connectivity are sensitive to assumptions on network weights, architectures and nonlinearities. For example we showed that in recurrent networks, perturbations are specifically diagnostic of recurrent connectivity. However, when other motifs such as feedback are added, responses depend on a mixture of these components.

We considered the case of single neuron excitatory activity perturbations, whereas patterned multi-neuron perturbations are possible experimentally. To approach this, one idea is to use optimal control theory and optimization to find optimal perturbation patterns to achieve certain objectives, such as to evoke the maximum norm response. Another objective could also be to maximise the difference in responses between two competing network models (similar to the approach of Golan *et al.* 2020). In nonlinear networks or with non-convex problems, simulation could be used to find these solutions.

In conclusion, our results make a case for the usefulness of perturbations guided by careful theoretical considerations, whilst acknowledging how the inferences made from results can change depending assumptions on network architecture and nonlinearities. We hope our results contribute to illuminating both the promise and limitations of activity perturbations in discriminating theories of cortical sensory processing.

## 4.2 Evaluation of predictions for the deep learning theory of perceptual learning

### 4.2.1 Orientation tuning curve changes were not specific to the trained stimulus orientation

The experimental predictions we set out, provoked by the theoretical work of Saxe (2015), predicted that OTC slope at 0° for neurons would increase the most in mice that learnt the task. Our findings did not support this. Statistically, within-mouse slope change at 0° compared between naive and proficient training stages did not differ from zero. Within-mouse slope changes comparing 0° and 90° did not differ either. Finally, slope change at 0° did not differ across learning status (learning vs. non-performing). This goes against the theoretical prediction that changes would target the most informative neurons, which were those with the steepest slopes at 0°.

Instead, we found that slopes increased for intermediate orientations that mice were exposed to on rewarded (correct) trials, i.e. orientations inside ±45 degrees. Slopes most prominently changed close to ±20 degrees for learning mice, by 0.006-0.009 units (response/degree) for CCW and CW sides respectively. The average slope change here was conversely -0.003 for non-performing mice. The difference at ±20° for learning mice was statistically significant from two perspectives. Firstly, within-subject slope change for learning mice, but not non-performing mice, was significantly different from zero. Secondly, the average change compared across learning status was statistically significant. As a reminder, slopes at ±20° were measured from neurons with modal POs between 2° to 38° and -38° to -2°. This corresponds to the neurons with the steepest slopes at 20° or -20°. The magnitude of slope increase within learning mice was similar to that found by Failor et al. in mouse area V1 after proficiency in a related orientation task (Failor *et al.* 2025).

The median OTC bandwidth significantly differed by learning status following training for neurons across all preferred orientations combined. Specifically, the median bandwidth changed by -2.67 vs +0.22 degrees on average for learning versus non-performing mice. A leftward shift in the main body of the OTC bandwidth distribution occurred specifically in all 4 visual areas for

learning mice, and interestingly, appeared to shift subtly rightwards for non-performing mice in V1, LM and LI only. These changes did not arise from changes in the frequency of extreme values, because they persisted after filtering out the tails of the distribution.

Furthermore, learning mouse bandwidth change was an inverse bell curve function centered around neurons with POs at 0°. Bandwidth reductions focused around 0° PO may explain the greater increases in OTC slope observed at orientations around, but less so at, the trained orientation. Bandwidth changes were therefore focused to neurons active during rewarded trials. For neurons with modal POs at 0±16°, the change in median bandwidth averaged across learning mice was nearly -5°. Across learning status, the difference at this point was -8.8°. Both the within-subject change and cross-learning status comparisons were statistically significant. In addition, bandwidth was still significantly lower at POs near 90° for learning mice, albeit with a much smaller magnitude. Overall, there was a broad sharpening of OTCs across the measured neural populations after task learning. More specifically, OTC bandwidth decreased as a function of preferred orientation distance from the rewarded angle in learning mice only. In the study in monkey V4 by Yang and Maunsell (2004) but not the study of V1 in the same laboratory by Ghose (2002), OTC bandwidth also changed as a function of distance of preferred orientation in trained but not untrained neurons, although the shape of the function was not as clear as in our results. Another possible framing of the observed changes is that the bandwidth difference was a function of temporal difference between time of exposure to a given angle and delivery of the water reward. One possibility is that OTC sharpening occurred due to differing attention, engagement or arousal levels between performing and non-performing mice, but this does not explain the dependence on preferred orientation.

Can the results of Failor and colleagues (2025) be reconciled with our own? The cited study found that the OTC slopes at the two cue orientations increased selectively in mouse V1. The task used by Failor was static, meaning stimulus orientation was not coupled to the angle of the wheel used to respond. Because of this, the most active neurons on a correct trial were likely those with PO close to the cue orientation. Slopes at the two cue orientations in this study changed the most for neurons within ±23 PO of each cue orientation. Similarly to our

results, they also found that slopes measured at the angle that lay between the two cue angles (68°) did not change. This angle was seen but likely spatially less attended to on correct trials. Taken together with our findings, including that slopes/bandwidths did not increase/decrease for non-performing mice, this could speculatively point towards changes that target neurons depending on neural activity at the time of or just preceding reward.

### 4.2.2 Orientation tuning curve changes were diffuse in visual space

The next specific prediction of the theory was that changes would be specifically targeted to neurons whose spatial receptive field was inside the stimulus location. This prediction arises in a fully connected deep network where units are spatially tuned, because feedback weights for the error signal are spatially specific. Previous data from primate V1 indeed showed that slope changes following orientation discrimination training were specific to trained but not 'naive' locations (A. Schoups *et al.* 2001).

Here, neither the observed slope nor bandwidth changes differed meaningfully or significantly when neurons classed as spatially trained vs. spatially naive were compared. This simple comparison essentially split the entire dataset into two, meaning it was powered by tens of thousands of neurons overall. Further analysis showed that Euclidean distance of the RF centre coordinate from the trained location centre did not relate to the magnitude of slope or bandwidth changes.

A smooth retinotopic gradient can be measured at the macro-scale (100s of microns) in visual cortical areas (Zhuang *et al.* 2017), but the individual RFs of neurons are actually locally diverse or heterogeneous, as seen here (figure 3.9g) and reported widely (Bonin *et al.* 2011). One possibility is that the mapping of a neuron to the macroscopic retinotopic gradient matters most in how it is targeted for modulation or learning. Our spatial analysis was limited to neurons with spatial RFs preferring the first 50 degrees of the left visual field in azimuth. Many 'untrained' neurons according to single neuron RF could have be spatially right in the stimulus location with respect to the macroscopic gradient. In future studies, using a larger stimulus screen

and a cranial window accessing medial V1 would allow more distal spatial locations in V1 to be sampled. Sampling from both cerebral hemispheres would be desirable, although not possible with the cranial window approach taken here.

### 4.2.3 No clear evidence of reverse hierarchy dynamics

Early visual areas V1 and LM changes were very similar in nature and in time. We did not observe that area LM changed before area V1, and the initial change across those areas did not differ in magnitude.

We did not detect changes of a greater magnitude or earlier in higher order areas LI or AL in comparison. This goes against the theoretical prediction that higher order areas (higher layers) would change the most, and change first.

Although V1 and LM did not seem to differ, sampling was poorer for the two higher-order areas. This was for technical reasons, discussed previously, as well as there being fewer orientation responsive detected overall, which has been reported previously (Marshel, Garrett, *et al.* 2011). This meant that our filtering criteria affected higher order areas more strongly. This points to a limitation of using orientation tuning as a read-out of learning across visual areas: higher order areas encode more complex stimulus features and display more complex feature tuning (Riesenhuber & Poggio 2002) although this is not well-understood in the mouse brain (Glickfeld & Olsen 2017).

Sampling was also poorer for later training stages in learning mice (4 or 2 mice per stage) because of the low-throughput nature of the behaviour task. We therefore did not have the statistical power to discern temporal learning dynamics of changes across areas.

### 4.2.4 Further predictions

One prediction was that early learning would spatially transfer more than late learning. Ultimately, our behavioural paradigm was not suitable for answering this question. One spatial

transfer session the day after imaging at each stage acquisition was included to probe spatial transfer. Here strictly one trial per novel location for CW and CCW orientations was sampled to avoid training mice on novel locations. Mice were clearly perturbed by changes in stimulus azimuth and altitude. This was evidenced by a large reduction in performance and increase in bias for trials at the trained location, depite having just displayed high proficiency in those trials. This was the case even though trials at the trained location were most frequent and spatial probe trials were relatively sparse. Mice may have experienced confusion or changed strategy following the change in task details, struggling to generalize task rules to the novel setting. It could also be because transfer trials were not rewarded as mice had learned to expect. We decided that spatial transfer sessions, because there was only a handful of sessions per mouse, did not provide a reliable measure of perceptual performance. This prediction therefore remains untested.

### 4.2.5 Further evaluation of the experimental paradigm

The behaviour analysis alone revealed that learning of the task and expression of perceptual abilities was complex and multifaceted for mice, unlike in the neural network model that was explicitly instructed to optimise CW vs CCW classification performance from the outset. Indeed, although mice in this study were ultimately able to report discriminations as low as 15°, this is substantially higher than for primates, who can develop discrimination thresholds under 1 degree (A. Schoups *et al.* 2001). This difference is stark given that orientation tuning bandwidths in V1 do not differ much between mice and primates (Van Hooser 2007; Niell & Stryker 2008). In addition, orientations can be decoded from 2-photon recordings of mouse V1 with extremely high precision (Stringer *et al.* 2021). It is not known how the presence of this stimulus information maps on to true behavioural abilities, as limitations in inter-cortical communication may limit read-out precision. Taken together, although the mice in this study expressed lower perceptual thresholds than in prior literature, possibly due to the staircase curriculum. However, it remains possible that the behavioural paradigms employed do not permit mice to express their true latent knowledge, causing threshold estimates to be inflated (Kuchibhotla *et al.* 2019).

### 4.2.6 Conclusion

To summarise, we made specific a-priori predictions for an implementation of the deep learning theory of perceptual learning and designed a behavioural and imaging paradigm to test them in the mouse visual cortex. Predictions were specific to one flavour of supervised gradient descent learning with a classification objective in a specific, layered serial network architecture. Overall, none of the predictions were shown to be directly supported by the neural dataset we collected. A caveat is that sampling of neurons was reduced in both higher order visual areas and in late training stages, meaning we did not have sufficient power to confidently falsify all predictions. Despite some limitations in the dataset, the overall pattern was that the simple theory we tested did not account for patterns in the neural results that were observed. This leads us to question both the model assumptions and the learning theory itself, opening the door to a whole range of data-driven modelling that will ultimately lead to more predictions, then more experiments.

Comparison of datasets from carefully controlled tasks that differ in specific details could be exploited to better understand learning in the brain. Under different theories, comparing neural changes under different task settings could distinguish different learning processes. This also points to the need for increased coordination of experiments in neuroscience research to make progress in this area. If behavioural paradigms differ in too many details, constraining theory with different sets of results becomes challenging. This is made even more challenging when combined with the a lack of standardisation in analysis methods for neural data. For the present study, we carefully considered tens to hundreds of design choices from stimulation paradigms, sampling settings, behavioural task, data pre-processing, analysis, visualisation and statistical approaches. The only elements that were near-standardised were production of widefield retinotopic maps, and 2-photon data pre-processing, both owing to open source tools shared by the academic community (Zhuang *et al.* 2017; Pachitariu, Stringer, Dipoppa, *et al.* 2017). More mature fields like astrophysics serve as inspiration to increase multi-centre collaborations, and to work towards standardisation of datasets to permit proper comparison. Understanding learning principles in brains within and across species remains one of the most challenging problems in science.

# Appendices

# .1 Behavioural analysis figures for all individual mice

Behaviour summary plots for each individual mouse are plotted here. Learning status is indicated in each caption.

The following figure legend is common to every figure in this section: (**a**) Proportion correct non-repeat trials over the course of training. (**b**) bias (probability of choosing CW at 0° minus 0.5) over the course of training (**c**) cumulative correct trials (or reward amount) for individual sessions (lines). Dark blue to light green line colour signifies increasing session number. The red line is unity. (**d**) psychometric curves for 5 groups of sessions (quintiled) whilst mouse was naive. Dark blue to light green indicates earlier to later sessions up to 45° acquisition. (**e**) psychometric curves for further training stages, after the introduction of sparse probe trials at non-modal cue stimulus angles to improve curve estimation. (**f**) perceptual orientation threshold determined from interpolation of psychometric curve at 70% correct value, computed separately for left and right sides of the curve.(**g**) psychometric curve slopes, computed separately for left and right sides of the curve. (**h**) average entropy of the stimulus orientation (stim pos entropy), or wheel position (wheel pos entropy) over sessions. (**i**) average reaction time (RT) across sessions. RT is the time between stimulus appearance (start of closed loop) and choice being made (stimulus orientation reaching 0° or $-\theta_{trial}$° degrees)).

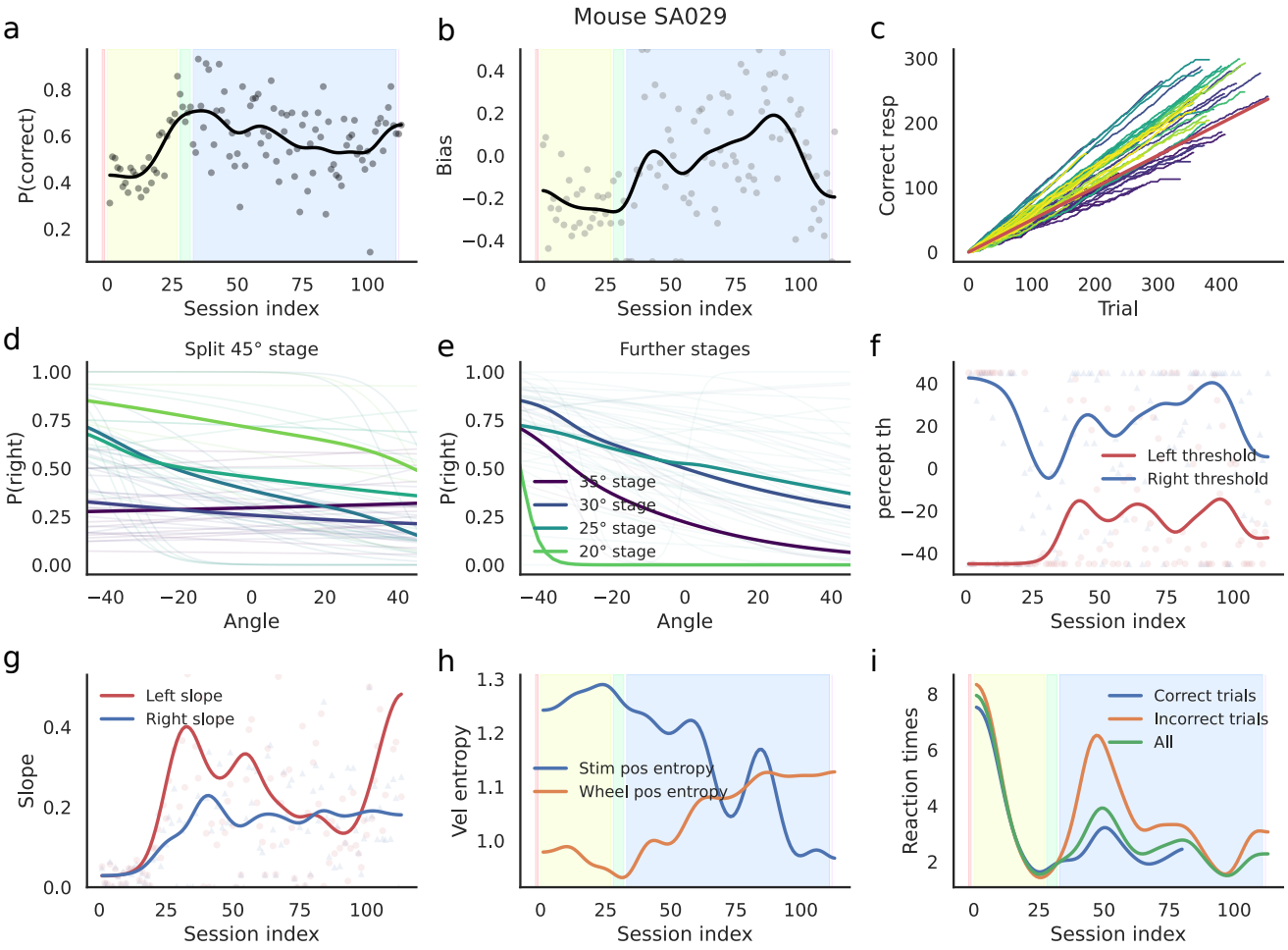



**Figure 1:** Mouse SA029 (learning)

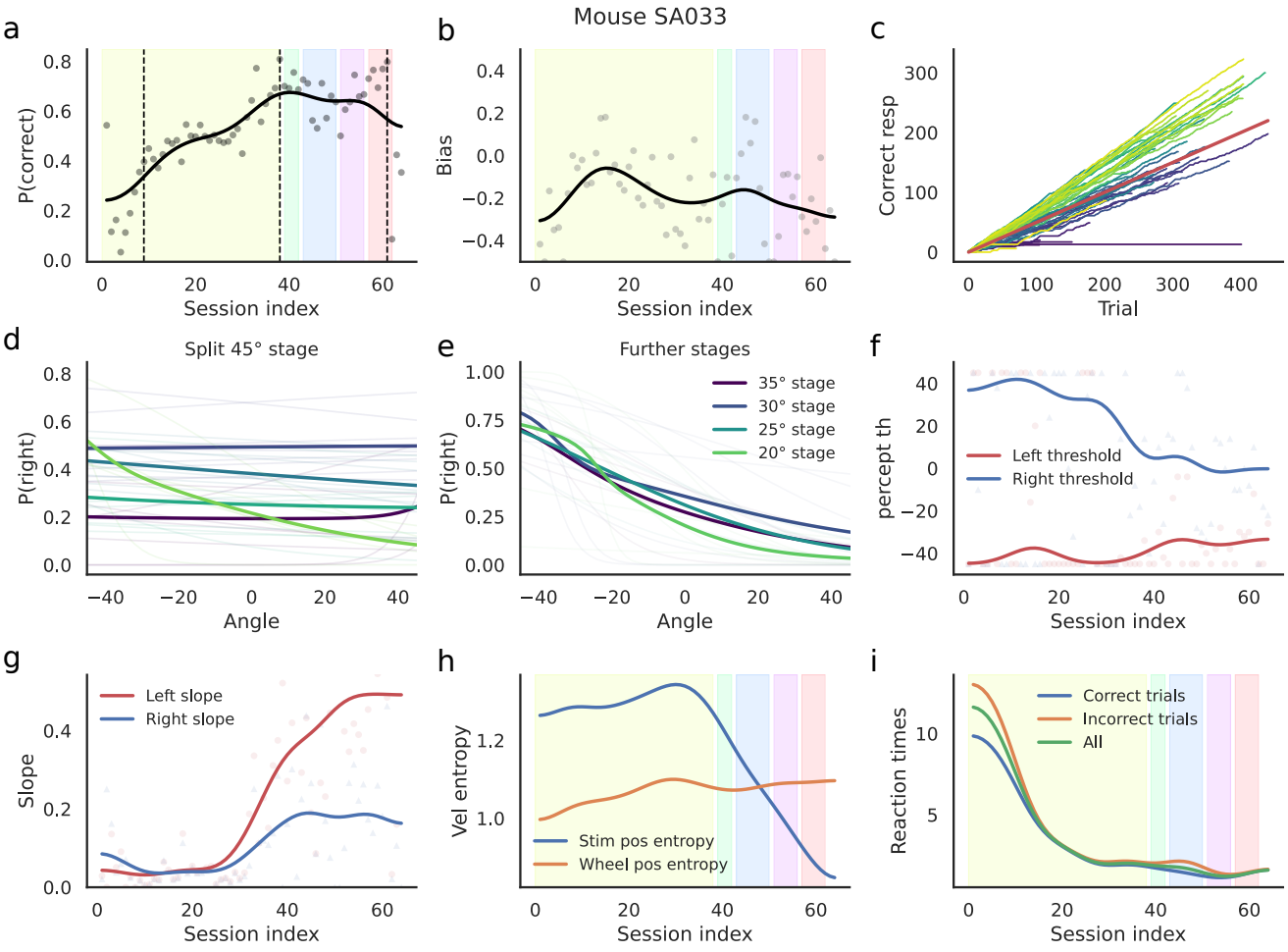



**Figure 2:** Mouse SA033 (learning)

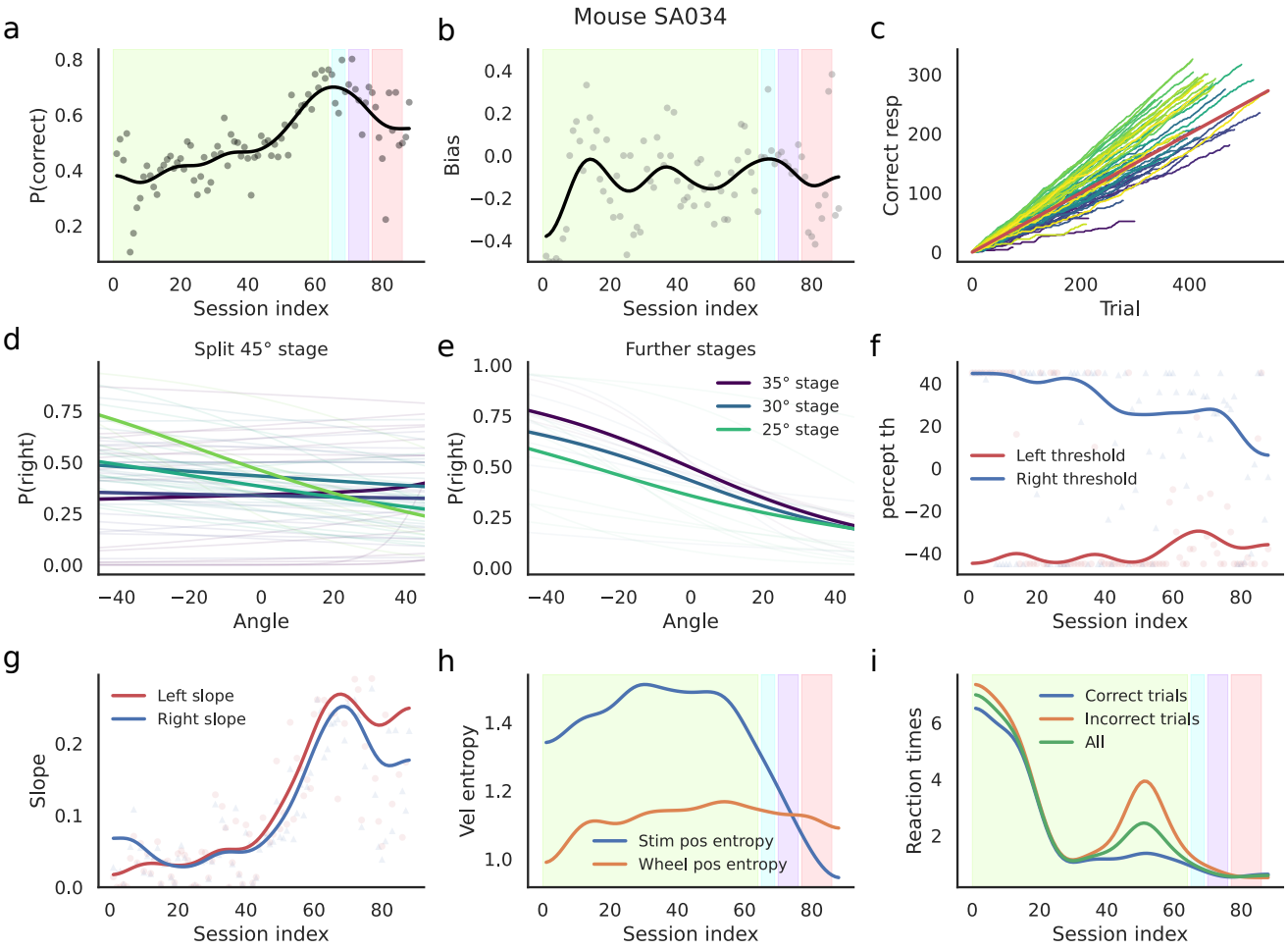



**Figure 3:** Mouse SA034 (learning)

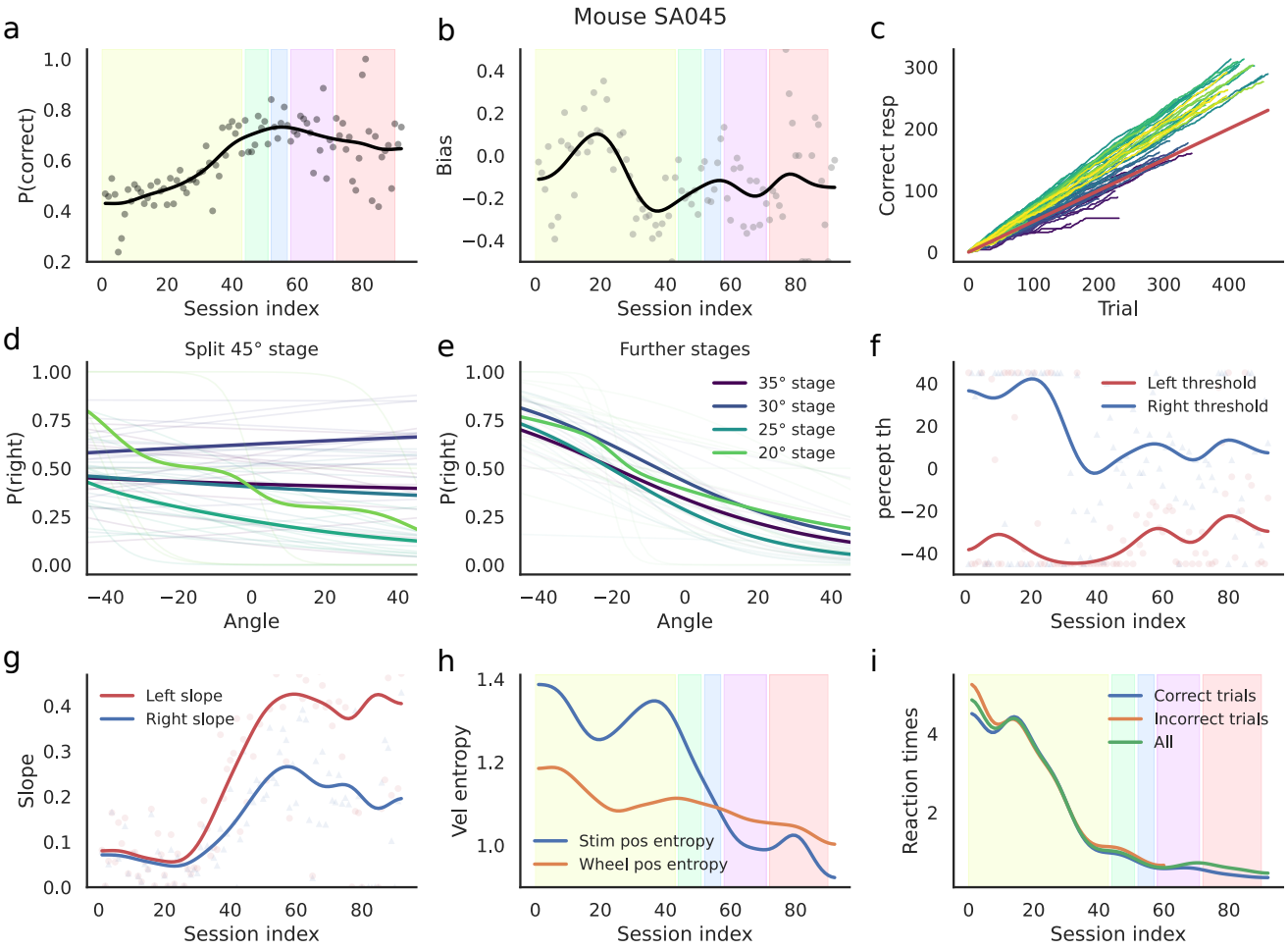



**Figure 4:** Mouse SA045 (learning)

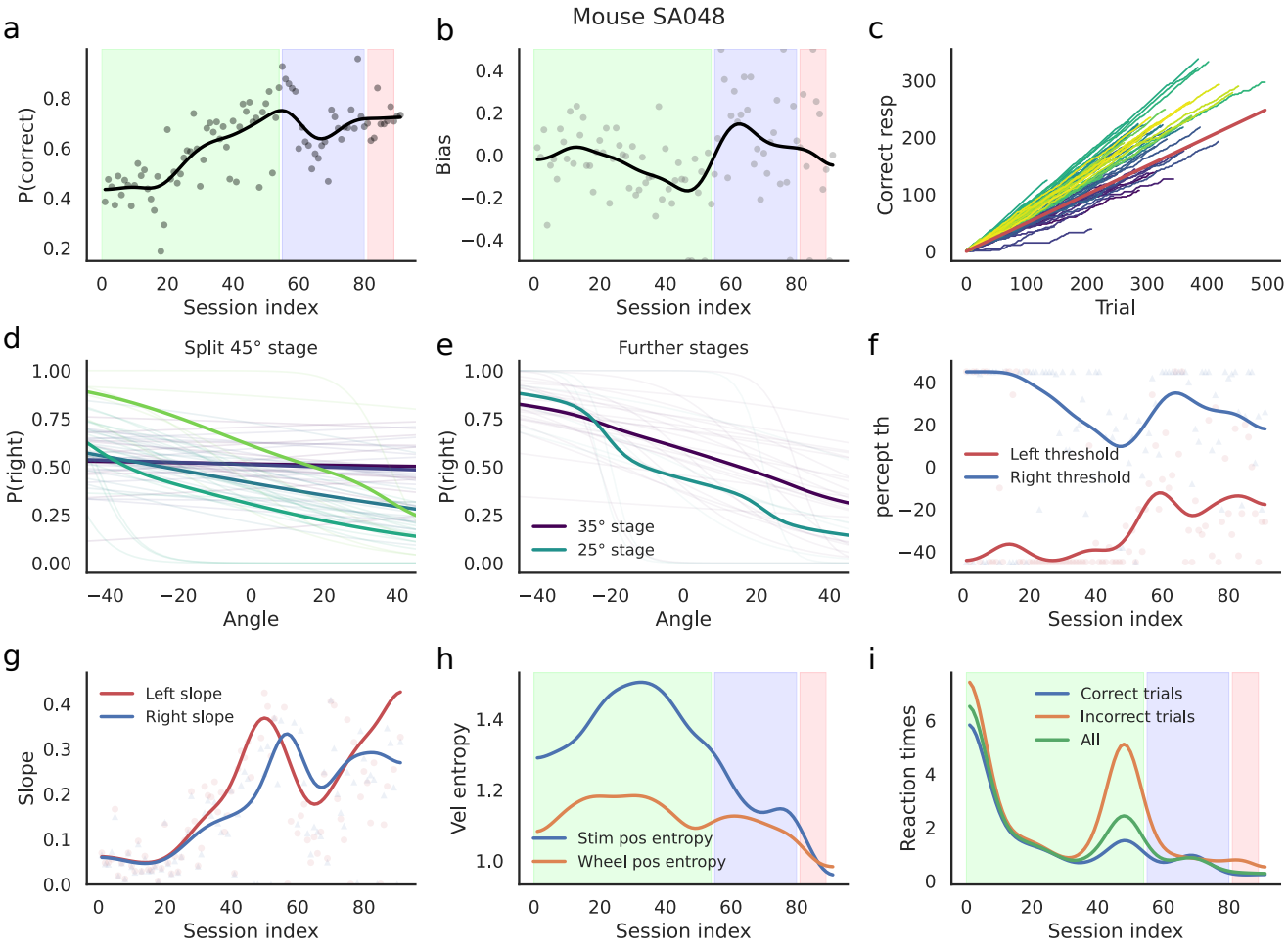



**Figure 5:** Mouse SA048 (learning)

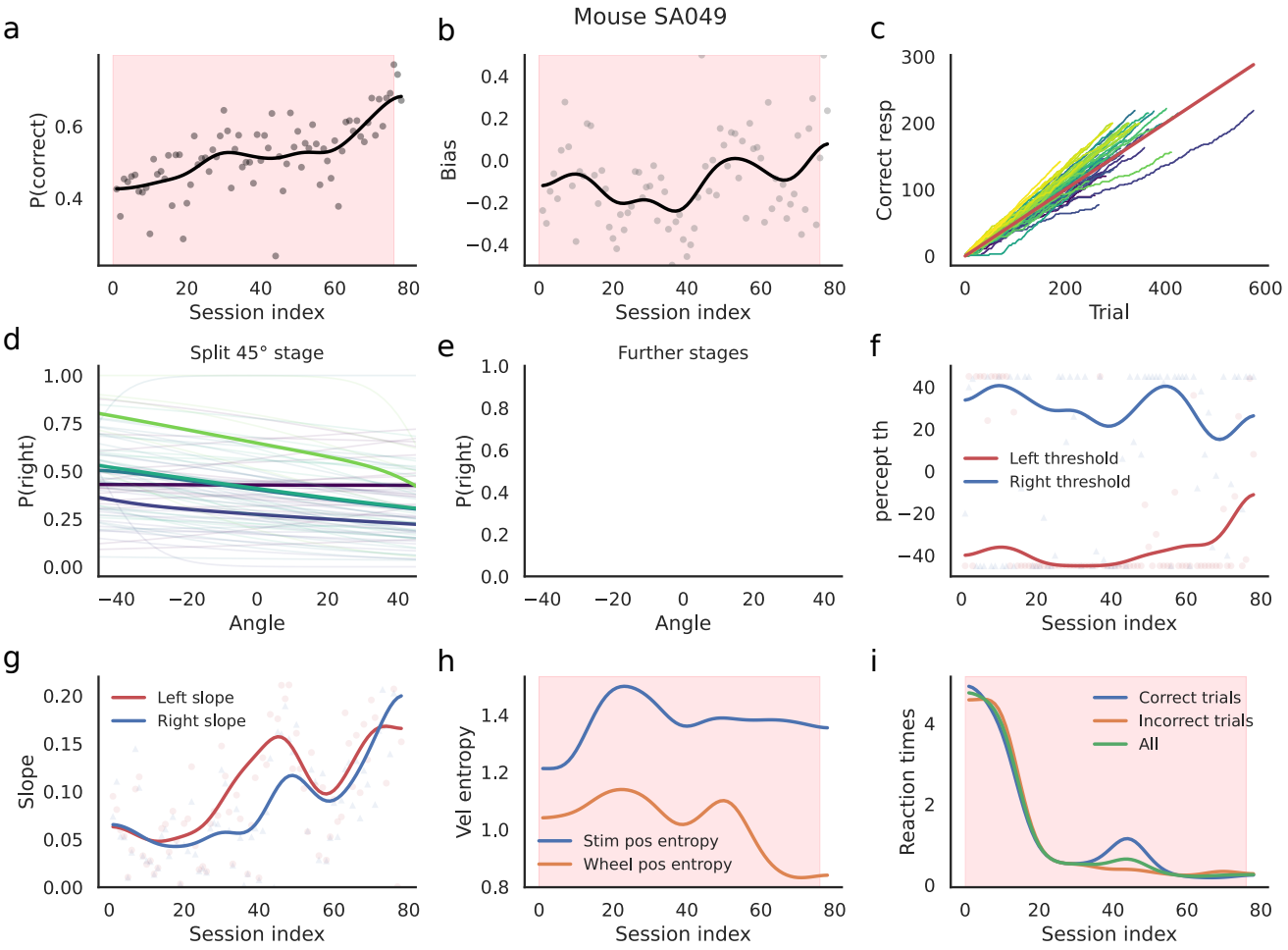



**Figure 6:** Mouse SA049 (learning)

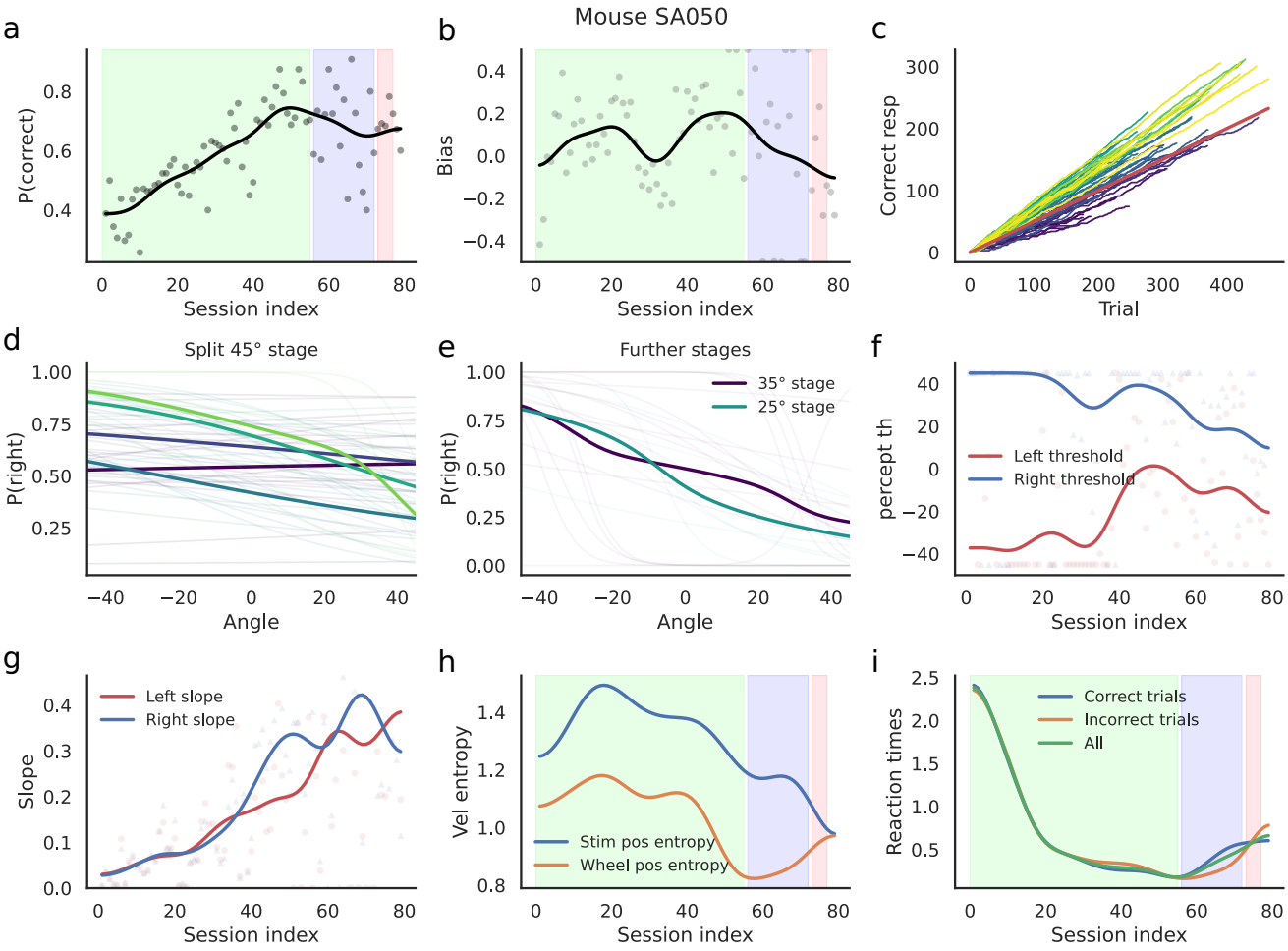



**Figure 7:** Mouse SA050 (learning)

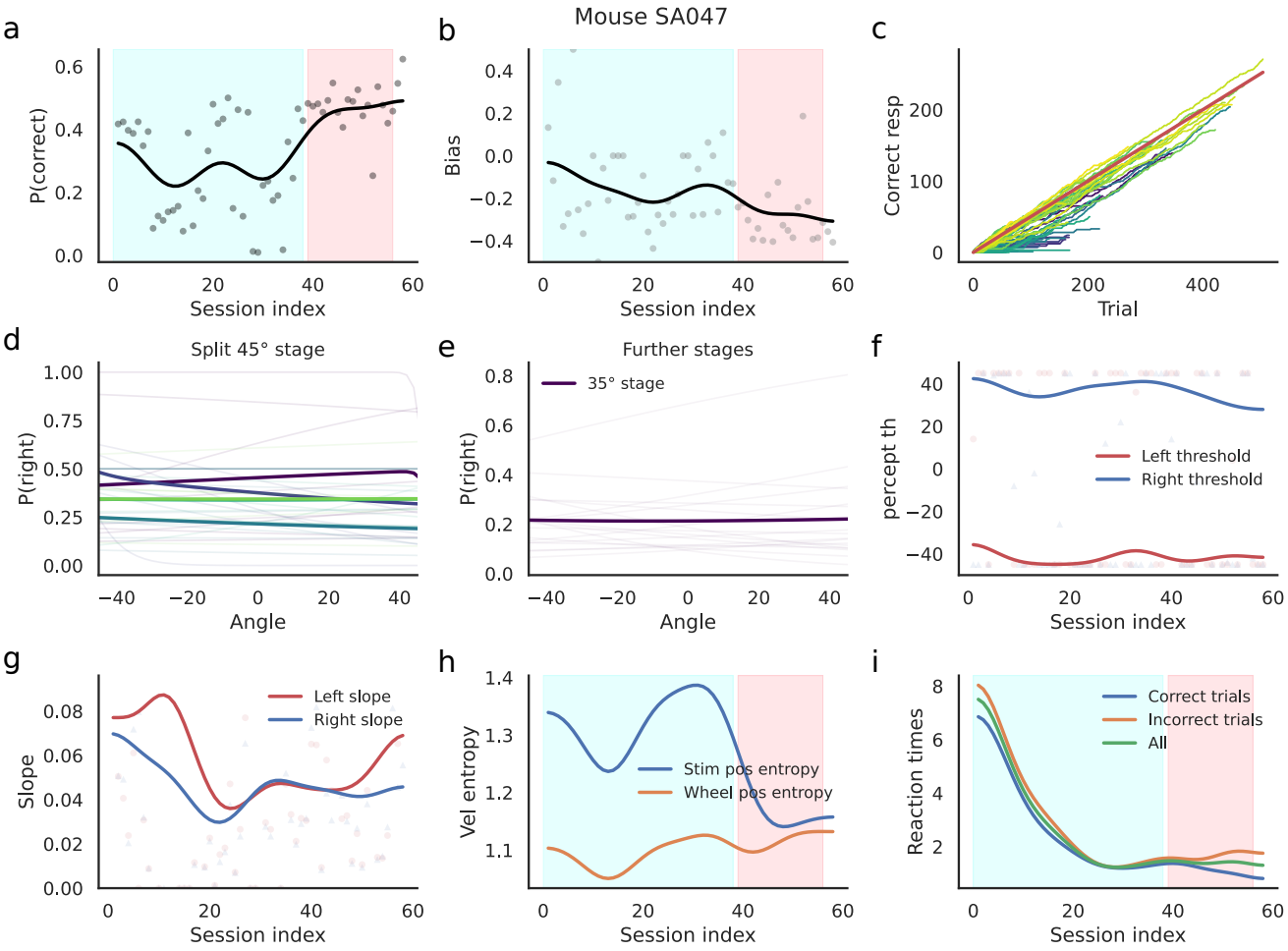



**Figure 8:** Mouse SA047 (non-performing, imaged after training)

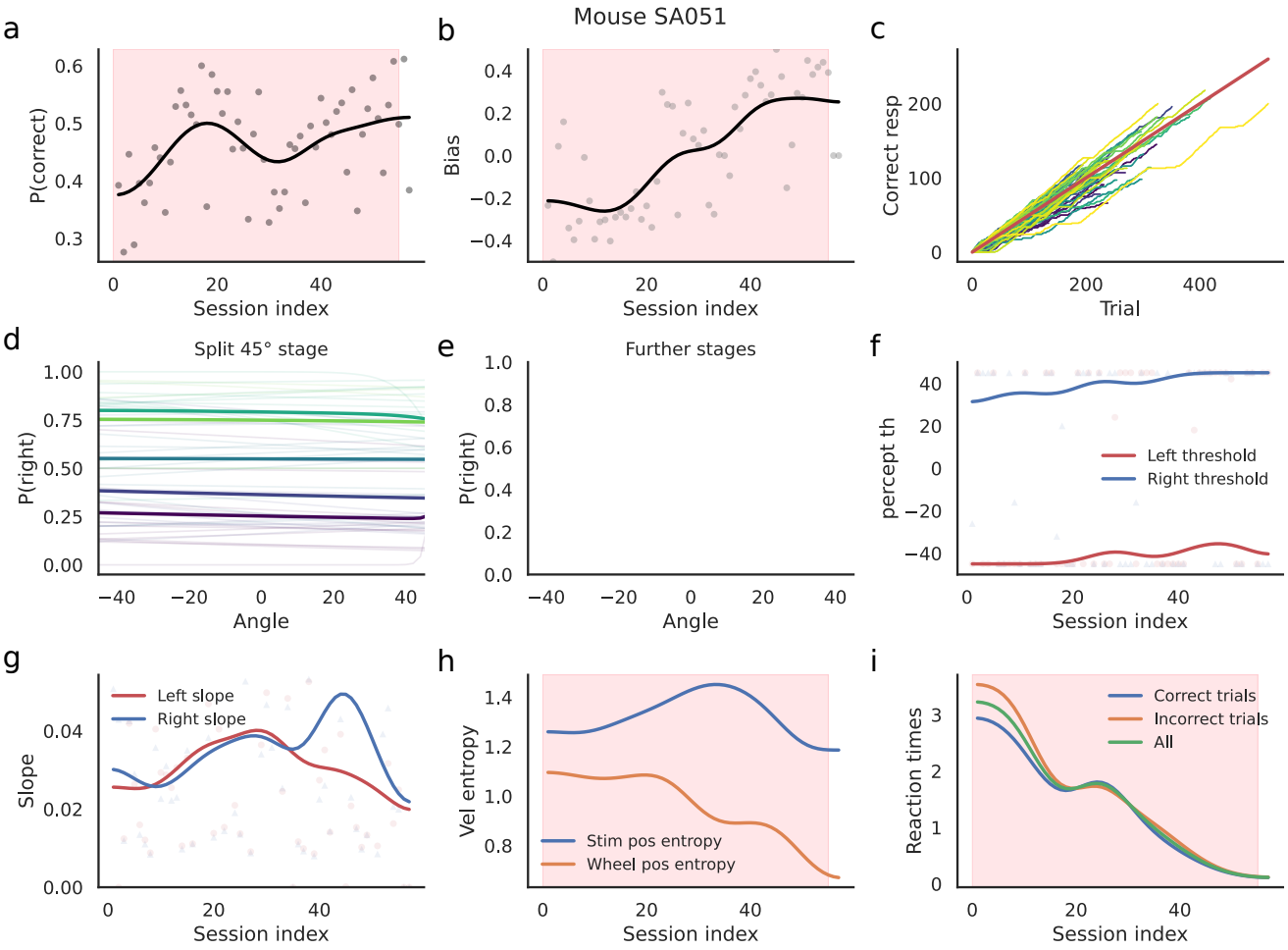



**Figure 9:** Mouse SA051 (non-performing, imaged after training)

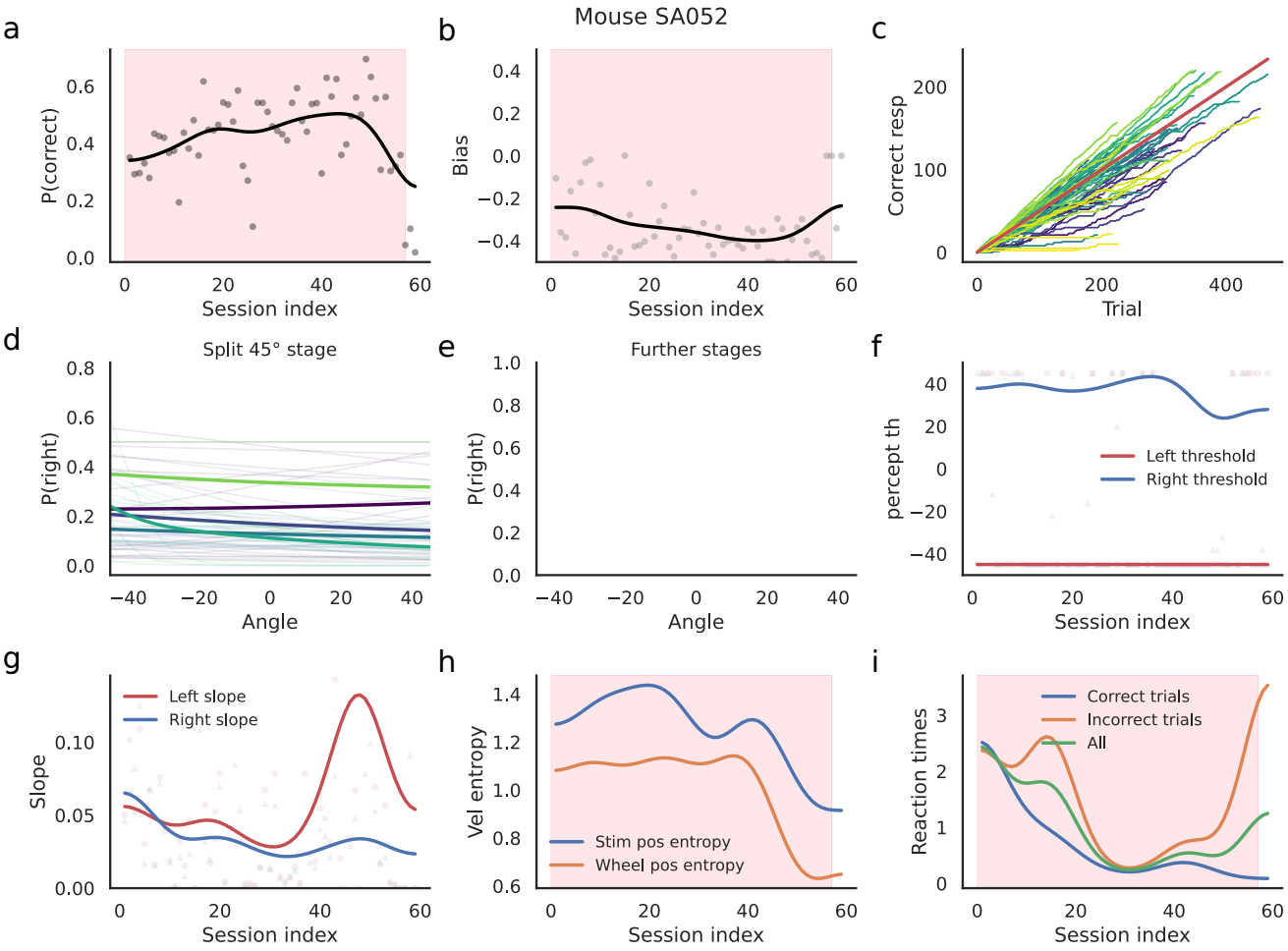



**Figure 10:** non-performing, imaged after training

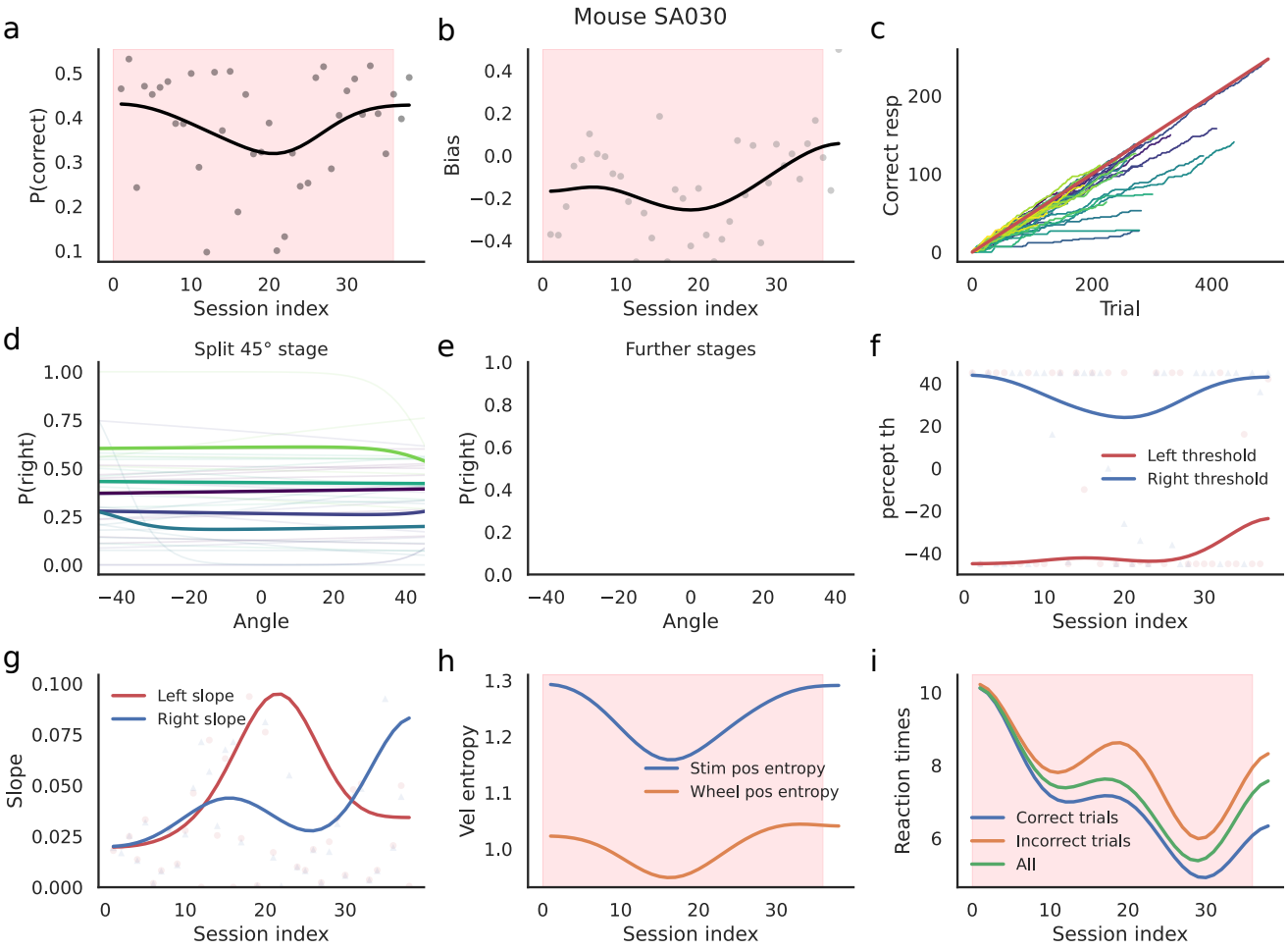



**Figure 11:** non-performing, imaged naive only

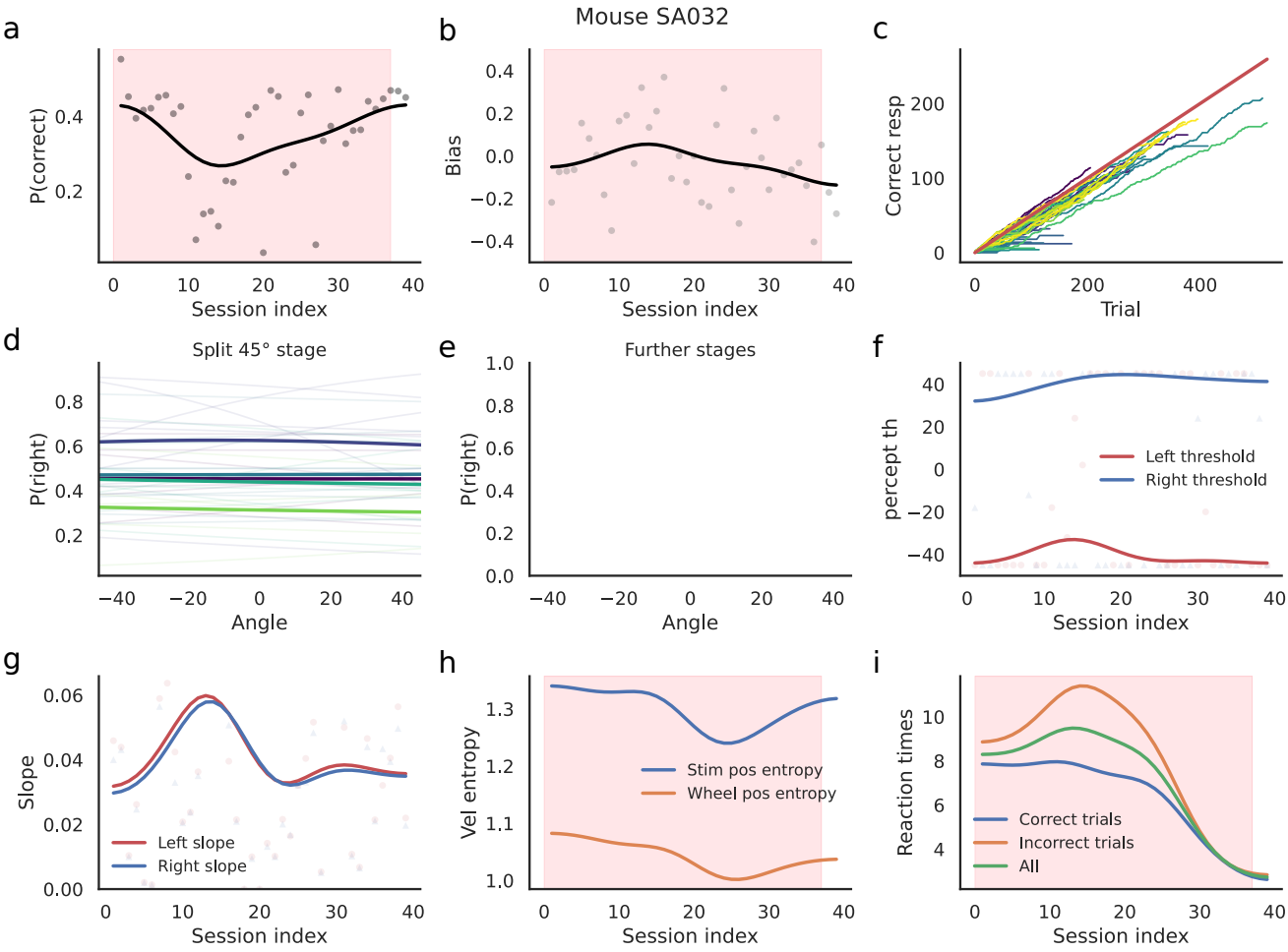



**Figure 12:** non-performing, imaged naive only

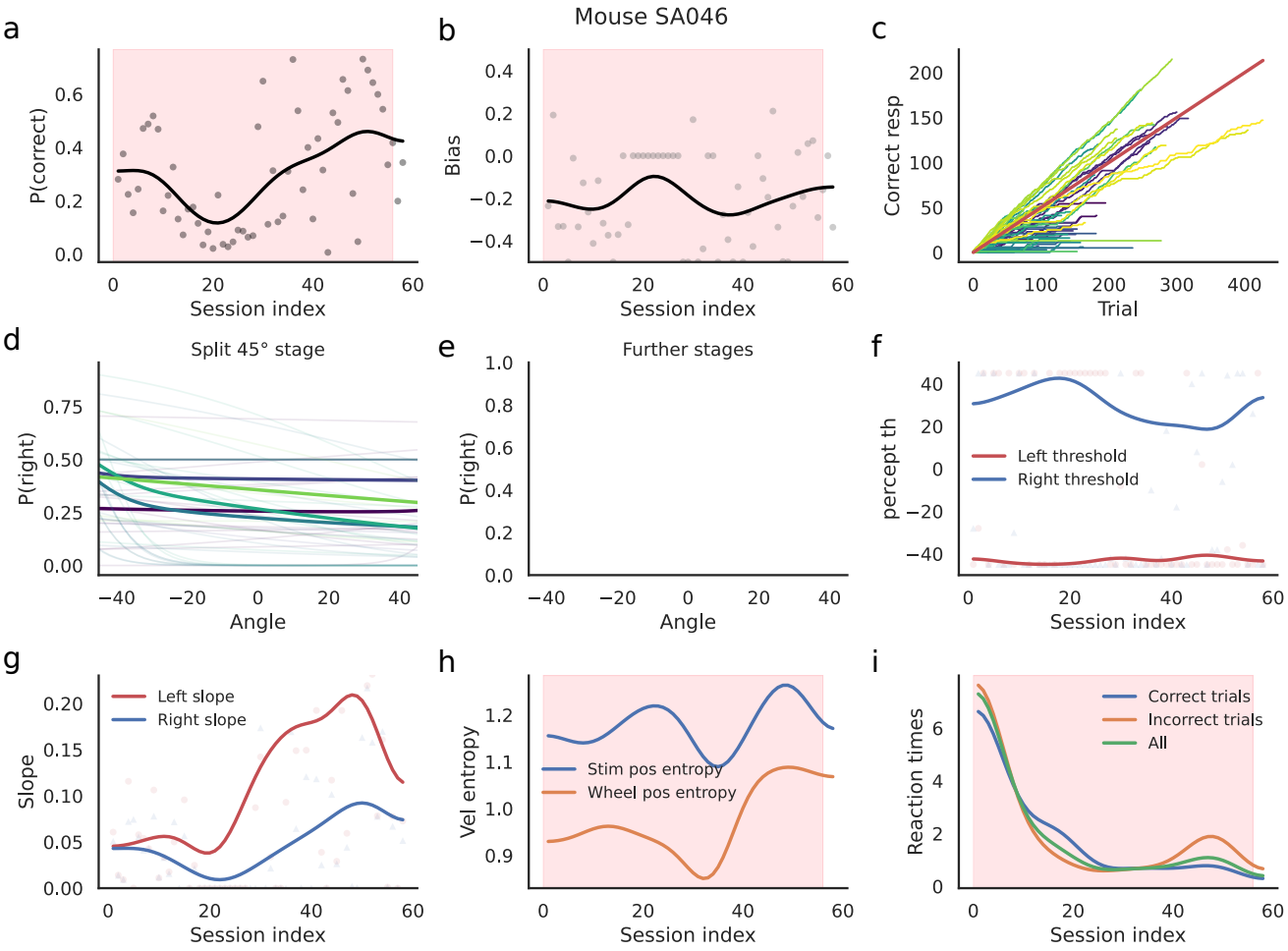



**Figure 13:** non-performing, imaged naive only

# .2 Supplementary figures for neural data

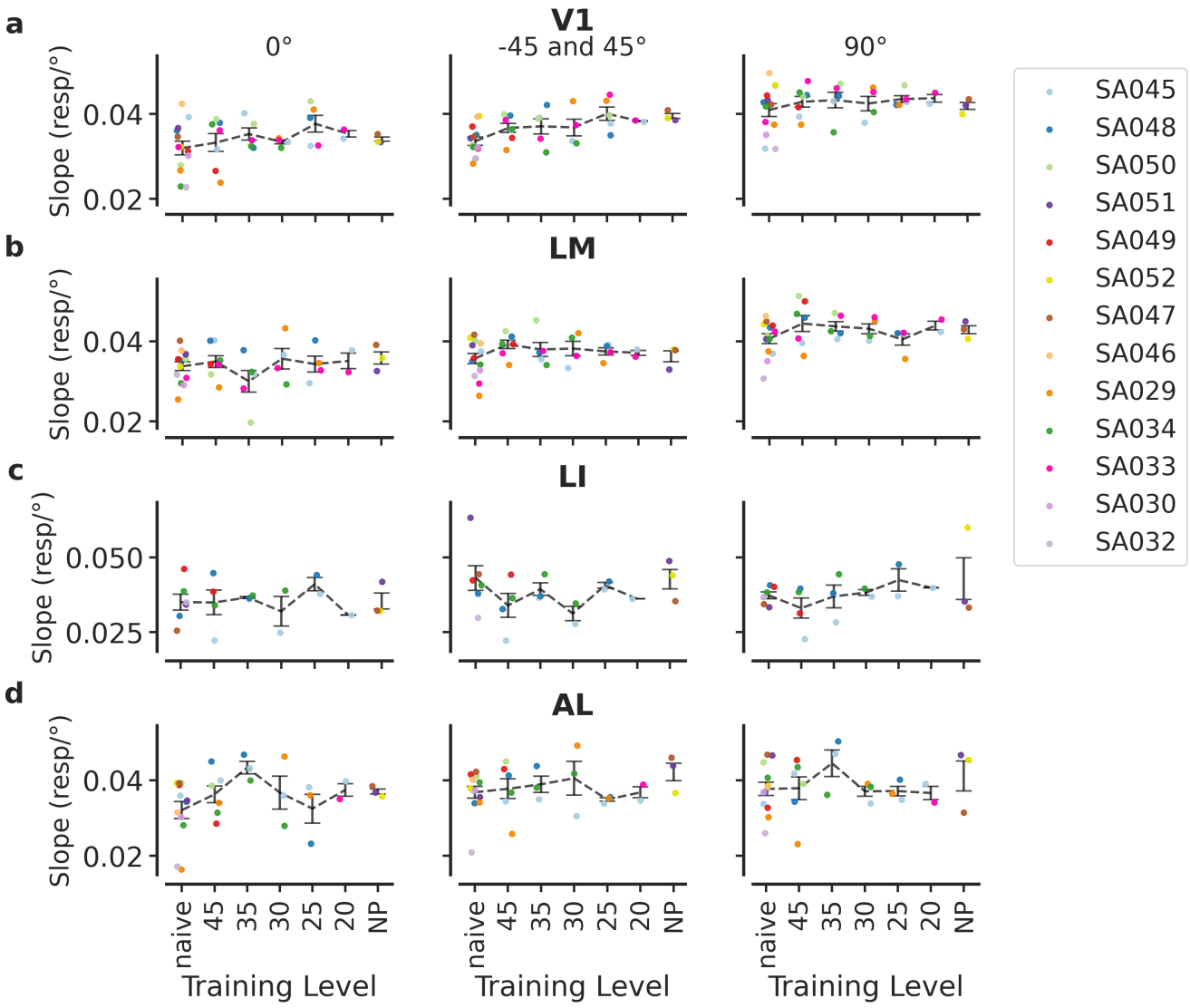



**Figure 14:** Raw slopes(**a**)

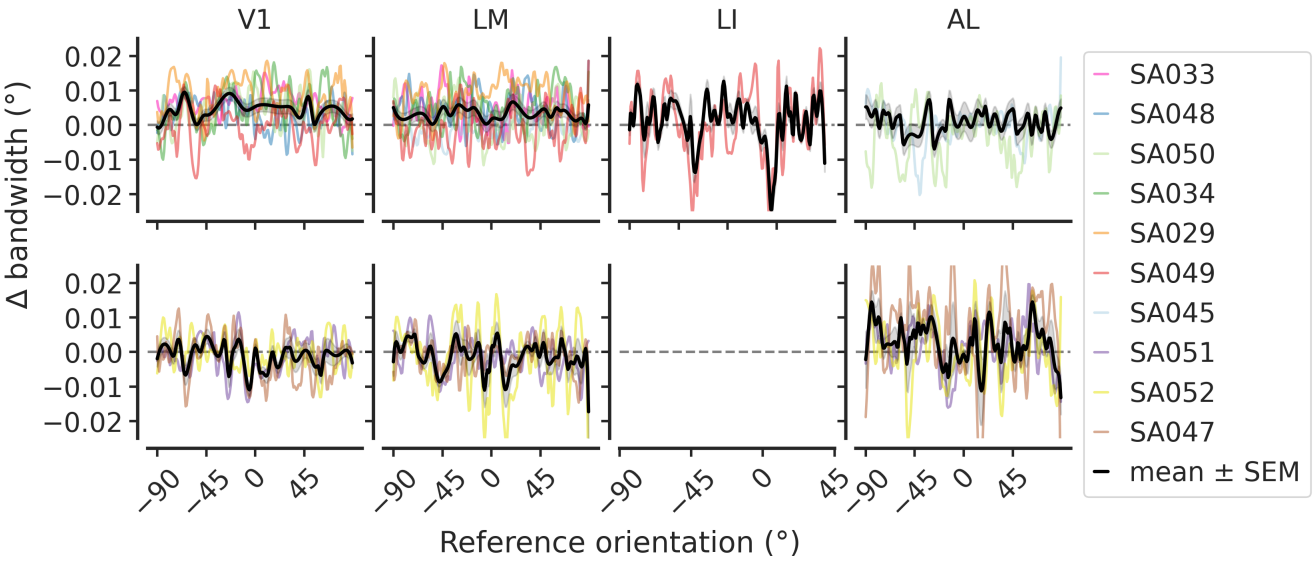



**Figure 15:** Slopes at different reference orientations for learning mice.

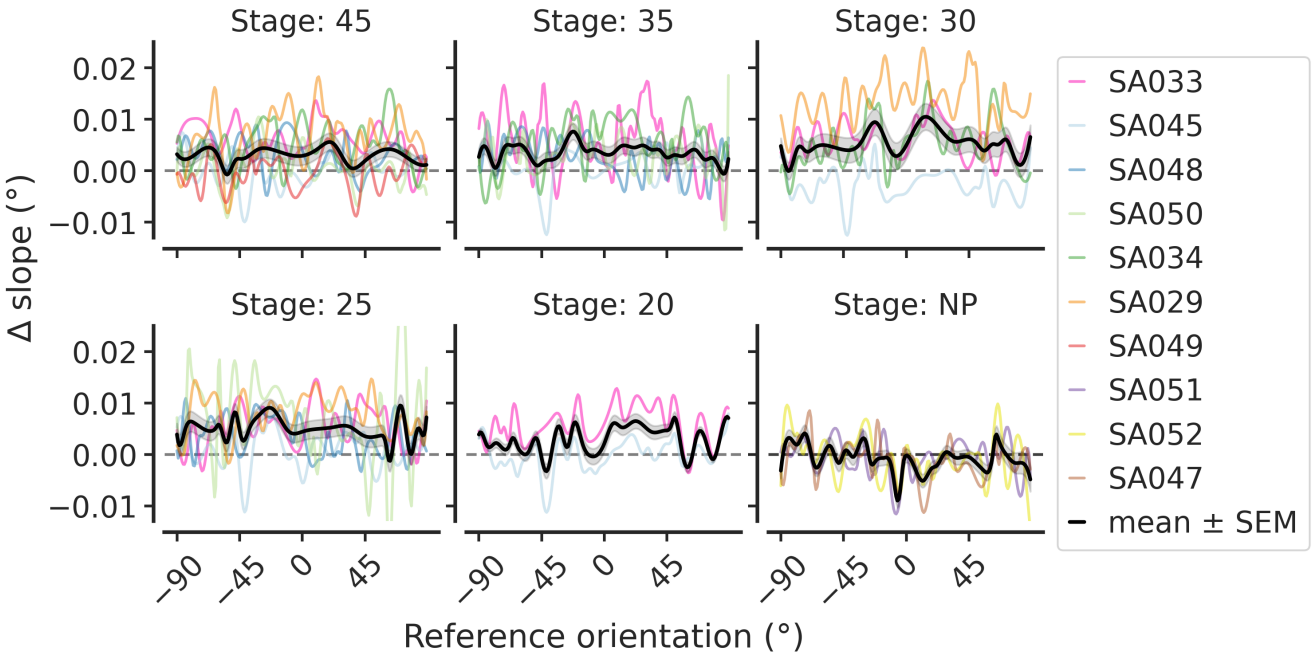



**Figure 16:** Slopes at different reference orientations for learning mice.

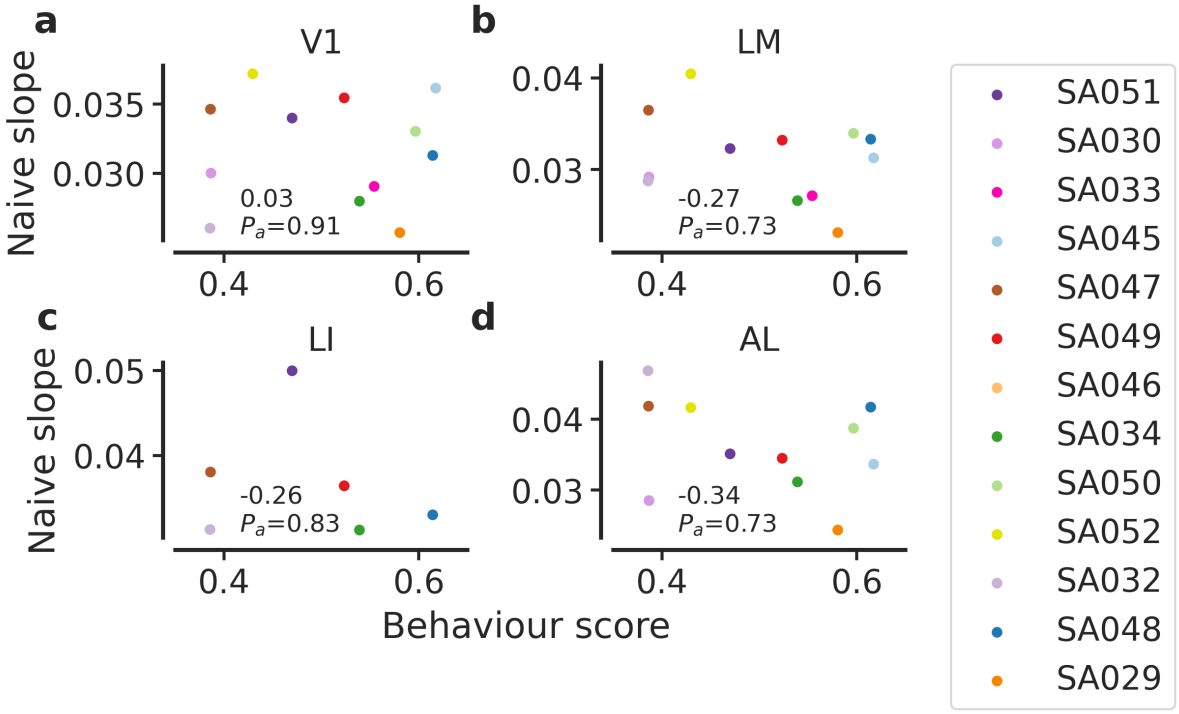



**Figure 17:** Each coloured data-point corresponds to one mouse. The legend shows the mouse ID. Note that data in different columns comprises different populations of neurons, because slopes were only taken for neurons whose PO was between ±3 and ±18 of the given reference orientation.)

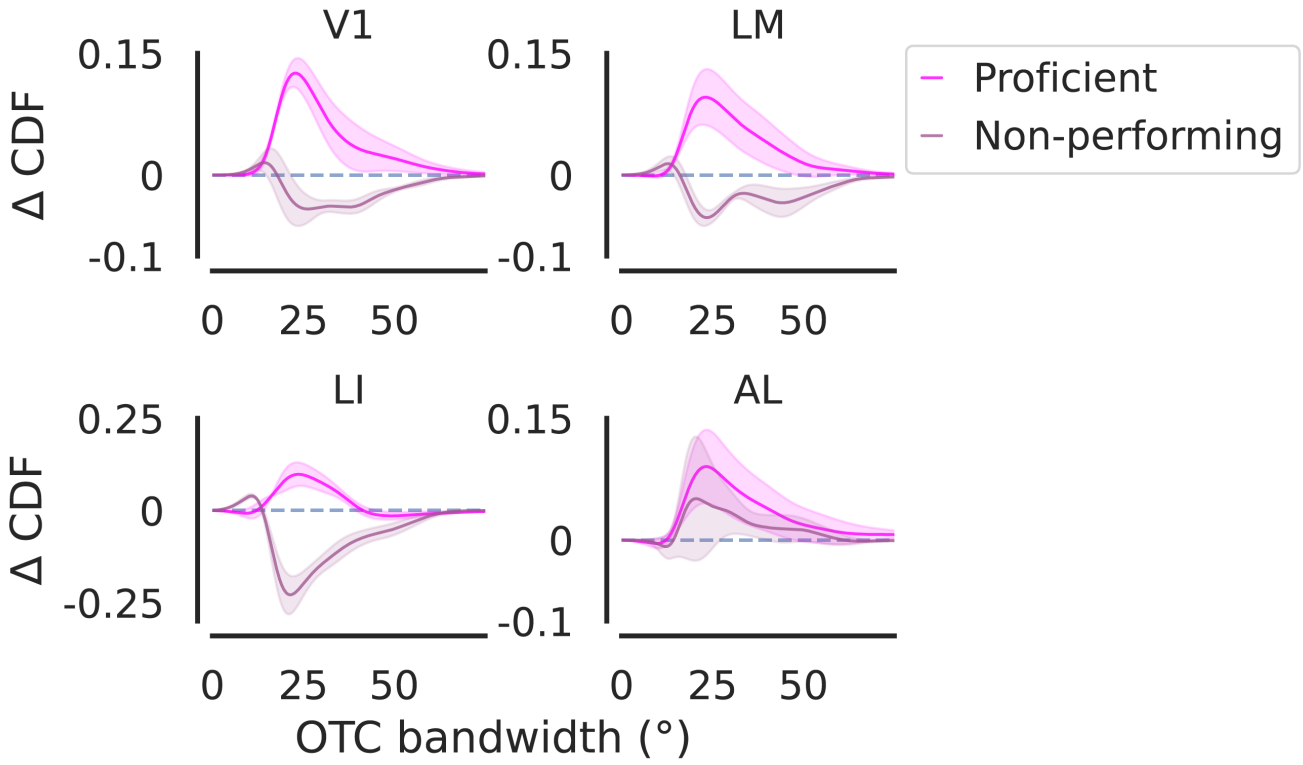



**Figure 18:** Orientation tuning curve (OTC) bandwidth distribution change from naive stage to proficient or non-performing stages. Change in cumulative distribution function (Δ CDF) of bandwidth from neurons measured at the naive stage to those measured at proficient stages in learning mice, or after training experience for non-performing mice. The Δ CDF represents the difference in cumulative probability of observing bandwidths up to a given value between the naive and later stages compared. Shading indicates 1 SEM across different mice.

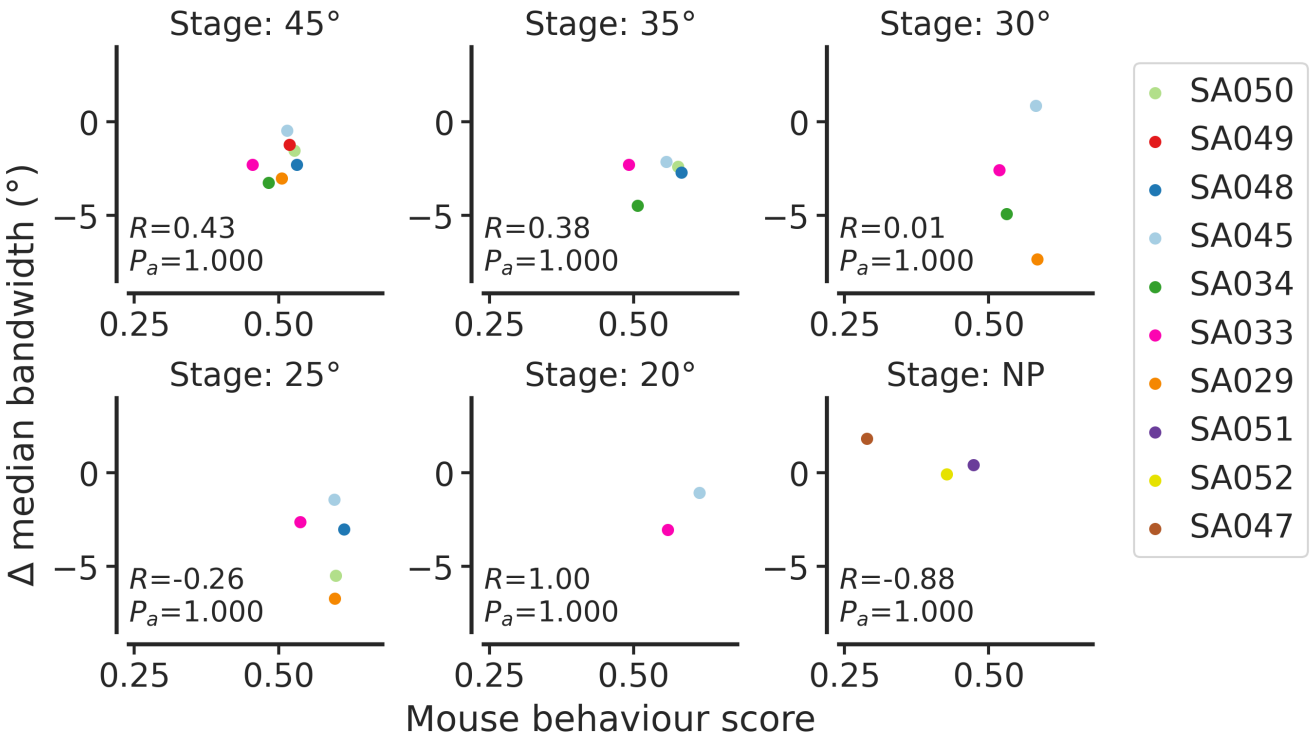



**Figure 19:** NP = non-performing.

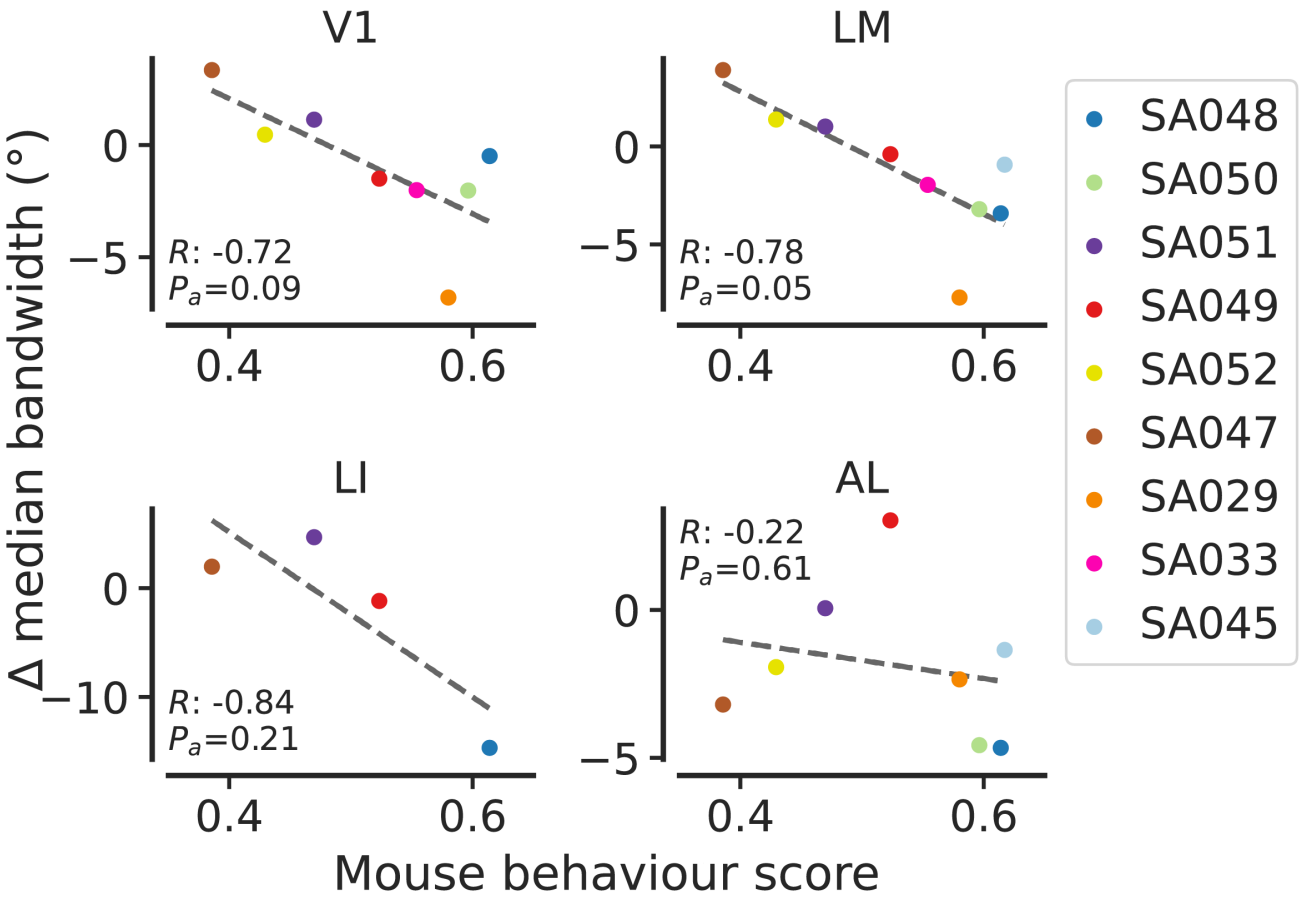



**Figure 20:** Orientation tuning curve (OTC) bandwidth distribution change from naive stage to proficient or non-performing stages. Change in cumulative distribution function (Δ CDF) of bandwidth from neurons measured at the naive stage to those measured at proficient stages in learning mice, or after training experience for non-performing mice. The Δ CDF represents the difference in cumulative probability of observing bandwidths up to a given value between the naive and later stages compared. Shading indicates 1 SEM across different mice.

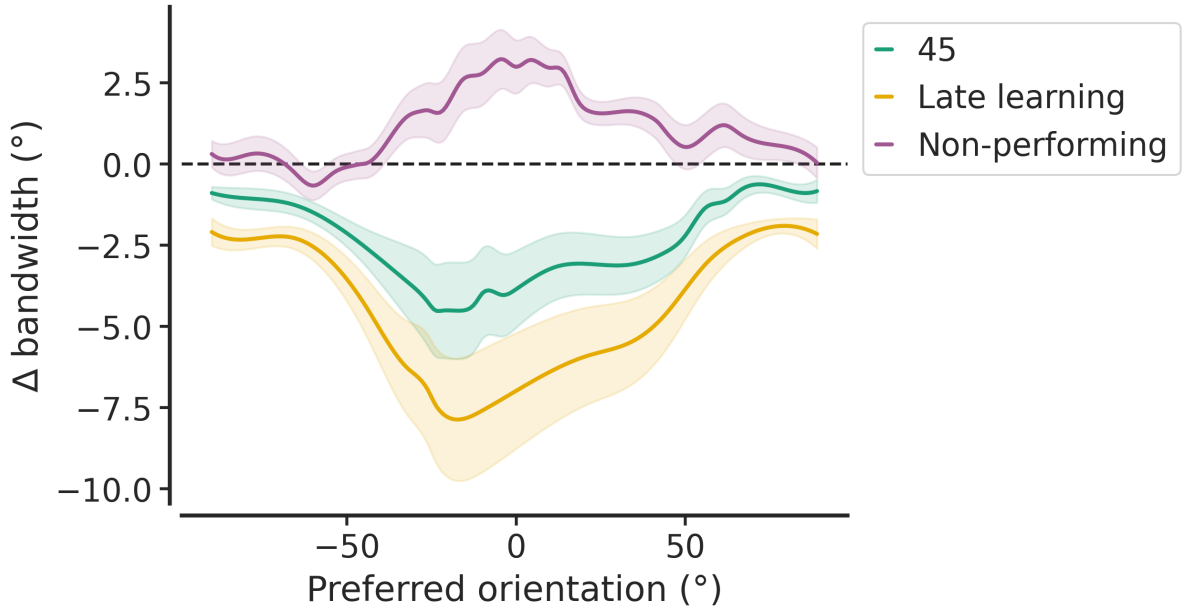



**Figure 21:** Change in median bandwidth as a function of preferred orientation, with all visual areas combined. For each preferred orientation, neurons with preferred orientation ± 16° were included in the calculation. Data is shown for the earliest proficient learning stage, 45 degrees (green), whilst data from later learning stages [35,30,25,20] were pooled together ('late learning'; yellow). The mean change across mice is displayed, around which shading denotes ± 1 SEM.

# Bibliography


1. Aberg, K. C. & Herzog, M. H. Different types of feedback change decision criterion and sensitivity differently in perceptual learning. *Journal of Vision* **12,** 3. ISSN: 1534-7362 (Mar. 2, 2012).
2. Ahissar, M. & Hochstein, S. The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences* **8.** Publisher: Elsevier, 457–464. ISSN: 1364-6613, 1879-307X (Oct. 1, 2004).
3. Aitchison, L. & Lengyel, M. With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology. Computational Neuroscience* **46,** 219–227. ISSN: 0959-4388 (Oct. 1, 2017).
4. Amit, Y. Deep Learning With Asymmetric Connections and Hebbian Updates. *Frontiers in Computational Neuroscience* **13.** Publisher: Frontiers. ISSN: 1662-5188 (Apr. 4, 2019).
5. Aslin, R. N. Statistical learning: a powerful mechanism that operates by mere exposure. *WIREs Cognitive Science* **8.** _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1373, e1373. ISSN: 1939-5086 (2017).
6. Assaf, Y., Bouznach, A., Zomet, O., Marom, A. & Yovel, Y. Conservation of brain connectivity and wiring across the mammalian class. *Nature Neuroscience* **23.** Publisher: Nature Publishing Group, 805–808. ISSN: 1546-1726 (July 2020).
7. Audette, N. J. & Schneider, D. M. Stimulus-Specific Prediction Error Neurons in Mouse Auditory Cortex. *Journal of Neuroscience* **43.** Publisher: Society for Neuroscience Section: Research Articles, 7119–7129. ISSN: 0270-6474, 1529-2401 (Oct. 25, 2023).
8. Baddeley, R. *et al.* Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **264.** Publisher: Royal Society, 1775–1783 (Dec. 22, 1997).
9. Ball, K. & Sekuler, R. A Specific and Enduring Improvement in Visual Motion Discrimination. *Science* **218.** Publisher: American Association for the Advancement of Science, 697–698 (Nov. 12, 1982).
10. Barlow. Possible principles underlying the transformation of sensory messages. *Sensory communication* **1.01,** 217–233 (1961).
11. Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* **76,** 695–711. ISSN: 1097-4199 (Nov. 2012).
12. Beck, J. M. *et al.* Probabilistic Population Codes for Bayesian Decision Making. *Neuron* **60.** Publisher: Elsevier, 1142–1152. ISSN: 0896-6273 (Dec. 26, 2008).
13. Beltramo, R. *et al.* Layer-specific excitatory circuits differentially control recurrent network dynamics in the neocortex. *Nature Neuroscience* **16.** Publisher: Nature Publishing Group, 227–234. ISSN: 1546-1726 (Feb. 2013).

14. Ben-Yishai, R., Bar-Or, R. L. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences* **92.** Publisher: National Academy of Sciences Section: Research Article, 3844–3848. ISSN: 0027-8424, 1091-6490 (Apr. 25, 1995).
15. Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T. & Lin, Z. *Towards Biologically Plausible Deep Learning* Aug. 9, 2016. arXiv: 1502.04156[cs].
16. Berridge, M. J. Neuronal Calcium Signaling. *Neuron* **21,** 13–26. ISSN: 0896-6273 (July 1, 1998).
17. Bhagat, J. *et al. Rigbox: an Open-Source Toolbox for Probing Neurons and Behavior* Oct. 4, 2019.
18. Bisley, J. W. The neural basis of visual attention. *The Journal of Physiology* **589,** 49–57. ISSN: 0022-3751 (Pt 1 Jan. 1, 2011).
19. Bolz, J. & Gilbert, C. D. Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature* **320.** Publisher: Nature Publishing Group, 362–365. ISSN: 1476-4687 (Mar. 1986).
20. Bonin, V., Histed, M. H., Yurgenson, S. & Reid, R. C. Local Diversity and Fine-Scale Organization of Receptive Fields in Mouse Visual Cortex. *Journal of Neuroscience* **31,** 18506–18521. ISSN: 0270-6474, 1529-2401 (Dec. 14, 2011).
21. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience* **8.** Publisher: Nature Publishing Group, 1263–1268. ISSN: 1546-1726 (Sept. 2005).
22. Breakspear, M. Dynamic models of large-scale brain activity. *Nature Neuroscience* **20.** Publisher: Nature Publishing Group, 340–352. ISSN: 1546-1726 (Mar. 2017).
23. Burgess, C. P. *et al.* High-Yield Methods for Accurate Two-Alternative Visual Psychophysics in Head-Fixed Mice. *Cell Reports* **20,** 2513–2524. ISSN: 2211-1247 (Sept. 5, 2017).
24. Campagnola, L. *et al.* Local connectivity and synaptic dynamics in mouse and human neocortex. *Science* **375.** Publisher: American Association for the Advancement of Science, eabj5861 (Mar. 11, 2022).
25. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nature Reviews Neuroscience* **13.** Publisher: Nature Publishing Group, 51–62. ISSN: 1471-0048 (Jan. 2012).
26. Chalk, M., Seitz, A. R. & Seriès, P. Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision* **10,** 2. ISSN: 1534-7362 (July 7, 2010).
27. Chapman, B. & Stryker, M. P. Development of orientation selectivity in ferret visual cortex and effects of deprivation. *Journal of Neuroscience* **13,** 5251–5262. ISSN: 0270-6474, 1529-2401 (Dec. 1, 1993).
28. Chen, T.-W. *et al.* Ultra-sensitive fluorescent proteins for imaging neuronal activity. *Nature* **499.** Publisher: Howard Hughes Medical Institute, 295 (July 7, 2013).
29. Chettih, S. N. & Harvey, C. D. Single-neuron perturbations reveal feature-specific competition in V1. *Nature* **567,** 334–340. ISSN: 1476-4687 (Mar. 2019).
30. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487.** Publisher: Nature Publishing Group, 51–56. ISSN: 1476-4687 (July 2012).
31. Citri, A. & Malenka, R. C. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. *Neuropsychopharmacology* **33.** Publisher: Nature Publishing Group, 18–41. ISSN: 1740-634X (Jan. 2008).

32. Cossell, L. *et al.* Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518,** 399–403. ISSN: 1476-4687 (Feb. 19, 2015).
33. Crist, R. E., Kapadia, M. K., Westheimer, G. & Gilbert, C. D. Perceptual Learning of Spatial Localization: Specificity for Orientation, Position, and Context. *Journal of Neurophysiology* **78.** Publisher: American Physiological Society, 2889–2894. ISSN: 0022-3077 (Dec. 1997).
34. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* **17.** Publisher: Nature Publishing Group, 1500–1509. ISSN: 1546-1726 (Nov. 2014).
35. D'Souza, R. D. *et al.* Hierarchical and nonhierarchical features of the mouse visual cortical network. *Nature Communications* **13,** 503. ISSN: 2041-1723 (Jan. 26, 2022).
36. Daie, K., Svoboda, K. & Druckmann, S. Targeted photostimulation uncovers circuit motifs supporting short-term memory. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 623785 (Apr. 30, 2019).
37. Dalgleish, H. W. P. *et al.* How many neurons are sufficient for perception of cortical activity? *eLife* **9** (ed Bathellier, B.) Publisher: eLife Sciences Publications, Ltd, e58889. ISSN: 2050-084X (Oct. 26, 2020).
38. Das, A. & Fiete, I. R. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience.* Publisher: Nature Publishing Group, 1–11. ISSN: 1546-1726 (Sept. 7, 2020).
39. Davis, G. W. & Bezprozvanny, I. Maintaining the Stability of Neural Function: A Homeostatic Hypothesis. *Annual Review of Physiology* **63.** Publisher: Annual Reviews, 847–869. ISSN: 0066-4278, 1545-1585 (Volume 63, 2001 Mar. 1, 2001).
40. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* Google-Books-ID: fLT4DwAAQBAJ. 477 pp. ISBN: 978-0-262-54185-5 (MIT Press, Aug. 12, 2005).
41. De Vries, S. E. J. *et al.* A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience* **23,** 138–151. ISSN: 1546-1726 (Jan. 2020).
42. De polavieja, G. G. Errors Drive the Evolution of Biological Signalling to Costly Codes. *Journal of Theoretical Biology* **214,** 657–664. ISSN: 0022-5193 (Feb. 21, 2002).
43. Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nature Neuroscience* **18.** Number: 9 Publisher: Nature Publishing Group, 1213–1225. ISSN: 1546-1726 (Sept. 2015).
44. Demany, L. Perceptual learning in frequency discrimination. *The Journal of the Acoustical Society of America* **78,** 1118–1120. ISSN: 0001-4966 (Sept. 1, 1985).
45. Denk, W., Strickler, J. H. & Webb, W. W. Two-Photon Laser Scanning Fluorescence Microscopy. *Science* **248.** Publisher: American Association for the Advancement of Science, 73–76 (Apr. 6, 1990).
46. Detorakis, G., Bartley, T. & Neftci, E. Contrastive Hebbian learning with random feedback weights. *Neural Networks* **114.** Publisher: Pergamon, 1–14. ISSN: 0893-6080 (June 1, 2019).
47. DeYoe, E. A. & Essen, D. C. V. Concurrent processing streams in monkey visual cortex. *Trends in Neurosciences* **11,** 219–226. ISSN: 0166-2236, 1878-108X (Jan. 1, 1988).
48. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How Does the Brain Solve Visual Object Recognition? *Neuron* **73.** Publisher: Elsevier, 415–434. ISSN: 0896-6273 (Feb. 9, 2012).

49. Doerig, A. *et al.* The neuroconnectionist research programme. *Nature Reviews Neuroscience* **24.** Publisher: Nature Publishing Group, 431–450. ISSN: 1471-0048 (July 2023).
50. Dosher, B. A., Jeter, P., Liu, J. & Lu, Z.-L. An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences* **110,** 13678–13683 (Aug. 13, 2013).
51. Dosher, B. A. & Lu, Z.-L. Mechanisms of perceptual learning. *Vision Research* **39,** 3197–3221. ISSN: 0042-6989 (Oct. 1, 1999).
52. Dosher, B. A. & Lu, Z.-L. Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences* **95,** 13988–13993 (Nov. 10, 1998).
53. Douglas, R. J., Martin, K. A. & Whitteridge, D. A Canonical Microcircuit for Neocortex. *Neural Computation* **1,** 480–488. ISSN: 0899-7667 (Dec. 1989).
54. Duncker, L., Driscoll, L., Shenoy, K. V., Sahani, M. & Sussillo, D. Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in neural information processing systems* **33,** 14387–14397 (2020).
55. Echeveste, R., Aitchison, L., Hennequin, G. & Lengyel, M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience* **23.** Publisher: Nature Publishing Group, 1138–1149. ISSN: 1546-1726 (Sept. 2020).
56. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415.** Publisher: Nature Publishing Group UK London, 429–433 (2002).
57. Fahey, P. G. *et al. A global map of orientation tuning in mouse visual cortex* Aug. 23, 2019.
58. Fahle, M. & Edelman, S. Long-term learning in vernier acuity: Effects of stimulus orientation, range and of feedback. *Vision Research* **33,** 397–412. ISSN: 0042-6989 (Feb. 1, 1993).
59. Failor, S. W., Carandini, M. & Harris, K. D. Visual experience orthogonalizes visual cortical stimulus responses via population code transformation. *Cell Reports* **44.** Publisher: Elsevier. ISSN: 2211-1247 (Feb. 25, 2025).
60. Fairhall, A. L., Lewen, G. D., Bialek, W. & de Ruyter van Steveninck, R. R. Efficiency and ambiguity in an adaptive neural code. *Nature* **412.** Publisher: Nature Publishing Group, 787–792. ISSN: 1476-4687 (Aug. 2001).
61. Favela, L. H. The dynamical renaissance in neuroscience. *Synthese* **199,** 2103–2127. ISSN: 0039-7857, 1573-0964 (Dec. 2021).
62. Feldman, D. E. Synaptic Mechanisms for Plasticity in Neocortex. *Annual Review of Neuroscience* **32.** Publisher: Annual Reviews, 33–55. ISSN: 0147-006X, 1545-4126 (Volume 32, 2009 July 21, 2009).
63. Feldman, J. A. & Ballard, D. H. Connectionist models and their properties. *Cognitive Science* **6,** 205–254. ISSN: 0364-0213 (July 1, 1982).
64. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, N.Y.* **1,** 1–47. ISSN: 1460-2199 (Jan. 1, 1991).
65. Fiorentini, A. & Berardi, N. Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research* **21,** 1149–1158. ISSN: 0042-6989 (Jan. 1, 1981).
66. FitzHugh, R. Impulses and Physiological States in Theoretical Models of Nerve Membrane. *Biophysical Journal* **1.** Publisher: Elsevier, 445–466. ISSN: 0006-3495, 1542-0086 (July 1, 1961).

67. Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLOS Computational Biology* **13,** e1005423. ISSN: 1553-7358 (Mar. 14, 2017).
68. Friston, K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360.** Publisher: Royal Society, 815–836 (Apr. 29, 2005).
69. Froudarakis, E. *et al.* The Visual Cortex in Context. *Annual Review of Vision Science* **5,** 317–339. ISSN: 2374-4642, 2374-4650 (Volume 5, 2019 Sept. 15, 2019).
70. Gal, E. *et al.* Rich cell-type-specific network topology in neocortical microcircuitry. *Nature Neuroscience* **20.** Publisher: Nature Publishing Group, 1004–1013. ISSN: 1546-1726 (July 2017).
71. Ghose, G. M., Yang, T. & Maunsell, J. H. R. Physiological Correlates of Perceptual Learning in Monkey V1 and V2. *Journal of Neurophysiology* **87.** Publisher: American Physiological Society, 1867–1888. ISSN: 0022-3077 (Apr. 1, 2002).
72. Glickfeld, L. L. & Olsen, S. R. Higher-Order Areas of the Mouse Visual Cortex. *Annual Review of Vision Science* **3,** 251–273. ISSN: 2374-4642, 2374-4650 (Volume 3, 2017 Sept. 15, 2017).
73. Glimcher, P. W. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences* **108.** Publisher: Proceedings of the National Academy of Sciences, 15647–15654 (supplement_3 Sept. 13, 2011).
74. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences* **117.** Publisher: National Academy of Sciences Section: Colloquium Paper, 29330–29337. ISSN: 0027-8424, 1091-6490 (Nov. 24, 2020).
75. Goltstein, P. M., Reinert, S., Bonhoeffer, T. & Hübener, M. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *Nature Neuroscience* **24,** 1441–1451. ISSN: 1546-1726 (Oct. 2021).
76. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* ISBN: 978-0-262-03561-3 (The MIT Press, 2016).
77. Gray, R. M. *Toeplitz and Circulant Matrices: A Review* Google-Books-ID: PrOi92L5dAUC. 105 pp. ISBN: 978-1-933019-23-9 (Now Publishers Inc, 2006).
78. Grienberger, C. & Konnerth, A. Imaging Calcium in Neurons. *Neuron* **73,** 862–885. ISSN: 0896-6273 (Mar. 8, 2012).
79. Grosof, D. H., Shapley, R. M. & Hawken, M. J. Macaque VI neurons can signal 'illusory' contours. *Nature* **365.** Publisher: Nature Publishing Group, 550–552. ISSN: 1476-4687 (Oct. 1993).
80. Güçlü, U. & Gerven, M. A. J. v. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience* **35.** Publisher: Society for Neuroscience Section: Articles, 10005–10014. ISSN: 0270-6474, 1529-2401 (July 8, 2015).
81. Guerguiev, J., Lillicrap, T. P. & Richards, B. A. Towards deep learning with segregated dendrites. *elife* **6.** Publisher: eLife Sciences Publications, Ltd, e22901 (2017).
82. Harris, K. D. & Mrsic-Flogel, T. D. Cortical connectivity and sensory coding. *Nature* **503.** Publisher: Nature Publishing Group, 51–58. ISSN: 1476-4687 (Nov. 2013).
83. Harris, K. D. & Shepherd, G. M. G. The neocortical circuit: themes and variations. *Nature Neuroscience* **18.** Publisher: Nature Publishing Group, 170–181. ISSN: 1546-1726 (Feb. 2015).
84. Helmholtz, H. v. *Handbuch der physiologischen Optik* Google-Books-ID: E3EZAAAAYAAJ. 922 pp. (L. Voss, 1867).

85. Hennequin, G., Vogels, T. P. & Gerstner, W. Non-normal amplification in random balanced neuronal networks. *Physical Review E* **86.** Publisher: American Physical Society, 011909 (July 11, 2012).

86. Hennig, M. H. The sloppy relationship between neural circuit structure and function. *The Journal of Physiology* **601.** _eprint: https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/JP282757, 3025–3035. ISSN: 1469-7793 (2023).

87. Hensch, T. K. CRITICAL PERIOD REGULATION. *Annual Review of Neuroscience* **27.** Publisher: Annual Reviews, 549–579. ISSN: 0147-006X, 1545-4126 (Volume 27, 2004 July 21, 2004).

88. Herzog, M. H. & Fahle, M. The role of feedback in learning a vernier discrimination task. *Vision Research* **37,** 2133–2141. ISSN: 0042-6989 (Aug. 1, 1997).

89. Ho, B. L. Effective construction of linear state-variable models from input/output functions. *Regelungstechnik* **14,** 545–548 (1966).

90. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology* **117,** 500–544. ISSN: 0022-3751 (Aug. 28, 1952).

91. Holtmaat, A. & Svoboda, K. Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience* **10.** Publisher: Nature Publishing Group, 647–658. ISSN: 1471-0048 (Sept. 2009).

92. Honey, C. J., Thivierge, J.-P. & Sporns, O. Can structure predict function in the human brain? *NeuroImage. Computational Models of the Brain* **52,** 766–776. ISSN: 1053-8119 (Sept. 1, 2010).

93. Huang, L. *et al.* Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. *eLife* **10** (eds Westbrook, G. L., Svoboda, K., Higley, M. & Sabatini, B. L.) e51675. ISSN: 2050-084X (Mar. 8, 2021).

94. Huang, Y. & Rao, R. P. N. Predictive coding. *WIREs Cognitive Science* **2.** _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.142, 580–593. ISSN: 1939-5086 (2011).

95. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology* **195,** 215–243. ISSN: 1469-7793 (1968).

96. Humeau, Y. & Choquet, D. The next generation of approaches to investigate the link between synaptic plasticity and learning. *Nature Neuroscience* **22.** Publisher: Nature Publishing Group, 1536–1543. ISSN: 1546-1726 (Oct. 2019).

97. Hunnicutt, B. J. *et al.* A comprehensive thalamocortical projection map at the mesoscopic level. *Nature neuroscience* **17,** 1276–1285. ISSN: 1097-6256 (Sept. 2014).

98. Isaac, J. T. R., Crair, M. C., Nicoll, R. A. & Malenka, R. C. Silent Synapses during Development of Thalamocortical Inputs. *Neuron* **18.** Publisher: Elsevier, 269–280. ISSN: 0896-6273 (Feb. 1, 1997).

99. Izhikevich, E. M. *Dynamical Systems in Neuroscience* 522 pp. ISBN: 978-0-262-09043-8 (MIT Press, 2007).

100. Jacot, A., Gabriel, F. & Hongler, C. *Neural Tangent Kernel: Convergence and Generalization in Neural Networks* in *Advances in Neural Information Processing Systems* **31** (Curran Associates, Inc., 2018).

101. Jin, M. & Glickfeld, L. L. Mouse Higher Visual Areas Provide Both Distributed and Specialized Contributions to Visually Guided Behaviors. *Current Biology* **30.** Publisher: Elsevier, 4682–4692.e7. ISSN: 0960-9822 (Dec. 7, 2020).

102. Kaas, J. H. *Evolutionary Neuroscience* Google-Books-ID: 8TjoDwAAQBAJ. 964 pp. ISBN: 978-0-12-820606-5 (Academic Press, May 30, 2020).

103. Kailath, T. *Linear systems* (Prentice-Hall, 1980).

104. Kalman, R. E. Canonical structure of linear dynamical systems. *Proceedings of the National Academy of Sciences* **48.** Publisher: Proceedings of the National Academy of Sciences, 596–600 (Apr. 1962).

105. Kandel Jessell, S. *Principles of Neural Science* 6th (Mar. 8, 2021).

106. Karni, A. & Sagi, D. Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences* **88.** Publisher: Proceedings of the National Academy of Sciences, 4966–4970 (June 1991).

107. Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. M. & Sagi, D. Dependence on REM Sleep of Overnight Improvement of a Perceptual Skill. *Science* **265.** Publisher: American Association for the Advancement of Science, 679–682 (July 29, 1994).

108. Katok, A., Katok, A. B. & Hasselblatt, B. *Introduction to the Modern Theory of Dynamical Systems* Google-Books-ID: 9nL7ZX8Djp4C. 828 pp. ISBN: 978-0-521-57557-7 (Cambridge University Press, 1995).

109. Katz, L. C. & Shatz, C. J. Synaptic Activity and the Construction of Cortical Circuits. *Science* **274.** Publisher: American Association for the Advancement of Science, 1133–1138 (Nov. 15, 1996).

110. Keller, G. B. & Mrsic-Flogel, T. D. Predictive Processing: A Canonical Cortical Computation. *Neuron* **100,** 424–435. ISSN: 08966273 (Oct. 2018).

111. Kerr, A. L., Cheng, S.-Y. & Jones, T. A. Experience-dependent neural plasticity in the adult damaged brain. *Journal of Communication Disorders*, S0021992411000335. ISSN: 00219924 (May 2011).

112. Kersten, D., Mamassian, P. & Yuille, A. Object Perception as Bayesian Inference. *Annual Review of Psychology* **55.** Publisher: Annual Reviews, 271–304. ISSN: 0066-4308, 1545-2085 (Volume 55, 2004 Jan. 1, 2004).

113. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology* **10.** Publisher: Public Library of Science, e1003915. ISSN: 1553-7358 (Nov. 6, 2014).

114. Khan, A. G. *et al.* Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. *Nature Neuroscience* **21.** Publisher: Nature Publishing Group, 851–859. ISSN: 1546-1726 (June 2018).

115. Kingdom, F. A. A. & Prins, N. *Psychophysics: A Practical Introduction* Google-Books-ID: 3sHQBAAAQBAJ. 348 pp. ISBN: 978-0-08-099381-2 (Academic Press, Jan. 4, 2016).

116. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences* **27.** Publisher: Elsevier, 712–719. ISSN: 0166-2236, 1878-108X (Dec. 1, 2004).

117. Knill, D. C. & Richards, W. *Perception as Bayesian Inference* Google-Books-ID: _cTLCgAAQBAJ. 530 pp. ISBN: 978-1-316-58252-7 (Cambridge University Press, Sept. 13, 1996).

118. Ko, H. *et al.* Functional specificity of local synaptic connections in neocortical networks. *Nature* **473,** 87–91. ISSN: 0028-0836 (May 5, 2011).

119. Kohn, A. Visual Adaptation: Physiology, Mechanisms, and Functional Benefits. *Journal of Neurophysiology* **97.** Publisher: American Physiological Society, 3155–3164. ISSN: 0022-3077 (May 2007).

120. Konishi, M., Igarashi, K. M. & Miura, K. Biologically plausible local synaptic learning rules robustly implement deep supervised learning. *Frontiers in Neuroscience* **17.** Publisher: Frontiers. ISSN: 1662-453X (Oct. 11, 2023).

121. Kreile, A. K., Bonhoeffer, T. & Hübener, M. Altered Visual Experience Induces Instructive Changes of Orientation Preference in Mouse Visual Cortex. *Journal of Neuroscience* **31,** 13911–13920. ISSN: 0270-6474, 1529-2401 (Sept. 28, 2011).

122. Kress, G. J. *et al.* Convergent cortical innervation of striatal projection neurons. *Nature Neuroscience* **16,** 665–667. ISSN: 1546-1726 (June 2013).

123. Kriegeskorte, N. & Golan, T. Neural network models and deep learning. *Current Biology* **29.** Publisher: Elsevier, R231–R236. ISSN: 0960-9822 (Apr. 1, 2019).

124. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *ImageNet Classification with Deep Convolutional Neural Networks* in *Advances in Neural Information Processing Systems* **25** (Curran Associates, Inc., 2012).

125. Kuchibhotla, K. V. *et al.* Dissociating task acquisition from expression during learning reveals latent knowledge. *Nature Communications* **10.** Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Decision;Learning and memory Subject_term_id: decision;learning-and-memory, 2151. ISSN: 2041-1723 (May 14, 2019).

126. Laboy-Juárez, K. J., Langberg, T., Ahn, S. & Feldman, D. E. Elementary motion sequence detectors in whisker somatosensory cortex. *Nature Neuroscience* **22,** 1438–1449. ISSN: 1546-1726 (Sept. 2019).

127. Lange, F. P. d., Heilbron, M. & Kok, P. How Do Expectations Shape Perception? *Trends in Cognitive Sciences* **22.** Publisher: Elsevier, 764–779. ISSN: 1364-6613, 1879-307X (Sept. 1, 2018).

128. Lappalainen, J. K. *et al.* Connectome-constrained networks predict neural activity across the fly visual system. *Nature* **634.** Publisher: Nature Publishing Group, 1132–1140. ISSN: 1476-4687 (Oct. 2024).

129. Law, C.-T. & Gold, J. I. Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nature Neuroscience* **11.** Publisher: Nature Publishing Group, 505–513. ISSN: 1546-1726 (Apr. 2008).

130. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361.** Publisher: Cambridge, MA USA, 1995 (1995).

131. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521.** Publisher: Nature Publishing Group, 436–444. ISSN: 1476-4687 (May 2015).

132. Levy, R. B. & Reyes, A. D. Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **32,** 5609–5619. ISSN: 1529-2401 (Apr. 18, 2012).

133. Lewicki, M. S. Efficient coding of natural sounds. *Nature Neuroscience* **5.** Publisher: Nature Publishing Group, 356–363. ISSN: 1546-1726 (Apr. 2002).

134. Liebana, S. *et al.* Dopamine encodes deep network teaching signals for individual learning trajectories. *Cell* **0.** Publisher: Elsevier. ISSN: 0092-8674, 1097-4172 (June 11, 2025).

135. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications* **7.** Number: 1 Publisher: Nature Publishing Group, 13276. ISSN: 2041-1723 (Nov. 8, 2016).

136. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nature Reviews Neuroscience* **21.** Publisher: Nature Publishing Group, 335–346. ISSN: 1471-0048 (June 2020).

137. Luo, L., Callaway, E. M. & Svoboda, K. Genetic Dissection of Neural Circuits: A Decade of Progress. *Neuron* **98.** Publisher: Elsevier, 256–281. ISSN: 0896-6273 (Apr. 18, 2018).

138. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9.** Publisher: Nature Publishing Group, 1432–1438. ISSN: 1546-1726 (Nov. 2006).

139. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience* **10.** Publisher: Frontiers. ISSN: 1662-5188 (Sept. 14, 2016).

140. Marr, D. *Vision: A computational investigation into the human representation and processing of visual information* (WH Freeman, San Francisco, 1982).

141. Mars, R. B. *et al.* Comparing brains by matching connectivity profiles. *Neuroscience & Biobehavioral Reviews* **60,** 90–97. ISSN: 0149-7634 (Jan. 1, 2016).

142. Marshel, J. H., Garrett, M. E., Nauhaus, I. & Callaway, E. M. Functional Specialization of Seven Mouse Visual Cortical Areas. *Neuron* **72,** 1040–1054. ISSN: 0896-6273 (Dec. 22, 2011).

143. Marshel, J. H., Kim, Y. S., *et al.* Cortical layer-specific critical dynamics triggering perception. *Science (New York, N.Y.)* **365,** eaaw5202. ISSN: 1095-9203 (Aug. 9, 2019).

144. Mastrogiuseppe, F. & Ostojic, S. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron* **99.** Publisher: Elsevier, 609–623.e29. ISSN: 0896-6273 (Aug. 8, 2018).

145. Max, K. *et al.* Learning efficient backprojections across cortical hierarchies in real time. *Nature Machine Intelligence* **6.** Publisher: Nature Publishing Group, 619–630. ISSN: 2522-5839 (June 2024).

146. May, A. Experience-dependent structural plasticity in the adult human brain. *Trends in Cognitive Sciences* **15.** Publisher: Elsevier, 475–482. ISSN: 1364-6613, 1879-307X (Oct. 1, 2011).

147. McClelland, J. L. Connectionist models and psychological evidence. *Journal of Memory and Language* **27,** 107–123. ISSN: 0749-596X (Apr. 1, 1988).

148. McGann, J. P. Associative learning and sensory neuroplasticity: how does it happen and what is it good for? *Learning & Memory* **22.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 567–576. ISSN: 1072-0502-01, 1549-5485 (Nov. 1, 2015).

149. Mikulasch, F. A., Rudelt, L. & Priesemann, V. Visuomotor Mismatch Responses as a Hallmark of Explaining Away in Causal Inference. *Neural Computation* **35,** 27–37. ISSN: 0899-7667 (Jan. 1, 2023).

150. Miller, K. D. Canonical computations of cerebral cortex. *Current Opinion in Neurobiology. Neurobiology of cognitive behavior* **37,** 75–84. ISSN: 0959-4388 (Apr. 1, 2016).

151. Miyamoto, D. *et al.* Top-down cortical input during NREM sleep consolidates perceptual memory. *Science* **352.** Publisher: American Association for the Advancement of Science, 1315–1318 (June 10, 2016).

152. Montúfar, G., Pascanu, R., Cho, K. & Bengio, Y. *On the Number of Linear Regions of Deep Neural Networks* in *Advances in Neural Information Processing Systems* **27** (Curran Associates, Inc., 2014).

153. Murphy, B. K. & Miller, K. D. Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns. *Neuron* **61,** 635–648. ISSN: 0896-6273 (Feb. 26, 2009).

154. Muzzu, T. & Saleem, A. B. Feature selectivity can explain mismatch signals in mouse visual cortex. *Cell Reports* **37,** 109772. ISSN: 2211-1247 (Oct. 5, 2021).

155. Näätänen, R., Paavilainen, P., Rinne, T. & Alho, K. The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology* **118,** 2544–2590. ISSN: 1388-2457 (Dec. 1, 2007).

156. Nagumo, J., Arimoto, S. & Yoshizawa, S. An Active Pulse Transmission Line Simulating Nerve Axon. *Proceedings of the IRE* **50,** 2061–2070. ISSN: 2162-6634 (Oct. 1962).

157. Niell, C. M. & Stryker, M. P. Highly Selective Receptive Fields in Mouse Visual Cortex. *Journal of Neuroscience* **28,** 7520–7536. ISSN: 0270-6474, 1529-2401 (July 23, 2008).

158. Nise, N. S. *Control Systems Engineering* Google-Books-ID: fj6MEAAAQBAJ. 978 pp. ISBN: 978-1-119-59017-0 (John Wiley & Sons, 2019).

159. Nó, L. d. *Studies on the structure of the cerebral cortex* Pages: 381 Publication Title: J. Psychol. Neurol. Volume: 45. 1933.

160. Oldenburg, I. A. *et al.* The logic of recurrent circuits in the primary visual cortex. *Nature Neuroscience* **27.** Publisher: Nature Publishing Group, 137–147. ISSN: 1546-1726 (Jan. 2024).

161. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Current Opinion in Neurobiology* **14,** 481–487. ISSN: 0959-4388 (Aug. 1, 2004).

162. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381.** Publisher: Nature Publishing Group, 607–609. ISSN: 1476-4687 (June 1996).

163. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron* **92.** Publisher: Elsevier, 530–543. ISSN: 0896-6273 (Oct. 19, 2016).

164. Otchy, T. M. *et al.* Acute off-target effects of neural circuit manipulations. *Nature* **528,** 358–363. ISSN: 1476-4687 (Dec. 2015).

165. Pachitariu, M., Stringer, C., Dipoppa, M., *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv* (July 20, 2017).

166. Pachitariu, M., Stringer, C. & Harris, K. D. Robustness of Spike Deconvolution for Neuronal Calcium Imaging. *Journal of Neuroscience* **38,** 7976–7985. ISSN: 0270-6474, 1529-2401 (Sept. 12, 2018).

167. Packer, A. M., Russell, L. E., Dalgleish, H. W. P. & Häusser, M. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nature Methods* **12.** Number: 2 Publisher: Nature Publishing Group, 140–146. ISSN: 1548-7105 (Feb. 2015).

168. Palmigiano, A. *et al. Common rules underlying optogenetic and behavioral modulation of responses in multi-cell-type V1 circuits* Jan. 22, 2023.

169. Palmigiano, A. *et al.* Structure and variability of optogenetic responses identify the operating regime of cortex. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.11.11.378729 (Nov. 11, 2020).

170. Petrov, A. A., Dosher, B. A. & Lu, Z.-L. The Dynamics of Perceptual Learning: An Incremental Reweighting Model. *Psychological Review* **112.** Place: US Publisher: American Psychological Association, 715–743. ISSN: 1939-1471 (2005).

171. Pogodin, R., Mehta, Y., Lillicrap, T. & Latham, P. E. *Towards Biologically Plausible Convolutional Networks* in *Advances in Neural Information Processing Systems* **34** (Curran Associates, Inc., 2021), 13924–13936.

172. Popper, K. R. Science as falsification. *Conjectures and refutations* **1.1963,** 33–39 (1963).

173. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nature Neuroscience* **16.** Publisher: Nature Publishing Group, 1170–1178. ISSN: 1546-1726 (Sept. 2013).

174. Pozzi, I., Bohté, S. & Roelfsema, P. *A Biologically Plausible Learning Rule for Deep Learning in the Brain* July 2, 2019. arXiv: 1811.01768[cs].

175. Prevedel, R. *et al.* Three-photon microscopy: an emerging technique for deep intravital brain imaging. *Nature Reviews Neuroscience.* Publisher: Nature Publishing Group, 1–17. ISSN: 1471-0048 (June 20, 2025).

176. Raiguel, S., Vogels, R., Mysore, S. G. & Orban, G. A. Learning to See the Difference Specifically Alters the Most Informative V4 Neurons. *Journal of Neuroscience* **26.** Publisher: Society for Neuroscience Section: Articles, 6589–6602. ISSN: 0270-6474, 1529-2401 (June 14, 2006).

177. Randlett, O. *et al.* Distributed Plasticity Drives Visual Habituation Learning in Larval Zebrafish. *Current Biology* **29.** Publisher: Elsevier, 1337–1345.e4. ISSN: 0960-9822 (Apr. 22, 2019).

178. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2.** Number: 1 Publisher: Nature Publishing Group, 79–87. ISSN: 1546-1726 (Jan. 1999).

179. Real, E., Asari, H., Gollisch, T. & Meister, M. Neural Circuit Inference from Function to Structure. *Current Biology* **27.** Publisher: Elsevier, 189–198. ISSN: 0960-9822 (Jan. 23, 2017).

180. Recanzone, G. H., Schreiner, C. E. & Merzenich, M. M. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *Journal of Neuroscience* **13.** Publisher: Society for Neuroscience Section: Articles, 87–103. ISSN: 0270-6474, 1529-2401 (Jan. 1, 1993).

181. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nature Neuroscience* **22.** Publisher: Nature Publishing Group, 1761–1770. ISSN: 1546-1726 (Nov. 2019).

182. Rieke, F., Bodnar, D. A. & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **262.** Publisher: Royal Society, 259–265 (Jan. 1997).

183. Riesenhuber, M. & Poggio, T. Neural mechanisms of object recognition. *Current Opinion in Neurobiology* **12,** 162–168. ISSN: 0959-4388 (Apr. 1, 2002).

184. Roberts, S. & Pashler, H. How persuasive is a good fit? A comment on theory testing. *Psychological Review* **107.** Place: US Publisher: American Psychological Association, 358–367. ISSN: 1939-1471 (2000).

185. Rose, D. & Blakemore, C. An analysis of orientation selectivity in the cat's visual cortex. *Experimental Brain Research* **20,** 1–17. ISSN: 1432-1106 (Apr. 1, 1974).

186. Rossi, L. F., Harris, K. D. & Carandini, M. Spatial connectivity matches direction selectivity in visual cortex. *Nature* **588,** 648–652. ISSN: 1476-4687 (Dec. 2020).

187. Rowland, J. M. *et al.* Propagation of activity through the cortical hierarchy and perception are determined by neural variability. *Nature Neuroscience* **26.** Publisher: Nature Publishing Group, 1584–1594. ISSN: 1546-1726 (Sept. 2023).

188. Rubin, D. B., Van Hooser, S. D. & Miller, K. D. The Stabilized Supralinear Network: A Unifying Circuit Motif Underlying Multi-Input Integration in Sensory Cortex. *Neuron* **85,** 402–417. ISSN: 0896-6273 (Jan. 21, 2015).

189. Ruderman, D. L. The statistics of natural images. *Network: Computation in Neural Systems* **5.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1088/0954-898X_5_4_006, 517–548. ISSN: 0954-898X (Jan. 1, 1994).

190. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323.** Publisher: Nature Publishing Group, 533–536. ISSN: 1476-4687 (Oct. 1986).

191. Rumelhart, D. E., McClelland, J. L. & Group, P. R. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations* ISBN: 978-0-262-29140-8 (The MIT Press, July 17, 1986).

192. Russell, L. E. *et al.* The influence of cortical activity on perception depends on behavioral state and sensory context. *Nature Communications* **15.** Publisher: Nature Publishing Group, 2456. ISSN: 2041-1723 (Mar. 19, 2024).

193. Saalmann, Y. B., Pigarev, I. N. & Vidyasagar, T. R. Neural Mechanisms of Visual Attention: How Top-Down Feedback Highlights Relevant Locations. *Science* **316.** Publisher: American Association for the Advancement of Science, 1612–1615 (June 15, 2007).

194. Sadeh, S. & Clopath, C. Patterned perturbation of inhibition can reveal the dynamical structure of neural processing. *eLife* **9** (eds Palmer, S. & Frank, M. J.) Publisher: eLife Sciences Publications, Ltd, e52757. ISSN: 2050-084X (Feb. 19, 2020).

195. Saleem, A. B. Two stream hypothesis of visual processing for navigation in mouse. *Current Opinion in Neurobiology. Systems Neuroscience* **64,** 70–78. ISSN: 0959-4388 (Oct. 1, 2020).

196. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience* **22.** Publisher: Nature Publishing Group, 55–67. ISSN: 1471-0048 (Jan. 2021).

197. Saxe, A. M., McClelland, J. L. & Ganguli, S. *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks* Feb. 19, 2014. arXiv: 1312.6120[cs].

198. Saxe, A. M. *Deep Linear Neural Networks: A Theory of Learning in the Brain and Mind* ISBN: 9798662564862. PhD thesis (Stanford University, United States – California, 2015). 207 pp.

199. Scala, F. *et al.* Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas. *Nature Communications* **10.** Publisher: Nature Publishing Group, 4174. ISSN: 2041-1723 (Sept. 13, 2019).

200. Schapiro, A. C., Turk-Browne, N. B., Norman, K. A. & Botvinick, M. M. Statistical learning of temporal community structure in the hippocampus. *Hippocampus* **26.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hipo.22523, 3–8. ISSN: 1098-1063 (2016).

201. Schoups, A. A., Vogels, R. & Orban, G. A. Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity. *The Journal of Physiology* **483.** _eprint: https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.1995.sp020623, 797–810. ISSN: 1469-7793 (1995).

202. Schoups, A., Vogels, R., Qian, N. & Orban, G. Practising orientation identification improves orientation coding in V1 neurons. *Nature* **412.** Number: 6846 Publisher: Nature Publishing Group, 549–553. ISSN: 1476-4687 (Aug. 2001).

203. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275.** Publisher: American Association for the Advancement of Science, 1593–1599 (Mar. 14, 1997).

204. Seitz, A. R. Perceptual learning. *Current Biology* **27.** Publisher: Elsevier, R631–R636. ISSN: 0960-9822 (July 10, 2017).

205. Sengpiel, F. & Bonhoeffer, T. Orientation specificity of contrast adaptation in visual cortical pinwheel centres and iso-orientation domains. *European Journal of Neuroscience* **15,** 876–886. ISSN: 1460-9568 (2002).

206. Shao, F. & Shen, Z. How can artificial neural networks approximate the brain? *Frontiers in Psychology* **13.** Publisher: Frontiers. ISSN: 1664-1078 (Jan. 9, 2023).

207. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience* **36.** Publisher: Annual Reviews, 337–359. ISSN: 0147-006X, 1545-4126 (Volume 36, 2013 July 8, 2013).

208. Shine, J. M. *et al.* Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nature Neuroscience* **22.** Publisher: Nature Publishing Group, 289–296. ISSN: 1546-1726 (Feb. 2019).

209. Simoncelli, E. P. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology* **13,** 144–149. ISSN: 0959-4388 (Apr. 1, 2003).

210. Simoncelli, E. P. & Olshausen, B. A. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience* **24.** Publisher: Annual Reviews, 1193–1216. ISSN: 0147-006X, 1545-4126 (Volume 24, 2001 Mar. 1, 2001).

211. Sorzano, C. O. S., Vargas, J. & Montano, A. P. *A survey of dimensionality reduction techniques* Mar. 12, 2014. arXiv: 1403.2877[stat].

212. Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* **87,** 245–251. ISSN: 1939-1455 (1980).

213. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576.** Publisher: Nature Publishing Group, 266–273. ISSN: 1476-4687 (Dec. 2019).

214. Stiles, J. in *Progress in Brain Research* (eds Braddick, O., Atkinson, J. & Innocenti, G. M.) 3–22 (Elsevier, Jan. 1, 2011).

215. Stringer, C., Michaelos, M., Tsyboulski, D., Lindo, S. E. & Pachitariu, M. High-precision coding in visual cortex. *Cell* **184.** Publisher: Elsevier, 2767–2778.e15. ISSN: 0092-8674, 1097-4172 (May 13, 2021).

216. Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* 3rd ed. 616 pp. ISBN: 978-0-429-39849-0 (Chapman and Hall/CRC, New York, Jan. 16, 2024).

217. Suárez, L. E., Richards, B. A., Lajoie, G. & Misic, B. Learning function from structure in neuromorphic networks. *Nature Machine Intelligence* **3.** Publisher: Nature Publishing Group, 771–786. ISSN: 2522-5839 (Sept. 2021).

218. Sussillo, D. Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology. Theoretical and computational neuroscience* **25,** 156–163. ISSN: 0959-4388 (Apr. 1, 2014).

219. Suvrathan, A. Beyond STDP — towards diverse and functionally relevant plasticity rules. *Current Opinion in Neurobiology. Neurobiology of Learning and Plasticity* **54,** 12–19. ISSN: 0959-4388 (Feb. 1, 2019).

220. Treves, A., Panzeri, S., Rolls, E. T., Booth, M. & Wakeman, E. A. Firing Rate Distributions and Efficiency of Information Transmission of Inferior Temporal Cortex Neurons to Natural Visual Stimuli. *Neural Computation* **11,** 601–631. ISSN: 0899-7667 (Apr. 1, 1999).

221. Van Gelder, T. What Might Cognition Be, If Not Computation? *The Journal of Philosophy* **92.** Publisher: Journal of Philosophy, Inc., 345–381. ISSN: 0022-362X (1995).

222. Van Hooser, S. D. Similarity and Diversity in Visual Cortex: Is There a Unifying Theory of Cortical Computation? *The Neuroscientist* **13,** 639–656. ISSN: 1073-8584 (Dec. 1, 2007).

223. Vaswani, A. *et al. Attention is All you Need* in *Advances in Neural Information Processing Systems* **30** (Curran Associates, Inc., 2017).

224. Verma, K. & Kumar, S. How Do We Connect Brain Areas with Cognitive Functions? The Past, the Present and the Future. *NeuroSci* **3.** Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, 521–532. ISSN: 2673-4087 (Sept. 2022).

225. Vinje, W. E. & Gallant, J. L. Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science* **287.** Publisher: American Association for the Advancement of Science, 1273–1276 (Feb. 18, 2000).

226. Vogels, R. Mechanisms of Visual Perceptual Learning in Macaque Visual Cortex. *Topics in Cognitive Science* **2,** 239–250. ISSN: 1756-8765 (2010).

227. Vogels, R. & Orban, G. A. The effect of practice on the oblique effect in line orientation judgments. *Vision Research* **25,** 1679–1687. ISSN: 0042-6989 (Jan. 1, 1985).

228. Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience* **2018.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2018/7068349, 7068349. ISSN: 1687-5273 (2018).

229. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation Through Neural Population Dynamics. *Annual Review of Neuroscience* **43,** 249–275. ISSN: 0147-006X, 1545-4126 (July 8, 2020).

230. Wang, Q. & Burkhalter, A. Area map of mouse visual cortex. *Journal of Comparative Neurology* **502,** 339–357. ISSN: 1096-9861 (2007).

231. Wang, Q., Gao, E. & Burkhalter, A. Gateways of Ventral and Dorsal Streams in Mouse Visual Cortex. *Journal of Neuroscience* **31,** 1905–1918. ISSN: 0270-6474, 1529-2401 (Feb. 2, 2011).

232. Watanabe, T., Náñez, J. E. & Sasaki, Y. Perceptual learning without perception. *Nature* **413.** Publisher: Nature Publishing Group, 844–848. ISSN: 1476-4687 (Oct. 2001).

233. Watson, B. O., Yuste, R. & Packer, A. M. *PackIO and EphysViewer: software tools for acquisition and analysis of neuroscience data* Pages: 054080 Section: New Results. May 18, 2016.

234. Wekselblatt, J. B., Flister, E. D., Piscopo, D. M. & Niell, C. M. Large-scale imaging of cortical dynamics during sensory perception and behavior. *APSselect* **3.** Publisher: American Physiological Society, 2852–2866 (Apr. 2016).

235. Wenliang, L. K. & Seitz, A. R. Deep Neural Networks for Modeling Visual Perceptual Learning. *Journal of Neuroscience* **38.** Publisher: Society for Neuroscience Section: Research Articles, 6028–6044. ISSN: 0270-6474, 1529-2401 (July 4, 2018).

236. Whitlock, J. R., Heynen, A. J., Shuler, M. G. & Bear, M. F. Learning Induces Long-Term Potentiation in the Hippocampus. *Science* **313.** Publisher: American Association for the Advancement of Science, 1093–1097 (Aug. 25, 2006).

237. Whittington, J. C. R. & Bogacz, R. Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences* **23,** 235–250. ISSN: 1364-6613 (Mar. 1, 2019).

238. Wiesel, T. N. & Hubel, D. H. Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Neurophysiology* **26.** Publisher: American Physiological Society, 1003–1017. ISSN: 0022-3077 (Nov. 1963).

239. Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An Internal Model for Sensorimotor Integration. *Science* **269.** Publisher: American Association for the Advancement of Science, 1880–1882 (Sept. 29, 1995).

240. Worden, F. G. & Marsh, J. T. Amplitude changes of auditory potentials evoked at cochlear nucleus during acoustic habituation. *Electroencephalography and Clinical Neurophysiology* **15,** 866–881. ISSN: 0013-4694 (Oct. 1, 1963).

241. Xing, Y., Zan, C. & Liu, L. Recent advances in understanding neuronal diversity and neural circuit complexity across different brain regions using single-cell sequencing. *Frontiers in Neural Circuits* **17.** Publisher: Frontiers. ISSN: 1662-5110 (Mar. 30, 2023).

242. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111.** Publisher: Proceedings of the National Academy of Sciences, 8619–8624 (June 10, 2014).

243. Yan, Y. *et al.* Perceptual training continuously refines neuronal population codes in primary visual cortex. *Nature Neuroscience* **17.** Publisher: Nature Publishing Group, 1380–1387. ISSN: 1546-1726 (Oct. 2014).

244. Yang, G., Pan, F. & Gan, W.-B. Stably maintained dendritic spines are associated with lifelong memories. *Nature* **462.** Publisher: Nature Publishing Group, 920–924. ISSN: 1476-4687 (Dec. 2009).

245. Yang, T. & Maunsell, J. H. R. The Effect of Perceptual Learning on Neuronal Responses in Monkey Visual Area V4. *Journal of Neuroscience* **24.** Publisher: Society for Neuroscience Section: Behavioral/Systems/Cognitive, 1617–1626. ISSN: 0270-6474, 1529-2401 (Feb. 18, 2004).

246. Zhang, L. I. & Poo, M.-m. Electrical activity and development of neural circuits. *Nature Neuroscience* **4.** Publisher: Nature Publishing Group, 1207–1214. ISSN: 1546-1726 (Nov. 2001).

247. Zhu, Z. & Wakin, M. B. On the Asymptotic Equivalence of Circulant and Toeplitz Matrices. *arXiv:1608.04820 [cs, math].* arXiv: 1608.04820 (Feb. 22, 2017).

248. Zhuang, J. *et al.* An extended retinotopic map of mouse cortex. *eLife* **6** (ed Kleinfeld, D.) Publisher: eLife Sciences Publications, Ltd, e18372. ISSN: 2050-084X (Jan. 6, 2017).

249. Znamenskiy, P. *et al.* Functional specificity of recurrent inhibition in visual cortex. *Neuron* **112.** Publisher: Elsevier, 991–1000 (2024).