

2022

# Unsupervised space-time learning in primary visual cortex

---

<https://hdl.handle.net/2144/45515>

*"Downloaded from OpenBU. Boston University's institutional repository."*

BOSTON UNIVERSITY  
SCHOOL OF MEDICINE

Dissertation

**UNSUPERVISED SPACE-TIME LEARNING IN PRIMARY VISUAL CORTEX**

by

**BYRON HOWARD PRICE**

B.A., Vanderbilt University, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2022

© 2022 by  
BYRON HOWARD PRICE  
All rights reserved

Approved by

First Reader

---

Jeffrey Gavornik, Ph.D.  
Assistant Professor of Biology

Second Reader

---

Arash Yazdanbakhsh, M.D., Ph.D.  
Research Assistant Professor

Third Reader

---

David Somers, Ph.D.  
Professor and Chair, Department of Psychological and Brain Sciences

Alice sighed wearily. 'I think you might do something better with the time,' she said, 'than waste it in asking riddles that have no answers.'

'If you knew Time as well as I do,' said the Hatter, 'you wouldn't talk about wasting *it*. It's *him*.'

'I don't know what you mean,' said Alice.

'Of course you don't!' the Hatter said, tossing his head contemptuously. 'I dare say you never even spoke to Time!'

'Perhaps not,' Alice cautiously replied: 'but I know I have to beat time when I learn music.'

'Ah! That accounts for it,' said the Hatter. 'He won't stand beating.'

- Lewis Carroll

## **DEDICATION**

I dedicate this work to my dad, who through long hours at the kitchen table, taught me a healthy skepticism of authority and a hardy respect for self-reliance.

## **ACKNOWLEDGMENTS**

I would like to thank my committee members, especially Jeff Gavornik, for guiding me along this journey to scientific independence. In addition, I thank my friends and family for their enduring support. Finally, much of the original research in this dissertation was completed as a team, with significant support from Cambria Jensen, Anthony Khoudary, and Scott Knudstrup.

# **UNSUPERVISED SPACE-TIME LEARNING IN PRIMARY VISUAL CORTEX**

**BYRON HOWARD PRICE**

Boston University School of Medicine, 2022

Major Professor: Jeffrey Gavornik, Ph.D., Assistant Professor of Biology

## **ABSTRACT**

The mammalian visual system is an incredibly complex computation device, capable of performing the various tasks of seeing: navigation, pattern and object recognition, motor coordination, trajectory extrapolation, among others. Decades of research has shown that experience-dependent plasticity of cortical circuitry underlies the impressive ability to rapidly learn many of these tasks and to adjust as required. One particular thread of investigation has focused on unsupervised learning, wherein changes to the visual environment lead to corresponding changes in cortical circuits. The most prominent example of unsupervised learning is ocular dominance plasticity, caused by visual deprivation to one eye and leading to a dramatic re-wiring of cortex. Other examples tend to make more subtle changes to the visual environment through passive exposure to novel visual stimuli. Here, we use one such unsupervised paradigm, sequence learning, to study experience-dependent plasticity in the mouse visual system. Through a combination of theory and experiment, we argue that the mammalian visual system is an unsupervised learning device.

Beginning with a mathematical exploration of unsupervised learning in biology, engineering, and machine learning, we seek a more precise expression of our fundamental

hypothesis. We draw connections between information theory, efficient coding, and common unsupervised learning algorithms such as Hebbian plasticity and principal component analysis. Efficient coding suggests a simple rule for transmitting information in the nervous system: use more spikes to encode unexpected information, and fewer spikes to encode expected information. Therefore, expectation violations ought to produce *prediction errors*, or brief periods of heightened firing when an unexpected event occurs. Meanwhile, modern unsupervised learning algorithms show how such expectations can be learned.

Next, we review data from decades of visual neuroscience research, highlighting the computational principles and synaptic plasticity processes that support biological learning and seeing. By tracking the flow of visual information from the retina to thalamus and primary visual cortex, we discuss how the principle of efficient coding is evident in neural activity. One common example is predictive coding in the retina, where ganglion cells with canonical center-surround receptive fields compute a prediction error, sending spikes to the central nervous system only in response to locally-unpredictable visual stimuli. This behavior can be learned through simple Hebbian plasticity mechanisms. Similar models explain much of the activity of neurons in primary visual cortex, but we also discuss ways in which the theory fails to capture the rich biological complexity.

Finally, we present novel experimental results from physiological investigations of the mouse primary visual cortex. We trained mice by passively exposing them to complex spatiotemporal patterns of light: rapidly-flashed sequences of images. We find evidence that visual cortex learns these sequences in a manner consistent with efficient coding, such

that unexpected stimuli tend to elicit more firing than expected ones. Overall, we observe dramatic changes in evoked neural activity across days of passive exposure. Neural responses to the first, unexpected sequence element increase with days of training while responses at other, expected time points either decrease or stay the same. Furthermore, substituting an unexpected element for an expected one or omitting an expected element both cause brief bursts of increased firing. Our results therefore provide evidence for unsupervised learning and efficient coding in the mouse visual system, especially because unexpected events drive prediction errors. Overall, our analysis suggests novel experiments, which could be performed in the near future, and provides a useful framework to understand visual perception and learning.

## PREFACE

The contents of this dissertation concern vision and the problem of perception more generally. A central hypothesis in the study of vision, dating back to the time of Hermann von Helmholtz, is the notion of perception as inference (H. Barlow, 2001b; H. B. Barlow, 1961; Friston, 2005; Peddie, 1925). To see a chair and recognize it as such is to perform a kind of statistical inference, discerning the chair from amongst a zoo of potential objects and doing so in the presence of considerable uncertainty. In order to perform such inferences, canonical models of the visual system posit a hierarchical computation (Carandini et al., 2005; David H Hubel & Wiesel, 1962; Olshausen & Field, 1996; Riesenhuber & Poggio, 1999; T N Wiesel, 1968). Visual neurons are feature detectors, signaling the presence of particular visual features, such as edges or textures, with neurons in regions further up the hierarchy representing more abstract information. Significant experimental evidence has helped to shape this hypothesis. For example, neurons in later visual cortical areas, such as inferior temporal cortex, respond more sparsely to visual inputs in broader regions of visual space than do neurons in earlier visual areas such as primary visual cortex (V1) (Rust & Dicarlo, 2010; Siegle et al., 2021).

In more recent years, the same general framework has been extended to what might be called the *deep learning theory of vision* (Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021; Cadena et al., 2019; Conwell, Buice, Alvarez, Katz, & Barbu, 2021; Kriegeskorte, 2015; Lindsay, 2021; Lindsey, Ocko, Ganguli, & Deny, 2019; Richards et al., 2019; Yamins & Dicarlo, 2016; Yamins et al., 2014; Zhuang et al., 2021). According to this

theory, visual information processing in the brain is comparable to the hierarchical processing of (deep) artificial neural networks. Convolutional neural networks trained to perform object recognition, for example, achieve state-of-the-art performance in their ability to predict visual cortical activity (Cadena et al., 2019; Yamins et al., 2014), far surpassing traditional models based on the notions of receptive fields or motion energy (Adelson & Bergen, 1985; Hughes, Schwartz, Pillow, Rust, & Simoncelli, 2006; M. Park & Pillow, 2011; Theunissen et al., 2001). In addition to this supervised learning approach, deep networks trained through unsupervised learning yield comparable state-of-the-art performance (Bakhtiari et al., 2021; Higgins et al., 2020; Zhuang et al., 2021). As we will explore in detail, while supervised learning attempts to discover a functional mapping between input-output data pairs, unsupervised learning seeks to learn the probability distribution of the data (Hinton & Sejnowski, 1999; Oja, 2002). A common corollary to the deep learning theory of vision is therefore: the visual system is a learning device that discovers the distribution of its inputs (DiCarlo, Zoccolan, & Rust, 2012).

The inputs to the visual system are extended in both space and time, and they are multi-modal, coming from many areas across the brain including auditory and motor cortices (Garner & Keller, 2021; Guitchounts, Masís, Wolff, & Cox, 2020; Leinweber, Ward, Sobczak, Attinger, & Keller, 2017; Poort et al., 2015; Stringer, Pachitariu, Steinmetz, Reddy, et al., 2019). It is generally accepted that the primate visual system processes information in two distinct systems: what and where/how, or the ventral and dorsal streams (Goodale & Milner, 1992; Milner & Goodale, 2008; Ungerleider & Mishkin, 1982). Object

recognition information is processed by the ventral stream, while movement information and visually-guided behaviors are processed in the dorsal stream. A similar dichotomy likely applies to the mouse visual system (Bakhtiari et al., 2021; Garrett, Nauhaus, Marshel, & Callaway, 2014; Marshel, Garrett, Nauhaus, & Callaway, 2011; Siegle et al., 2021). Due to this segregation of space-like information in the ventral stream and time-like information in the dorsal stream, many models of ventral stream object recognition are based on static images (Bell & Sejnowski, 1997; Cadena et al., 2019; Carandini et al., 2005; Olshausen & Field, 1996; Yamins et al., 2014; Zhuang et al., 2021). The temporal component of the visual input is merely believed to allow for the integration of noisy sensory data, in order to accumulate evidence. However, given the prevalence of saccadic eye movements (Kuang, Poletti, Victor, & Rucci, 2012; Rucci, 2008; Rucci & Victor, 2015) and the ubiquity of motion within the natural environment (both through visual flow and external movement), temporal information may also be relevant to ventral stream computations. We thus arrive at our central hypothesis: *the visual system is an unsupervised learning device that discovers the distribution of its inputs, which are functions of both space and time*. In essence, time matters for seeing. We will explore this idea in more detail throughout and attempt to make it more mathematically precise.

To explore our hypothesis, the dissertation is divided into four sections: 1) Unsupervised Learning; 2) Vision; 3) Expectation Violations Produce Error Signals in Mouse V1; and 4) Conclusions and Future Directions. The first section provides an overview of unsupervised learning, with a focus on models that are relevant to the visual system, in particular

predictive coders and autoencoders. The second section is a review of experimental data from the visual system. We discuss Barlow's efficient coding hypothesis (H. B. Barlow, 1961), predictive coding (Elias, 1955; Srinivasan, Laughlin, & Dubs, 1982), and data supporting both the hierarchical feature extraction and deep learning theories of vision. We also provide relevant background to understand the third section, including sequence learning and other experience-dependent plasticity paradigms (Dudek & Bear, 1992; Frenkel et al., 2006; Gavornik & Bear, 2014). The third section contains novel results from *in vivo* experiments performed in our lab. These experiments broadly conform to the central hypothesis, as we find evidence that V1 learns pertinent features of spatiotemporal sequences that are out-of-distribution relative to the natural environment. The final section discusses experiments that might be performed to further investigate our hypothesis and to better understand sequence learning.

## TABLE OF CONTENTS

DEDICATION.....	v
ACKNOWLEDGMENTS.....	vi
ABSTRACT.....	vii
PREFACE.....	x
TABLE OF CONTENTS.....	xiv
LIST OF TABLES.....	xvii
LIST OF FIGURES.....	xviii
LIST OF ABBREVIATIONS.....	xix
UNSUPERVISED LEARNING.....	1
Hebbian Learning.....	3
Predictive Coders.....	11
A Predictive Coder with Local Anti-Hebbian Plasticity.....	16
PCA.....	20
A Local Implementation of PCA.....	24
Deep Unsupervised Learning.....	28
Variational Autoencoders.....	29
Self-Supervised Learning.....	31
VISION.....	35
Retina.....	36
Structure.....	36
Function.....	38

Thalamus.....	44
Structure.....	44
Function.....	46
Primary Visual Cortex.....	48
Structure.....	48
Function.....	52
Visual Cortical Plasticity.....	62
Ocular Dominance Plasticity & Monocular Deprivation.....	63
Sequence Learning.....	68
EXPECTATION VIOLATIONS PRODUCE ERROR SIGNALS IN MOUSE V1.....	78
Introduction.....	78
Results.....	80
Experimental Design and Sequence Stimulus.....	80
A Statistical Model (MbTDR) Captures Stimulus-Dependent Neural Variability...	82
MbTDR Reveals Coordinated Training-Dependent Changes in Neural Activity....	87
Orientation Tuning Does Not Shift Significantly with Training.....	89
Unexpected Omissions Cause Negative Prediction Errors.....	90
Unexpected Substitutions Cause Positive Prediction Errors.....	94
Reliable Temporal Information is Contained in the Neural Code.....	98
Discussion.....	101
Methods.....	107
Supplemental Information and Figures.....	127

CONCLUSIONS.....	134
Infinite Time, Infinite Resources.....	137
Finite Time, Finite Resources.....	141
Fin.....	143
APPENDIX.....	144
Probability and Information Theory.....	144
Laws of Probability.....	144
Random Variables.....	148
Expectation.....	149
Statistics.....	150
Statistical Inference.....	151
Maximum Likelihood Estimation.....	152
Information Entropy.....	153
Mutual Information.....	155
Efficient Coding.....	156
BIBLIOGRAPHY.....	163
CURRICULUM VITAE.....	193

## LIST OF TABLES

Supplemental Table 1: MbTDR Information.....	132
Supplemental Table 2: Summary of Statistical Tests .....	133

## LIST OF FIGURES

Figure 1: Deep Neural Network Spatial Predictive Coder.....	15
Figure 2: Local Predictive Coder Network.....	18
Figure 3: Visual Representation of PCA .....	22
Figure 4: Sequence Stimulus Representation in a Model of dLGN and V1 .....	72
Figure 5: Experimental Design .....	81
Figure 6: MbTDR Captures Stimulus-Dependent Neural Variability .....	84
Figure 7: MbTDR Reveals Coordinated Training-Dependent Change in Neural Activity .....	88
Figure 8: Unexpected Omissions Cause Negative Prediction Errors .....	91
Figure 9: Unexpected Substitutions Cause Positive Prediction Errors .....	96
Figure 10: Predictions Span Element Transitions.....	98
Figure 11: Temporal Information in the V1 Neural Code .....	100
Supplemental Figure 1: Post-Mortem Histology .....	127
Supplemental Figure 2: Absence of Stimulus-Aligned Movement .....	128
Supplemental Figure 3: Orientation Tuning of the Second Sequence Element.....	129
Supplemental Figure 4: Orientation Tuning of the Negative Prediction Error.....	130
Supplemental Figure 5: MbTDR Bases .....	131

## LIST OF ABBREVIATIONS

AMPARs	$\alpha$ -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors
BCM	Bienenstock, Cooper, Munro theory of synaptic plasticity
BDNF	Brain-derived neurotrophic factor
CNN	Convolutional neural network
CPP	3-(2-carboxypiperazin-4-yl)propyl-1-phosphonic acid (NMDAR antagonist)
dLGN	Dorsal lateral geniculate nucleus of the thalamus
GABA	gamma-Aminobutyric acid
IT	Inferior temporal cortex
LFP	Local field potential
LTD	Long-term depression
LTP	Long-term potentiation
mAChRs	Muscarinic acetylcholine receptors
NMDARs	N-methyl-D-aspartate receptors
ODP	Ocular dominance plasticity
PCA	Principal component analysis
PV	Parvalbumin expressing interneuron
RGC	Retinal ganglion cell
SOM	Somatostatin expressing interneuron
SRP	Stimulus-selective response potentiation
STDP	Spike-timing dependent plasticity

TRN	Thalamic reticular nucleus
V1	Primary visual cortex
VAE	Variational autoencoder
VEP	Visually-evoked potential (as measured in the local field potential)
VIP	Vasoactive intestinal peptide expressing interneuron

## UNSUPERVISED LEARNING

Unsupervised learning is a statistical procedure, differentiated from supervised learning and reinforcement learning. In supervised learning, data come in input-output pairs and the goal is to learn a functional mapping from input to output. The prototypical example of this is object recognition and the well-known ImageNet dataset (Deng et al., 2009). ImageNet contains thousands of images of common objects, such as dogs and cars, along with labels that identify the objects. Supervised learning algorithms discover a functional mapping from the images to the labels, attempting to accurately predict the label given the image. In reinforcement learning, an agent, broadly construed, navigates an environment, receiving rewards and punishments along the way (Collins & Cockburn, 2020). The goal is for the agent to maximize its long-run reward and minimize its long-run punishment.

By comparison, unsupervised learning is somewhat more nebulous (Hinton & Sejnowski, 1999; Oja, 2002). It is often referred to as learning without a teacher. In the example of ImageNet, unsupervised algorithms might attempt to cluster images into groupings of related objects, but without knowing the labels. Or, they could try to find a compact low-dimensional representation of the images, perhaps for data-efficient storage. They might aim to learn a *generative* model, so that novel images (deep fakes) can be synthesized. Though seemingly disparate tasks, the connective thread is that unsupervised learning algorithms attempt to learn the probability distribution of the data.

In recent years, self-supervised learning has gained popularity, though it falls under the umbrella of unsupervised learning. Self-supervised learning uses unlabeled data, like images, in creative ways. For example, in natural language processing, self-supervised algorithms remove words or letters from sentences, attempting to reconstruct the original text. Modern algorithms like Barlow Twins (Zbontar, Jing, Misra, LeCun, & Deny, 2021) and SimCLR (T. Chen, Kornblith, Norouzi, & Hinton, 2020) use data distortions, such as rotations and blurring, seeking to learn a low-dimensional representation that is invariant to such changes. It has recently been proved that a class of *contrastive* self-supervised models explicitly learns the distribution of a generative latent-variable model (Oord, Li, & Vinyals, 2018; Zimmermann, Sharma, Schneider, Bethge, & Brendel, 2021).

These self-supervised approaches are relevant because they are very good models of object recognition and the mammalian visual system. On the standard ImageNet object recognition task, self-supervised approaches now match the most competitive supervised algorithms (Tomasev et al., 2022). With self-supervised learning for object recognition, a deep neural network model is trained to learn a low-dimensional representation of the distribution of the image data, with no labels. Once training is complete, a straightforward linear classification is performed from the learned low-dimensional representation to the labels. That this approach matches supervised algorithms is very impressive for two reasons: 1) linear classification done on the raw images, or on the raw images after principal component analysis (PCA), performs terribly; and 2) self-supervised algorithms have no access to labels and no explicit goal of learning about objects, but they seem to learn very

abstract features of the data. As a model of the visual system, researchers have taken a similar approach (Bakhtiari et al., 2021; Higgins et al., 2020; Zhuang et al., 2021). They first train a self-supervised neural network on a dataset of images or videos, and then use the low-dimensional representation generated by the model, along with linear regression, to predict the activity of visual neurons. This self-supervised approach achieves similar performance to deep neural networks trained through supervised learning to directly predict neural activity from images or videos.

In this section, we provide an accessible introduction to unsupervised learning, progressively building to more complicated models. The purpose is to establish a mathematical foundation to more precisely define our hypothesis, that the visual system is an unsupervised learning device. The models described here will allow us to imagine experiments that could test our hypothesis and will establish a useful framework for understanding the subsequent section on Vision.

### **Hebbian Learning**

Biological learning, whether supervised or unsupervised, is accompanied by lasting changes to the nervous system (Gerstner & Kistler, 2002; Hebb, 1949; Hennequin, Agnes, & Vogels, 2017; H. Markram, Gerstner, & Sjöström, 2012). Such changes include modifications to neuronal connectivity, to the weight or strength of synaptic connections, to the intrinsic electrical properties of neurons, and likely to extra-neuronal properties of the local environment such as the distribution of glia and blood vessels. The canonical model of unsupervised learning focuses only on changes to synaptic weights, as originally

proposed by Donald Hebb (Gerstner & Kistler, 2002; Hebb, 1949). In this view, the strength of a synapse,  $w$ , changes as some function of pre- and post-synaptic firing rates ( $x(t)$  and  $z(t)$ , respectively):

$$\frac{dw}{dt} = \eta f[x(t), z(t)]$$

$\eta$  is a small positive value representing the learning rate. The traditional Hebbian learning rule uses a simple linear form for the function  $f$ :

$$f[x(t), z(t)] = x(t)z(t)$$

Thus, when the pre-synaptic and post-synaptic neurons show correlated firing,  $x(t)z(t)$  tends to be positive and the synaptic weight increases. Donald Hebb described this as the pre-synaptic neuron increasing its efficiency in causing the post-synaptic neuron to fire (Hebb, 1949). Modern accounts often express this as an implementation of associative learning (Dayan & Abbott, 2005).

There is also an Anti-Hebbian learning rule that makes a minor modification to  $f$ :

$$f[x(t), z(t)] = -x(t)z(t)$$

Now, anti-correlated activity causes the weight to increase. As we will outline below, this rule implements predictive coding when instantiated in a specific feedforward inhibitory network architecture (Hosoya, Baccus, & Meister, 2005).

A wide variety of learning rules fit this general framework, where synaptic weight modifications are functions of pre- and post-synaptic neural activity, for example, neural implementations of principal component analysis (PCA) (H. B. Barlow & Földiák, 1989;

Falconbridge, Stamps, & Badcock, 2006; Oja, 1982; Pehlevan & Chklovskii, 2019; Sanger, 1989). These tend to require a more general consideration of  $f$ , such that weight modifications are functions of additional parameters:

$$\frac{dw}{dt} = \eta f[x(t), z(t), w, \theta(t)]$$

Now, the weight changes as a function of the weight itself, and an additional time-varying parameter,  $\theta$ . The only restriction on  $\theta(t)$  is that it be some function of locally-available information such as the activity of other synapses on the post-synaptic neuron. Synaptic weight is typically included in the modification function in order to stabilize post-synaptic firing rates. The traditional Hebbian rule allows for run-away excitation, with the pre-synaptic neuron increasingly able to cause firing in the post-synaptic neuron. To that end, Erkki Oja proposed the following rule (Oja, 1982):

$$f[x(t), z(t), w] = x(t)z(t) - wz(t)^2$$

The second factor,  $wz(t)^2$ , is a weight decay term that prevents run-away excitation and unbounded increases in the weight. As we will see later, a network instantiated with this learning rule performs PCA.

In *three-factor* learning rules,  $\theta$  is typically a function that gates learning. For example, it could be an indicator function signaling the presence of a neuromodulator such as dopamine or acetylcholine (Kuśmierz, Isomura, & Toyozumi, 2017):

$$f[x(t), z(t), \theta(t)] = x(t)z(t)\theta(t)$$

$$\theta(t) \triangleq I(d(t) > \delta)$$

where the indicator function,  $I(\cdot)$ , is 1 when dopamine levels,  $d(t)$ , are greater than some threshold,  $\delta$ , and zero otherwise. Thus, Hebbian plasticity only occurs when a gating signal authorizes it to occur. This kind of learning rule is generally used to implement supervised and reinforcement learning algorithms, though it can be used for unsupervised learning as well. Biological examples of three-factor learning rules come by way of *heterosynaptic plasticity* (Harvey & Svoboda, 2007; Larsen & Sjöström, 2015; Lynch, Dunwiddie, & Gribkoff, 1977; L. Wang & Maffei, 2014). In heterosynaptic plasticity, the activity or weight of other local synapses influences the weight of the target synapse. This has been observed in many systems and brain regions, including the dopamine system, where the presence of dopamine can reduce the synaptic weight of neighboring inhibitory synapses (Ishikawa et al., 2013).

An apparent departure from the Hebbian-learning framework comes from spike-timing dependent plasticity (STDP) (Bi & Poo, 1998; Caporale & Dan, 2008; H. Markram et al., 2012; Henry Markram, Luebke, Frotscher, & Sakmann, 1997; Sjöström, Turrigiano, & Nelson, 2001). In STDP, the precise relative timing of pre- and post-synaptic spike pairs determines the synaptic weight change:

$$\frac{dw}{dt} = \eta g(t_z - t_x)$$

$$g(x) \triangleq \begin{cases} a_+ \exp\left(-\frac{x}{\tau_+}\right) & \text{for } x \geq 0 \\ a_- \exp\left(\frac{x}{\tau_-}\right) & \text{for } x < 0 \end{cases}$$

where  $t_x$  and  $t_z$  are the timing of pre- and post-synaptic spikes, respectively, and  $\tau$  and  $a$  are parameters governing the timing and magnitude of the synaptic weight change when

pre-synaptic spikes come before post-synaptic spikes (+) and when they come after (-). This rule, and variations of it, suggest that spike timing, not firing rate, is crucial to plasticity: pre just before post typically increases the weight ( $a_+ > 0$ ), while pre just after post decreases it ( $a_- < 0$ ).

Though the STDP weight modification is not written as a function of pre- and post-synaptic firing rates, it still falls under the Hebbian framework. In line with Hebb's original idea, if the pre-synaptic neuron reliably fires just before the post-synaptic neuron, then it likely has a causal influence on the post-synaptic neuron's firing. STDP increases the efficacy of this causal influence. Furthermore, the firing rate of a neuron can be taken to arbitrary temporal precision, with each instant in time simply signaling spike or no spike. In this regime, the function  $g$  (technically a function of the relative spike timing) can be thought of as a function of pre- and post-synaptic firing rates. Along these lines, recent theoretical work suggests STDP is approximately equivalent to the following model (Bengio, Mesnard, Fischer, Zhang, & Wu, 2015):

$$\frac{dw}{dt} = \eta x(t) \frac{dz}{dt}$$

where  $\frac{dz}{dt}$  is the first time derivative of the post-synaptic firing rate. So, we regain the canonical form of a Hebbian learning rule.

Another extension of Hebbian learning is the Bienenstock, Cooper, Munro (BCM) model of synaptic plasticity (Bienenstock, Cooper, & Munro, 1982; Cooper & Bear, 2012; Intrator

& Cooper, 1992). The BCM rule posits a sliding threshold that controls whether the synapse strengthens or weakens, and a decay term that stabilizes the synaptic weight. The basic BCM rule is:

$$\frac{dw}{dt} = \eta[x(t)z(t)(z(t) - \theta) - \epsilon w]$$

$$\theta \triangleq \mathbb{E}[z(t)^2]$$

(see Appendix on *Probability* for information on the expectation operator,  $\mathbb{E}[\cdot]$ ) In the BCM rule, we recognize the traditional Hebbian term,  $x(t)z(t)$ . However, it is now multiplied by an additional factor,  $(z(t) - \theta)$ . This implies that correlated firing will increase the weight when  $z(t) > \theta$  and decrease the weight when  $z(t) < \theta$ . In essence, the synapse strengthens only when the pre-synaptic neuron contributes to the post-synaptic neuron firing well above its average rate. The choice of threshold can vary, but this version sets the threshold to the average power of the output. In practice, the threshold is a function only of recent activity and the resulting sliding threshold establishes a form of *metaplasticity*: plastic changes in the threshold change the sign and strength of traditional plasticity (Bear, 2003; Yger & Gilson, 2015). The BCM rule, when instantiated in a neural network that accepts natural scenes as input, is also a good model of the visual system.

Biological examples of Hebbian learning are abundant. The most commonly-studied forms are known as long-term potentiation (LTP) and long-term depression (LTD) (Bliss & Lømo, 1973; Dudek & Bear, 1992; Malenka & Bear, 2004; Nicoll, 2017; Sjöström et al., 2001). In LTP, pre- and post-synaptic excitatory neuron pairs are simultaneously stimulated, such that pre- and post-synaptic firing are reliably correlated, and synaptic

transmission co-occurs with a sustained post-synaptic depolarization. This leads to a prolonged increase in the functional strength of the synapse (Bliss & Lømo, 1973; Nicoll, 2017). In many cases, results closely match the STDP version of Hebbian learning (Bi & Poo, 1998; Caporale & Dan, 2008). Thus, both LTP and LTD can occur depending on the precise relative timing of pre- and post-synaptic spiking. Subsequent research has shown that LTP is ubiquitous across the nervous system, but its precise mathematical form can be significantly more complex than the models presented here (Larsen & Sjöström, 2015; H. Markram et al., 2012; Sjöström et al., 2001). For example, synaptic modification is often a function of spike timing, stimulation frequency, and the number of coincident spikes incoming from other pre-synaptic neurons (Dudek & Bear, 1992; Sjöström et al., 2001). Most of these studies are performed *in vitro*, but there is also evidence *in vivo* of bidirectional Hebbian synaptic weight modifications (El-Boustani et al., 2018).

There are also examples of biological learning that do not obey the Hebbian framework. One commonly cited example is homeostatic plasticity through synaptic scaling (Abbott & Nelson, 2000; Turrigiano, 2008; Turrigiano, Leslie, Desai, Rutherford, & Nelson, 1998). Synaptic scaling maintains a fixed total excitatory synaptic weight across all synapses on a given neuron. In essence, synaptic strengths are continually re-normalized such that potentiated synapses can only ever gather a higher percentage of the total resource budget. This prevents unbounded synaptic weights and maintains an approximate target baseline firing rate (Turrigiano, 2008). Though this form of plasticity is markedly different from STDP, especially in its mechanistic underpinning, it still generally fits within the Hebbian

framework (Keck et al., 2017). Synaptic scaling can be thought of as a form of heterosynaptic plasticity, since plasticity at one synapse is a function of synaptic weights at other synapses.

Other forms of plasticity are more difficult to reconcile with Hebbian learning. Prominent examples involve changes to the intrinsic excitability of neurons, including the relationship between input current and firing rate (F-I curves). A recent paper observed learning-related sparsening of neural activity in primary visual cortex (Failor, Carandini, & Harris, 2021). As mice learned to discriminate two visual stimuli, population activity for just those two stimuli became sparser. Their data was consistent with an effective increase in the firing threshold for weakly-tuned neurons. Neurons that initially responded weakly to the stimulus set, with low firing rates that did not reveal a preference for any given stimulus, became unresponsive after learning. This effective change in the gain of responsiveness could be modeled by a single function that dynamically changed across trials and applied equally to all neurons. Because there were no changes to synaptic weights, this cannot be modeled with a Hebbian learning rule. Other studies of plasticity, and especially cortical development, emphasize connectivity between neurons, dendritic spine turnover, and synaptic pruning (Bhatt, Zhang, & Gan, 2009; Chechik, Meilijson, & Ruppin, 1998; Scholl, Connon, Ryan, Kamasawa, & Fitzpatrick, 2021; Segal et al., 2020; Xu et al., 2009; C. Zhang, Kolodkin, Wong, & James, 2017). Though related to Hebbian learning, the relevant quantity is no longer the synaptic weight but rather the simple presence or absence of a connection. What causes two neurons to connect in the first place? What causes axons and

dendrites to create new connections or sever old ones? Answers to these questions will likely be more complex than a simple Hebbian learning rule.

Given what we have explored so far, an obvious question that arises is: what computations do these rules perform? For example, if the synaptic weights of a network of neurons change according to LTP with synaptic scaling, what will that network do? The next two sections attempt to answer this question in reverse. Starting with two common algorithms, the predictive coder and PCA, we ask what types of learning rules are required to perform the requisite computations. In the end, we find that relatively simple local Hebbian and Anti-Hebbian learning rules implement surprisingly complex global computations.

### **Predictive Coders**

In engineering and telecommunications, predictive coders are used for efficient wired or wireless transmission (Elias, 1955; Nassar, 2001). The fundamental idea is to use a compressed form of the data for transmission (a form with lower information entropy), and then reconstruct the original data on the receiving end. To do so, the predictive coder uses data collected from the past to predict the future, transmitting only the error in that prediction. For example, assume the data to be transmitted,  $x_t \in \mathbb{R}^1$ , is generated in real-time and sent as it arrives. In naturalistic data, there are often low-frequency correlations present in the data: two consecutive frames in a movie are on average almost identical. Given this statistical structure, a simple model of the data is a first-order linear Gaussian autoregressive process:

$$x_{t+1}|x_t \sim \text{Normal}(x_{t+1} | \alpha x_t + \beta, \sigma^2)$$

$$x_0 \sim \text{Normal}(x_0 \mid \mu, \tau^2) \\ \text{with } |\alpha| < 1$$

We could simply transmit the data as is at each timestep. However, as we will see, less information can be transmitted, with no loss on the receiving end, by exploiting the statistical structure of the model. In order to understand the information cost of transmission, recall the definition of differential information entropy for the Gaussian distribution (see Appendix):

$$h(X) = \frac{1}{2} + \frac{1}{2} \log_2(2\pi\sigma^2)$$

Thus, we have for  $x_0$ :

$$h(x_0) = \frac{1}{2} + \frac{1}{2} \log_2(2\pi\tau^2)$$

For subsequent timesteps, we must calculate the marginal distribution of the data. For  $t = 1$ :

$$p(x_1) = \int p(x_1|x_0)p(x_0)dx_0 \\ x_1 \sim \text{Normal}(x_1 \mid \alpha\mu + \beta, \sigma^2 + \tau^2\alpha^2)$$

For later timesteps, a geometric series emerges, converging to:

$$x_t \sim \text{Normal}(x_t \mid \frac{\beta}{1-\alpha}, \frac{\sigma^2}{1-\alpha^2})$$

The average differential entropy of sending the raw data is therefore:

$$h(x_t) = \frac{1}{2} + \frac{1}{2} \log_2(2\pi \frac{\sigma^2}{1-\alpha^2})$$

The key engineering insight comes from recognizing that we might transmit the data using Algorithm 1 (see also Figure 2).

---

**Algorithm 1** Predictive Coder Transmission
 

---

Let the parameters of the model,  $\theta = \{\alpha, \beta\}$ , be known and distributed to both the source and receiver. Data,  $x_t$ , is emitted at the source, compressed, and then transmitted to the receiver and reconstructed.

- 1: **procedure** PredictiveCoder( $\theta$ )
  - 2:   Upon emission, transmit  $x_0$  as is from source to receiver. Store value in memory.
  - 3:   **for**  $t = 1, \dots, T$
  - 4:     At source and receiver, calculate prediction:  $\hat{x}_t \leftarrow \alpha x_{t-1} + \beta$
  - 5:     At source, upon emission of  $x_t$ , calculate prediction error:  $r_t \leftarrow x_t - \hat{x}_t$
  - 6:     At source, transmit  $r_t$  to receiver
  - 7:     At receiver, recover the original data:  $x_t \leftarrow \hat{x}_t + r_t$
  - 8:     At source and receiver, store  $x_t$  in memory
  - 9:   **end for**
  - 10: **end procedure**
- 

As long as source and the receiver both have access to the model, then the data generated at the source can be exactly recovered by the receiver. This procedure is actually utilized to transmit television signals.

The residuals, which are transmitted instead of the raw data, have an average differential entropy governed by the variance of the conditional density,  $p(x_{t+1}|x_t)$ :

$$h(r_t) = h(x_t|x_{t-1}) = \frac{1}{2} + \frac{1}{2} \log_2(2\pi\sigma^2)$$

This yields an improvement in the average amount of transmitted information:

$$\Delta h \triangleq h(x_t) - h(r_t) = \frac{1}{2} \log_2 \left( \frac{1}{1 - \alpha^2} \right) \geq 0$$

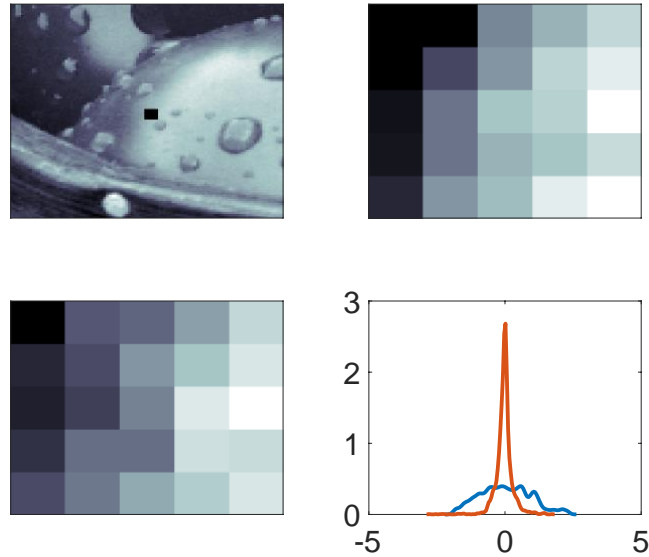
If  $\alpha = 0.97$ , then each transmission saves about 2 bits relative to sending the raw data. For high-dimensional data streams, this will constitute substantial savings in energy and resources. Furthermore, there is no loss of information. All of the information in the original data is contained in the transmitted prediction errors. The predictive coder is

therefore a compression algorithm, which removes statistical dependences from data and reduces entropy: it creates an efficient encoding of the data with minimal redundancy (see Appendix on *Efficient Coding*). The crucial operation is the exploitation of autocorrelations within the time series. If  $\alpha = 0$ , then there are no first-order temporal correlations in the data to exploit and there is no possibility for compression under this model.

As an example of predictive coding and its usefulness for compression, we built a spatial predictive coder using a dataset of natural images. In the spatial domain, the computational goal is to use data from one spatial location to predict data in a different spatial location. Once such a model has been fit, we could transmit the residuals as described in Algorithm 1, or use the model to fill in missing data. For example, image sensors often have dead pixels, which are an inevitable consequence of the manufacturing process. A spatial predictive coder would use surrounding pixels to fill in what would have been detected by a dead pixel. In the process of learning what needs to be filled in, the model learns about the structure of the data.

We took (50 by 50)-pixel regions of images and removed a central (5 by 5)-pixel frame, attempting to predict the held-out central frame with a deep neural network. Figure 1 illustrates the results. The top row shows an example image and its omitted central frame, while the bottom row is the predicted central frame along with density estimates for the raw data and residuals across a set of test images. While not able to perfectly reconstruct the central frame, the model captures many statistical regularities within the dataset and

therefore allows for compression: the entropy of the residual distribution is significantly reduced relative to the raw data distribution.



**Figure 1: Deep Neural Network Spatial Predictive Coder**

**Top Left:** Natural scene with 5x5 central frame omitted. **Top Right:** True omitted central frame, on a different color scale than the scene at left in order to show details. **Bottom Left:** Predicted central frame. Prediction was created by a deep neural network trained through backpropagation on a set of natural images (with the image at Top Left held out). The input to the network is the original image (minus the central frame) and the desired output is the central frame. Training loss was the sum of squared residuals between the true central frames and the predicted central frames. **Bottom Right:** Kernel density estimates showing entropy reduction (x-axis is normalized pixel value & y-axis is probability density). In blue, the distribution of the raw data from the central frame across a set of 100 test images not used to train the network. In orange, the distribution of residuals between the model output and the raw data. The raw data has a standard deviation about three times greater than that of the residuals. This  $\sim 3$ -fold decrement in standard deviation constitutes a  $\sim 1.5$ -bit decrement in the entropy of the distribution.

The best possible model minimizes the entropy of the residual distribution, leaving no opportunity for further compression. In probability and information theory, this optimal distribution is known as a factorized distribution, wherein each component is statistically independent of all other components (Cover & Thomas, 2005; Wasserman, 2004). In the

predictive coding model, the residuals across timepoints are statistically independent because their joint distribution can be written in a factorized form:

$$p(r_1, \dots, r_T) = \prod_{t=1}^T p(r_t)$$

This is a general condition that must be met for statistical independence, and in general, data transmitted in this way will be more efficient.

### *A Predictive Coder with Local Anti-Hebbian Plasticity*

In this section, we derive a simple biologically-plausible neural network that performs predictive coding (Hosoya et al., 2005; Srinivasan et al., 1982). For an artificial neural network to be biologically plausible, it is often required to meet two constraints (Dayan & Abbott, 2005; Gerstner & Kistler, 2002):

- 1) Local computation: each neuron in the network computes only some function of its weighted synaptic inputs. This requirement is satisfied in most artificial neural networks.
- 2) Local plasticity: the strength of a synaptic connection can only change based on information available locally at the synapse. This is rarely satisfied in artificial neural networks.

The predictive coding network will satisfy these constraints and transmit prediction errors at its output layer.

To begin, consider the multivariate form of the linear autoregressive Gaussian model:

$$\begin{aligned} \mathbf{x}_{t+1} | \mathbf{x}_t &\sim \text{Normal}(\mathbf{x}_{t+1} | \mathbf{A}\mathbf{x}_t + \mathbf{b}, \mathbf{\Sigma}) \\ \mathbf{x}_0 &\sim \text{Normal}(\mathbf{x}_0 | \boldsymbol{\mu}, \mathbf{\Lambda}) \\ \theta &\triangleq \{\mathbf{A}, \mathbf{b}, \mathbf{\Sigma}, \boldsymbol{\mu}, \mathbf{\Lambda}\} \end{aligned}$$

where  $\mathbf{x}_t \in \mathbb{R}^d$ . Previously, we assumed the model parameters were known. To actually learn parameters from data, we must perform some kind of optimization. This can be done by maximum likelihood estimation (Appendix). The log likelihood function is:

$$\mathcal{L}(\theta; \mathbf{x}) = \log(P(\mathbf{x}|\theta)) = \log(P(\mathbf{x}_0 | \boldsymbol{\mu}, \boldsymbol{\Lambda})) + \sum_{t=0}^{T-1} \log(P(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{A}, \mathbf{b}, \boldsymbol{\Sigma}))$$

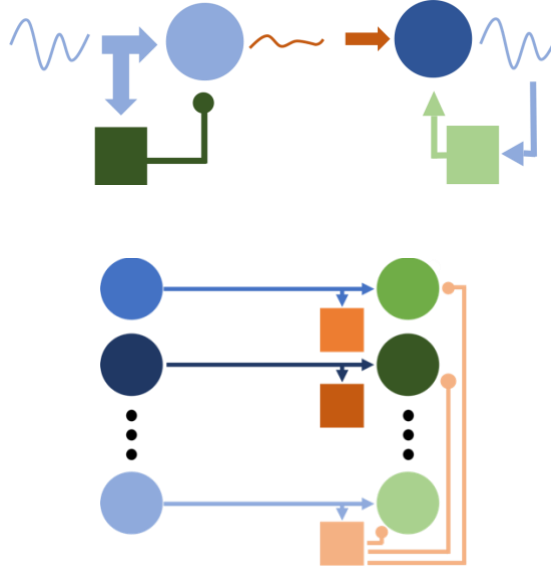
After some manipulation of variables, and ignoring  $\boldsymbol{\mu}$  and  $\boldsymbol{\Lambda}$ , we have:

$$\mathcal{L}(\theta; \mathbf{x}) = -\frac{1}{2} \log(|\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{t=0}^{T-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t - \mathbf{b})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t - \mathbf{b}) + \text{const}$$

By maximizing this scalar quantity with respect to the parameters,  $\theta$ , we learn optimal estimates for those parameters. When considered with respect to  $\mathbf{A}$  and  $\mathbf{b}$ , maximizing this log likelihood is identical to minimizing the cost:

$$C \triangleq \sum_t (\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{b})^T (\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{b})$$

The thing to be minimized is the squared model residual, or squared prediction error, where the residual is  $\mathbf{r}_t \triangleq \mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{b}$ . A relatively simple neural network model can take data,  $\mathbf{x}_t$ , as input and output this residual (Figure 2). A network connected in this way minimizes the desired cost function by minimizing the squared activity of its output neurons, effectively an energy constraint.



**Figure 2: Local Predictive Coder Network**

**Top:** Schematic representation of the predictive coding algorithm, as used in engineering and telecommunications (see Algorithm 1). Data (light blue squiggle) enters the system on the left, where it is passed to a prediction engine (dark green square) and a transmitter (light blue circle). The transmitter computes the residual (orange squiggle) between data and prediction, and then sends it to the receiver (dark blue circle). The receiver then inverts this process to reconstruct the data, but note that all of the information in the original signal is also present in the residual. **Bottom:** Visualization of a neural network that performs linear predictive coding. Inputs,  $\mathbf{x}_t$ , come in on the left (blue neurons, where each neuron represents one entry in the vector  $\mathbf{x}_t$ ). A set of inhibitory neurons (orange) receives the input and introduces a one time-step delay into the network. The output neurons (green) receive both the input at time  $t$  (blue to green connections) and a delayed input, from  $t - 1$ , transformed through the inhibitory neurons by a weight matrix  $\mathbf{A}$  (orange-to-green connections). Neurons at the output transmit the residual between the current input,  $\mathbf{x}_t$ , and the predicted input,  $\mathbf{A}\mathbf{x}_{t-1}$ . Only those connections emanating from one inhibitory neuron, *light orange at bottom*, depicted for clarity (these represent one column of  $\mathbf{A}$ ).

Instead of performing a global optimization on an entire dataset, we can also learn the parameters online, as data arrives, by gradient descent with updates:

$$\begin{aligned} \mathbf{A} &\leftarrow \mathbf{A} - \eta \frac{\partial C}{\partial \mathbf{A}} \\ \mathbf{b} &\leftarrow \mathbf{b} - \eta \frac{\partial C}{\partial \mathbf{b}} \end{aligned}$$

Each time a new data point arrives, we update our estimates of the parameters until we arrive at a steady state in which the derivatives equal zero. Taking the appropriate derivatives:

$$\begin{aligned}\frac{\partial C}{\partial \mathbf{A}} &= -2(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{b})\mathbf{x}_{t-1}^T \\ \frac{\partial C}{\partial \mathbf{b}} &= -2(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{b})\end{aligned}$$

Note we can absorb the value of 2 into the learning rate,  $\eta$ . The update for  $\mathbf{A}$  is a function only of local information, in particular the activity of connected pre- and post-synaptic pairs:

$$\Delta \mathbf{A} \triangleq \eta \frac{\partial C}{\partial \mathbf{A}} = -\eta \mathbf{r}_t \mathbf{x}_{t-1}^T$$

This is a traditional Anti-Hebbian learning rule, since the effect of the inputs is inhibitory and the weight change is only a function of pre- and post-synaptic activity (Dayan & Abbott, 2005; Földiák, 1990; Gerstner & Kistler, 2002). We therefore have a biologically-plausible implementation of predictive coding. All of the information available in the inputs,  $\mathbf{x}_t$ , is preserved in the outputs,  $\mathbf{r}_t$ . In essence, the network learns to eliminate redundancy in the inputs and then to output independent, and therefore fully uncorrelated, signals. The crucial computation requires the inhibitory inputs to become fully decorrelated from the output neurons. The retina appears to perform a similar predictive coding operation (Hosoya et al., 2005; Palmer, Marre, Berry, & Bialek, 2015; Srinivasan et al., 1982), and later visual areas may use related principles (Rao & Ballard, 1999; Michael W Spratling, 2010; Zmarz & Keller, 2016). This connection will be explored further in the section on Vision.

## PCA

Most dimensionality reduction techniques, including principal components analysis (PCA), are examples of unsupervised learning. Dimensionality reduction maps a high-dimensional dataset into a lower-dimensional subspace, compressing the data, reducing redundancy, easing the computational load of subsequent processing steps, and allowing for efficient data transfer. From an information theoretic perspective, dimensionality reduction reduces the entropy of the source distribution, so far as the data allows. Importantly, there is significant evidence that nervous systems compress incoming sensory data, for similar reasons of energy, time, and information efficiency (Atick & Redlich, 1992; H. B. Barlow, 1989; Dan, Atick, & Reid, 1996; Olshausen & Field, 1996; Rao & Ballard, 1999; Srinivasan et al., 1982; Sterling & Laughlin, 2015).

PCA is the most prominent dimensionality reduction technique. It is the prototypical example of a class of unsupervised learning models called autoencoders. Assume data,  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T \in \mathbb{R}^{N \text{ by } d}$ , has been collected and centered so that its columns have zero mean, then PCA solves the following optimization problem:

$$\begin{aligned} \hat{\mathbf{V}} &= \arg \min_{\mathbf{V}} \|\mathbf{X} - (\mathbf{X}\mathbf{V})\mathbf{V}^T\|_2^2 \\ \text{s. t. } \mathbf{V}^T\mathbf{V} &= \mathbf{I} \ \& \ \text{rank}(\mathbf{V}) \leq \text{rank}(\mathbf{X}) \end{aligned}$$

In this case, we pre-specify the rank of  $\mathbf{V}$  and find the optimum for that rank. Deciding upon an optimal rank requires some form of rank penalization or cross-validation procedure. Note, if  $\mathbf{X} \in \mathbb{R}^{N \text{ by } d}$  and the rank of  $\mathbf{V}$  is  $q$ , then  $\mathbf{V} \in \mathbb{R}^{d \text{ by } q}$ . The columns of

$\hat{\mathbf{V}}$  are known as eigenvectors of the sample covariance matrix,  $\mathbf{X}^T \mathbf{X}$ , or the right singular vectors of the data matrix,  $\mathbf{X}$ .

There are many algorithms to find the optimal eigenvector matrix,  $\hat{\mathbf{V}}$ , but perhaps the most common is to perform an eigen-value/-vector decomposition of the data covariance matrix.

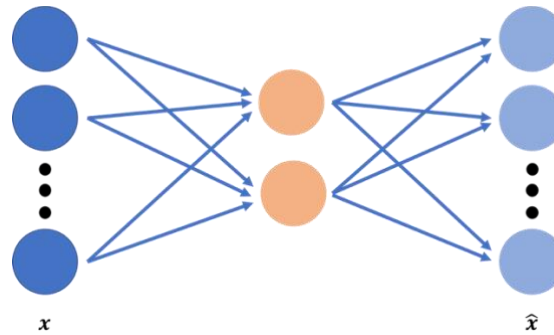
The decomposition solves:

$$(\mathbf{X}^T \mathbf{X}) \mathbf{v}_i = \lambda_i \mathbf{v}_i$$

where  $\lambda_i$  is the eigenvalue associated with eigenvector  $\mathbf{v}_i$ . In this formulation,  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_q)$  and we can define  $\mathbf{\Lambda} \triangleq \text{diag}(\lambda_1, \dots, \lambda_q)$ . Solving this equation, first for the eigenvalues and then for the eigenvectors, yields a solution to the optimization problem above. In addition, assuming  $d > q$ , the eigenvectors and eigenvalues provide an optimal low-rank approximation to the covariance matrix:

$$\mathbf{C} \triangleq \mathbb{E}[\mathbf{X}^T \mathbf{X}] - \mathbb{E}[\mathbf{X}^T] \mathbb{E}[\mathbf{X}] \approx \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$$

In essence, PCA is a data compression and reconstruction algorithm. It projects the data into an orthonormal low-dimensional subspace, and then projects it back into the original space, attempting to preserve as much variance as possible to create a good reconstruction (see Figure 3).



**Figure 3: Visual Representation of PCA**

The data,  $\mathbf{x}$ , is projected into a latent 2-dimensional space by matrix multiplication with  $\mathbf{V}^T$ , and then projected back into the original space by multiplication with  $\mathbf{V}$ . The columns of  $\mathbf{V}$  are basis functions that span the high-dimensional space and capture statistical correlations across dimensions. Projection down to the 2-dimensional space creates a bottleneck, forcing the eigenvectors to preserve information for the subsequent reconstruction,  $\hat{\mathbf{x}}$ . If we assume the relationship between the input and latent space is nonlinear, then this visualization also serves as a more general representation of an autoencoder. In that context, the left side is known as the encoder (from data to latent), while the right side is known as the decoder (from latent to reconstructed data).

PCA also admits of a probabilistic formulation, which can be understood as learning a generative model of the data (C. M. Bishop, 1999). The term generative is used because data can be simulated/synthesized that bears a resemblance to the original data. Often, the generative model includes a set of latent variables, which are not observed but contribute in a direct way to the observations. The latent variables can heuristically be thought of as the underlying causes of the data. Tipping and Bishop first proposed the probabilistic model, which is identical to traditional PCA under certain conditions (C. Bishop & Tipping, 1999). We present the model here to provide an example of a simple generative model, which gives the requisite background for understanding the visual system as a device that learns a generative model.

Tipping and Bishop begin by representing the latent subspace with a simple factorized probability distribution:

$$p(\mathbf{z}) = \text{Normal}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}_q)$$

where  $q$  is the dimensionality of the subspace ( $q=2$  in the example from Figure 3). If the original data is  $d$ -dimensional, then we model the data conditional on the latent subspace as:

$$p(\mathbf{x}|\mathbf{z}) = \text{Normal}(\mathbf{x} \mid \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$$

Integration of the two distributions yields the marginal distribution of the data:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} = \text{Normal}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}_d)$$

Optimization solves for those values of  $\boldsymbol{\theta} \triangleq \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}$  that maximize  $p(\mathbf{x})$ , the likelihood of the data (Appendix). Tipping and Bishop showed that

$$\widehat{\mathbf{W}}_{ML} = \mathbf{V}(\boldsymbol{\Lambda} - \sigma^2 \mathbf{I}_q)^{1/2} \mathbf{R}$$

where  $\mathbf{V}$  is the matrix of traditional PCA eigenvectors,  $\boldsymbol{\Lambda}$  is the diagonal matrix of eigenvalues, and  $\mathbf{R}$  is an arbitrary orthogonal rotation matrix ( $q$  by  $q$ ). This probabilistic method converges to the traditional method as  $\sigma^2 \rightarrow 0$ .

To infer the latent variables from the data, we have by Bayes' theorem (Appendix):

$$p(\mathbf{z}|\mathbf{x}) = \text{Normal}(\mathbf{z} \mid \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x} - \boldsymbol{\mu}), \sigma^2 \mathbf{M}^{-1})$$

$$\mathbf{M} \triangleq \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbf{I}_q$$

If we know  $\mathbf{W}$ , or have a good estimate of it, then we can infer the latent factors,  $\mathbf{z}$ . With a sufficient number of latent factors, all of the information in the original data is available in

those latent factors. In many cases of practical importance, storing or transmitting only the latent factors provides a significant information savings.

### *A Local Implementation of PCA*

Based on the work of Erkki Oja (Oja, 1982), Terence Sanger proposed a biologically-plausible neural network implementation of PCA sometimes known as Sanger's Rule or the generalized Hebbian algorithm (Gorrell, 2006; Oja, 1992; Sanger, 1989). Oja and Sanger envisioned a way for an artificial neural network to naturally, automatically, perform PCA as data arrives, learning the eigenvectors in real time and outputting the latent variables. Such a network is an input-output mapping function. The network receives input data,  $\mathbf{x}$ , and transforms it through multiplication with a weight matrix to yield a compressed representation,  $\mathbf{z}$ , at the output:

$$\mathbf{z} = \mathbf{W}^T \mathbf{x}$$

There might be  $d$  neurons at the input,  $\mathbf{x} \in \mathbb{R}^d$ , and  $q$  at the output,  $\mathbf{z} \in \mathbb{R}^q$ , such that  $\mathbf{W} \in \mathbb{R}^{d \text{ by } q}$ . We can envision the input layer as faithfully transmitting inputs (perhaps images) through its activity. The weight matrix then governs the activity at the output layer, which is effectively the latent space from the probabilistic interpretation of PCA. If we assume the input layer is indexed by  $i$  and the output layer by  $j$ , then the weight of the synaptic connection from neuron  $i$  to  $j$  is:

$$w_{ij}$$

and we can rewrite the matrix equation above for an output neuron as:

$$z_j = \sum_{i=1}^d w_{ij} x_i$$

Thus, the set of activities in the input layer,  $\mathbf{x} = (x_1, \dots, x_d)^T$ , controls the activity at the output layer through a weighted sum. The generalized Hebbian algorithm then specifies a synaptic weight modification, or update, that occurs each time an input is passed through the network:

$$\Delta w_{ij} = \eta z_j (x_i - \sum_{k=1}^j w_{ik} z_k)$$

where  $\eta$  is the learning rate. This is a modification of the traditional Hebbian learning rule, which we recover by removing the  $\{-\sum_{k=1}^j w_{ik} z_k\}$  term. The learning rule is known as Sanger's rule.

Sanger proved that a linear network operating under this modification rule will converge to a steady state, such that the columns of the weight matrix are equivalent to the eigenvectors of the data covariance matrix. That is,

$$\mathbf{W} = \mathbf{V}$$

where  $\mathbf{V}$  is the same eigenvector matrix presented above. Unlike traditional PCA algorithms, however, Sanger's rule requires neither the data covariance matrix nor the entire dataset. Data passes sequentially through the network and modifies the synapses with each pass, automatically driving the weights toward the eigenvectors. Though the inputs to the network may be highly correlated, the outputs,  $\mathbf{z}$ , will be fully uncorrelated.

Following the proof from Oja (Oja, 1982), imagine there is only a single output neuron, with the weight modification rule in vector form:

$$\Delta \mathbf{w} = \eta (\mathbf{x} - \mathbf{z}\mathbf{w})\mathbf{z}$$

$$z \triangleq \mathbf{x}^T \mathbf{w}$$

This is known as Oja's rule. Under the rule, the synaptic weights converge to the first eigenvector of the data covariance matrix, which we now prove.

If the sequence of inputs is a stationary process with zero mean and unchanging covariance,

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \mathbf{0} \\ \mathbb{E}[\mathbf{x}\mathbf{x}^T] &= \mathbf{\Sigma}\end{aligned}$$

then the long-run weight modification, with respect to the distribution of  $\mathbf{x}$ , is:

$$\begin{aligned}\mathbb{E}[\Delta \mathbf{w}] &= \mathbb{E}[\eta(\mathbf{x} - (\mathbf{x}^T \mathbf{w})\mathbf{w})\mathbf{x}^T \mathbf{w}] \\ &= \eta \mathbb{E}[\mathbf{x}\mathbf{x}^T] \mathbf{w} - \eta \mathbb{E}[(\mathbf{x}^T \mathbf{w})\mathbf{w}(\mathbf{x}^T \mathbf{w})] \\ &= \eta \mathbf{\Sigma} \mathbf{w} - \eta (\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}) \mathbf{w}\end{aligned}$$

where we used the facts that  $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mathbf{\Sigma}$ ,  $\mathbf{x}^T \mathbf{w}$  is a scalar quantity, and  $\mathbf{x}^T \mathbf{w} = \mathbf{w}^T \mathbf{x}$ .

The condition required for steady-state convergence of the synaptic weights is:

$$\mathbb{E}[\Delta \mathbf{w}] = \mathbf{0}$$

Each iteration,  $\mathbf{w} \leftarrow \mathbf{w} - \Delta \mathbf{w}$ , is a gradient descent update that converges to this optimum / steady state (technically, we would also need to prove that this optimum is indeed a local or global minimum). Therefore, the following equality holds at equilibrium:

$$(\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}) \mathbf{w} = \mathbf{\Sigma} \mathbf{w}$$

The only stable solution for  $\mathbf{w}$  is the first eigenvector,  $\mathbf{v}_1$ , of the covariance matrix. To see this, recall that for a single eigenvector, the optimal approximation to the covariance matrix is,

$$\mathbf{\Sigma} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T$$

Substituting into the equation above:

$$(\mathbf{v}_1^T \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T \mathbf{v}_1) \mathbf{v}_1 = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T \mathbf{v}_1$$

Since  $\mathbf{v}^T \mathbf{v} = 1$  for any eigenvector,

$$\lambda_1 \mathbf{v}_1 = \lambda_1 \mathbf{v}_1 .$$

And, therefore, repeated application of Oja's update transforms the weight vector from any arbitrary starting values to the first eigenvector of the covariance of the input data. Sanger's proof for the general case, with many output neurons, begins with this derivation and proves by induction that the remaining eigenvectors emerge as more output neurons are added. Therefore, a network operating under this synaptic learning principle will perform optimal linear dimensionality reduction on incoming data.

Oja's rule is a local Hebbian learning rule. It simply adds a correction factor to the traditional Hebbian learning rule, normalizing the input weights and preventing them from growing without bound. When the learning rule is generalized to many output neurons (Sanger's rule), however, it is no longer biologically plausible. Careful examination of Sanger's rule reveals the computation requires non-local plasticity, as the weight update between two neurons depends on the synaptic weight of many other connections in the network.

PCA, and its neural network implementations, are very relevant to the visual system. Briefly, Zylberberg et al. showed that a Sanger's rule network with an additional sparsity constraint on output layer activity can be implemented with local learning rules (Zylberberg, Murphy, & Deweese, 2011). In essence, they prove that Sanger's non-local rule reduces to a local rule when the activity of the outputs is very sparse. When trained on

whitened natural images, the network's output neurons learn receptive fields that are comparable to those found in macaque V1. Similar studies have shown that autoencoder networks trained under a variety of sparsity constraints also create retina- and V1-like representations (Atick & Redlich, 1992; Bienenstock et al., 1982; Falconbridge et al., 2006; Földiák, 1990; Ocko, Lindsey, Ganguli, & Deny, 2018; Olshausen & Field, 1996, 1997; Rao & Ballard, 1999). These theoretical findings support the hypothesis that the visual system is an unsupervised learning device.

### **Deep Unsupervised Learning**

Despite their success in modeling certain aspects of the early visual system, the predictive coding and PCA algorithms described above are linear models. Linearity is a nice property for theoretical derivations, but its simplicity limits computational complexity. In addition, biological neurons are not linear: they are under energy and other biophysical constraints that limit their average and peak firing rates (Dayan & Abbott, 2005; S. B. Laughlin, 2001; Simon B. Laughlin, De Ruyter Van Steveninck, & Anderson, 1998); they exhibit firing thresholds; their dendrites do not simply sum inputs (Bicknell & Häusser, 2021; Gidon et al., 2020; Sorg et al., 2016; Takahashi et al., 2020), among other nonlinear features. Deep neural networks may therefore be a more appropriate way to model complex networks of biological neurons, as both are highly nonlinear. Though often criticized for being “black boxes”, in that a human-interpretable explanation of their computations is rarely possible, deep neural networks provide very good models of the visual system (Bakhtiari et al., 2021; Cadena et al., 2019; Higgins et al., 2020; Yamins & Dicarlo, 2016; Yamins et al., 2014;

Zhuang et al., 2021). Here, we will present two unsupervised deep neural networks models that have been shown to accurately model the mammalian visual system.

### *Variational Autoencoders*

The autoencoder is a general class of models, of which PCA is the representative example. As depicted in Figure 3, autoencoders compress data into a low-dimensional representation and then attempt to reconstruct it. The variational autoencoder (VAE) is a specific class of autoencoder that makes several simplifying and regularizing assumptions about the data and latent space (Doersch, 2016; Kingma & Welling, 2014). As with PCA, the VAE is a latent-variable model and its latent space has the same distribution as PCA:

$$p(\mathbf{z}) = \text{Normal}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}_q)$$

In the VAE, however, the relationship between the latent and observed data is a nonlinear function that is parameterized by a deep neural network,  $\mathbf{x} = f(\mathbf{z})$ :

$$p(\mathbf{x}|\mathbf{z}) = \text{Normal}(\mathbf{x} \mid f(\mathbf{z}), \sigma^2 \mathbf{I}_d)$$

Because of the complex functional dependence between  $\mathbf{z}$  and  $\mathbf{x}$ , the integral,

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$$

is no longer tractable, but we still wish to find the parameters that maximize  $p(\mathbf{x})$ . Variational inference solves this problem with a clever approximation (C. M. Bishop, 2006; Blei, Kucukelbir, & McAuliffe, 2017). It introduces a surrogate posterior distribution,  $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$ , and writes:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) \frac{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} d\mathbf{z}$$

$$= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \left[ \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \right]$$

Taking the log of both sides:

$$\log(p(\mathbf{x})) = \log(\mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \left[ \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \right])$$

By Jensen's inequality (Wasserman, 2004),

$$\log(p(\mathbf{x})) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \left[ \log\left(\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}\right) \right]$$

Jensen's inequality states that for any convex function,  $\varphi(x)$ ,  $\varphi(\mathbb{E}[x]) \leq \mathbb{E}[\varphi(x)]$ . If our convex function is  $\varphi(x) = -\log(x)$ , then  $\log(\mathbb{E}[x]) \geq \mathbb{E}[\log(x)]$ .

The quantity on the right is often written:

$$ELBO(q) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} [\log(p(\mathbf{x}|\mathbf{z}))] - \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})} \left[ \log\left(\frac{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{z})}\right) \right]$$

The term ELBO means evidence lower bound, because the quantity is a lower bound on  $\log(p(\mathbf{x}))$ , sometimes referred to as the evidence in Bayesian inference (C. M. Bishop, 2006). Maximizing the ELBO with respect to the distribution  $q$  maximizes a lower bound on the quantity that we originally set out to maximize,  $p(\mathbf{x})$ . In addition, the difference between the ELBO and the evidence is:

$$\log(p(\mathbf{x})) = ELBO(q) + KL(q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})|p(\mathbf{z}|\mathbf{x}))$$

The KL divergence on the right-hand side is a non-negative quantity that is effectively a distance measure between distributions (Kullback & Leibler, 1951). Based on its properties,  $ELBO(q) = \log(p(\mathbf{x}))$  only when  $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{z}|\mathbf{x})$ .

The reason to use the surrogate distribution  $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$  is that  $p(\mathbf{z}|\mathbf{x})$  cannot be computed directly. One must choose a family of distributions,  $\mathcal{Q}$ , such that  $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$  approximates  $p(\mathbf{z}|\mathbf{x})$  and allows the ELBO to be computed during stochastic gradient ascent/descent. Because  $p(\mathbf{x}|\mathbf{z})$  and  $p(\mathbf{z})$  are Gaussian, in the present case

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = \text{Normal}(\mathbf{z}|g(\mathbf{x}), h(\mathbf{x}))$$

where  $g(\mathbf{x})$  and  $h(\mathbf{x})$  are also nonlinear functions computed by deep neural networks. Relating this back to the autoencoder,  $\mathbf{z} = g(\mathbf{x})$  is often called the encoder, and  $\mathbf{x} = f(\mathbf{z})$  the decoder. These are the left and right side of the autoencoder depicted in Figure 3.

In practice, the VAE has an uncanny ability to create convincing deep fakes, and the latent representations can be used for a variety of transfer learning tasks, such as object recognition or localization. In Neuroscience, a recent paper from Higgins et al. showed that the latent variables of a variational autoencoder trained on images of faces accurately captured the firing variability of single neurons in inferior temporal cortex (IT) of macaque monkeys (Higgins et al., 2020). They trained their model in an unsupervised way on images of faces, and then showed that individual latent factors had a strong correspondence to single units recorded in IT, explaining up to 25% of the firing variance. Future research will likely discover other correspondences between VAEs and biological neural data.

### *Self-Supervised Learning*

The goal of many self-supervised learning methods is to discover a low-dimensional data representation that is invariant to a variety of distortions and augmentations. These were originally inspired by our impressive human ability to recognize objects and faces invariant

to lighting conditions, angle, size, color, contrast, etc. One representative model is known as Barlow twins (Zbontar et al., 2021), so named for Horace Barlow. The derivation of Barlow Twins involves an information bottleneck, and Barlow famously hypothesized that the optic nerve acts as a bottleneck for visual information leaving the retina (H. B. Barlow, 1961).

The information bottleneck principle formalizes a variety of problems involving compressed latent representations of data (Tishby, Pereira, & Bialek, 2000). For Barlow twins, we have some set of original images,  $X$ , a set of distorted images,  $Y$ , along with a compressed latent representation,  $Z$ . The information bottleneck attempts to minimize the following loss functional:

$$\mathcal{L} = I(Z; Y) - \beta I(Z; X)$$

The loss is a function of the parameters that govern the mapping from the distorted images to the latent representation ( $\mathbf{z} = f(\mathbf{y})$ ), with the function  $f$  typically implemented by a deep neural network. Based on this objective function, the learned latent representation,  $Z$ , will be minimally informative of the distortions,  $Y$ , and maximally informative of the original images,  $X$  (see Appendix for definition of the mutual information,  $I(X; Y)$ ). This formulation establishes the desired invariance to distortion.

Based on information theoretic identities (Appendix), the loss becomes

$$\mathcal{L} = H(Z) - H(Z|Y) - \beta[H(Z) - H(Z|X)]$$

Neural networks are deterministic functions of their inputs, so  $H(Z|Y) = 0$  (given  $Y$ , there is no uncertainty on  $Z$ ). And, therefore,

$$\frac{\mathcal{L}}{\beta} = H(Z|X) + \frac{1-\beta}{\beta} H(Z)$$

Computing these entropies for high-dimensional data can be very difficult, so the authors make the simplifying assumption that  $Z$  be Gaussian. The entropy of a Gaussian distribution is a function of the log determinant of its covariance matrix, which is easy to compute. Given a few more manipulations, they arrive at the following procedure for training the self-supervised model. They pass two distorted versions of the same image through two identical neural networks (hence “twins”) and force the networks to give outputs whose cross-correlation equals the identity matrix. This creates a low-dimensional representation that is invariant to distortion (the outputs of the two networks must be essentially the same) and decorrelated (each output dimension is statistically independent of the others, up to second order), the latter being comparable to Barlow’s notion of efficient coding through redundancy reduction. Related *contrastive learning* methods have a similar general flavor (Oord et al., 2018; Zimmermann et al., 2021).

Recent work in Neuroscience has shown that deep neural networks trained with these self-supervised approaches achieve state-of-the-art performance on prediction of neural activity in the macaque ventral stream (Zhuang et al., 2021). Singer et al. show that an artificial neural network trained to predict future video frames has internal units with spatiotemporal receptive fields that resemble those found in V1 (Singer et al., 2018). In addition, a contrastive predictive coding algorithm trained on a dataset of natural videos accurately predicts neural activity across mouse visual cortical areas (Bakhtiari et al., 2021). The contrastive predictive coding algorithm creates a latent representation and a related context

variable, and then uses the context to predict future values of the latent variable. The same authors showed that a network with two parallel pathways automatically segregates video data into ventral-like object information and dorsal-like motion information. Hence their title, *The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning*. These results are consistent with the notion of the visual system as an unsupervised learning device and suggest ways in which the visual system might perform the necessary computations: attempting to reconstruct sensory inputs from a low-dimensional representation; passing data through a bottleneck with statistically independent outputs; predicting the future; comparing different views of the same scene, perhaps before and after a saccade; guessing what lies behind an occlusion; pretending there are occlusions and guessing what lies behind those; learning a common latent space across modalities.

## VISION

The visual system is a highly complex and interconnected set of brain regions, all involved in the various tasks of seeing: navigating the environment, guiding movement, avoiding predation, capturing prey, identifying objects and conspecifics and mates, entraining circadian rhythms, controlling pupillary reflexes, among others. Light enters the visual system through the retina, where it is parsed into parallel information streams that are mostly distributed to thalamus, but also superior colliculus, amygdala, hypothalamus, and a variety of other subcortical regions (Dhande, Stafford, Lim, & Huberman, 2015; Kandel et al., 2014; Seabrook, Burbridge, Crair, & Huberman, 2017).

In this section, we will provide a thorough review of the visual system by focusing on the structure and function of the three critical early visual regions: retina, thalamus, and primary visual cortex. Our focus will ultimately be on the mouse visual system, though general information from research on non-human primates, cats, rabbits, guinea pigs, and salamanders will also be included. The overall framing of this section is in relation to the previous section on unsupervised learning, so the goal is to think about vision in the context of building models and efficiently encoding information. The section ends with a review of synaptic plasticity in the visual system, which is the biological basis of learning.

## **Retina**

### *Structure*

The retina is a fine tissue layer laid across the back of the eye. Light enters the eye through the pupil, where it is projected onto the retina in the manner of a *camera obscura*. In mammals, light travels through several layers of cell bodies and neuropil before arriving at the photoreceptors. These layers, from the back of the eye toward the front, are known as the photoreceptor layer, outer nuclear layer, outer plexiform layer, inner nuclear layer, inner plexiform layer, ganglion cell layer, and nerve fiber layer (Kandel et al., 2014; Kolb, 2003; Masland, 2001). There are also membranes on either side of these layers, which sustain the structure of the retina, absorb scattered light, and provide nutritive functions. Nuclear layers contain cell bodies, while plexiform layers consist almost entirely of axons and dendrites. There are only five primary neuronal cell types in the retina: photoreceptors, bipolar cells, horizontal cells, amacrine cells, and ganglion cells. Information generally flows in a feedforward direction from photoreceptors to bipolar cells to ganglion cells, and then through the nerve fiber layer toward the optic nerve and ultimately the rest of the central nervous system.

Vision starts at the photoreceptors, rods and cones, which absorb photons and transduce electromagnetic waves into a chemical signaling cascade. The culmination of this process is a graded decrease in the steady-state release of glutamate from photoreceptor synaptic terminals. The primary targets of these axon terminals are the dendrites of bipolar cells in the outer plexiform layer. There are about 10 types of bipolar cell, generally characterized

by the type of glutamate receptor they express, their physiological response to light, and also their morphology (Kolb, 2003; Masland, 2001). Photoreceptors also synapse onto horizontal cells, which provide lateral inhibitory feedback to other photoreceptors. The bipolar cells then provide graded synapses onto ganglion cell dendrites and amacrine cells in the inner plexiform layer. The amacrine cells, of which there are at least 30 types, provide lateral inhibition onto bipolar cells and sometimes directly onto ganglion cells (Balasubramanian & Sterling, 2009; Kolb, 2003; Masland, 2001). There are about 30 types of ganglion cell in the mouse retina, the first retinal cell to signal via action potential (Dhande et al., 2015; Seabrook et al., 2017). Each type of ganglion cell has a unique morphology, especially regarding the size and location of its dendritic arbor, and its unique set of connections with different types of bipolar and amacrine cells. Each type also tiles the entire retina in a mosaic and so they seem to transmit parallel information streams to cortical and sub-cortical targets (Hoon, Okawa, Della Santina, & Wong, 2014; Kolb, 2003).

Despite its relatively simple feedforward architecture, the retina is a highly complex signal processing device (Balasubramanian & Sterling, 2009; Hosoya et al., 2005; Meister & Berry, 1999; Palmer et al., 2015). In the human retina, for example, there are about  $10^8$  photoreceptors and only  $10^6$  axonal fibers traversing the optic nerve (Meister & Berry, 1999). These fibers therefore seem to transmit a multiplexed and compressed representation of visual information to the central nervous system (Balasubramanian & Sterling, 2009; Meister & Berry, 1999). Despite decades of research cataloging the various

cell types, their interconnections, and their precise functional roles, there remain many areas of active investigation (Hoon et al., 2014; Seabrook et al., 2017).

### *Function*

Perhaps the most fundamental functional goal of the retina is to efficiently encode visual information (Atick, 1992; Atick & Redlich, 1990, 1992; Balasubramanian & Sterling, 2009; H. B. Barlow, 1961, 1989; H. B. Barlow & Földiák, 1989; Meister & Berry, 1999; Ocko et al., 2018; Simoncelli, 2003; Sterling & Laughlin, 2015; van Hateren, 1992). This theory, known as efficient coding, was originally proposed by Fred Attneave and Horace Barlow (Attneave, 1954; H. B. Barlow, 1961, 1989). Visual inputs at the photoreceptor array are highly correlated in space and time due to the statistical structure of natural scenes. From an information-theoretic perspective, the array therefore encodes visual information very inefficiently (see Appendix on *Efficient Coding*). Information inefficiencies lead to wasted space, energy, and resources (Balasubramanian & Sterling, 2009; Sterling & Laughlin, 2015). In order to improve efficiency, the retina performs filtering operations that ultimately remove many of those input correlations. This establishes a ganglion cell neural code with much less redundancy than that of the photoreceptor array (Atick & Redlich, 1992; Meister & Berry, 1999). Mathematically, the computational goal of reducing redundancy and efficiently encoding visual information is equivalent to building an accurate autoencoder (Appendix) (Atick & Redlich, 1992; Ocko et al., 2018; Olshausen & Field, 1996). The retina might therefore be understood as transmitting the latent variables of an unsupervised model of the visual environment.

In order to understand the significance of this interpretation of retinal function, and the extent to which efficient coding provides an adequate explanation of the data, we will briefly explore some of the characteristics of light-evoked neural activity in the retina. Neurons along the visual pathways are often characterized by their responses to light, or receptive fields. The receptive field provides a summary of a neuron's behavior and indicates what spatiotemporal pattern of light most effectively triggers firing in that cell (Hughes et al., 2006; D L Ringach, 2004; T N Wiesel, 1968). The receptive field is always depicted in the relative coordinates of a *retinotopic map*. The fixed physical location of photoreceptors on the retina establishes this map and the subsequent wiring of neurons along the visual pathways is tightly constrained to obey the original coordinate system. This means that neighboring neurons in later visual areas receive inputs emanating from neighboring photoreceptors.

Beginning with photoreceptors, these respond to increments and decrements of light shone directly onto their outer segments, which contain the photo-sensitive pigment rhodopsin (Kolb, 2003). In the dark, photoreceptors release a steady stream of glutamate that activates the class of OFF bipolar cells by way of ionotropic glutamate receptors. OFF bipolar cells therefore respond to darkness and light decrements with a depolarization, while ON bipolar cells depolarize to light increments by way of a sign inversion. The latter class expresses hyperpolarizing metabotropic glutamate receptors, so the steady stream of photoreceptor glutamate in darkness inhibits them (Masland, 2001). Bipolar cells are further subdivided into transient and sustained classes, acting as high-pass and low-pass filters, respectively

(Kandel et al., 2014; Masland, 2001). In the primate fovea, bipolar cells are in one-to-one correspondence with photoreceptors, while at larger retinal eccentricities, bipolar cells begin to integrate information from many photoreceptors (Kolb, 2003).

Retinal ganglion cells (RGCs) are the output cells of the retina. These receive direct input from bipolar cells and are divided into subtypes based on their functional responses to light, i.e., their receptive fields. As with bipolar cells, there are ON and OFF RGCs, along with a class of ON-OFF RGCs. The majority of RGCs in the primate retina respond to light with a canonical center-surround, or difference of Gaussians, profile (Meister & Berry, 1999). ON RGCs are activated by light shone on their corresponding photoreceptor(s) (receptive field center), while light shone onto neighboring photoreceptors (surround) inhibits them. The reverse is true of OFF RGCs, which are activated when light shone on their center turns off. Such behavior is often described as excitatory center and antagonistic, or inhibitory, surround, with surround inhibition mediated via horizontal and amacrine interneurons. ON-OFF RGCs respond to both increments and decrements of light. In the temporal domain, a related effect occurs, such that the neurons respond transiently to changes in luminance, decaying to baseline within about 200-300ms and remaining quiet in static scenes. RGCs thereby compute the difference between center and surround, and between present and past.

From the section on *Unsupervised Learning*, we immediately recognize that these computations implement a predictive coder in space and time (Elias, 1955; Hosoya et al.,

2005; Srinivasan et al., 1982). Due to the statistical structure of natural scenes, if both center and surround are stimulated by light of comparable intensity, then the region of space corresponding to the center is predictable from the surround and therefore redundant. This is also true in the time domain: if the present is predictable from the recent past then it is redundant. RGCs only send spikes to the rest of the central nervous system when locally-unpredictable information is present in their receptive field centers. In a strict information-theoretic sense, as proven in the sections on *Predictive Coders* and *Efficient Coding*, the retina therefore compresses information and reduces redundancy. Furthermore, a plausible biological implementation is provided by Anti-Hebbian plasticity at the inhibitory amacrine-to-RGC synapse.

Retinal computation exhibits further complexity through parallelization and adaptation. In primates, for example, ON and OFF groups are subdivided into midget and parasol cells. Midget cells are characterized by their dense packing on the retina, small receptive fields, and low-pass temporal filtering (derived from the sustained class of bipolar cell). Alternatively, parasol cells are more sparsely packed, have large receptive fields, and act as high-pass filters (Dacey, 1994; E. Kaplan & Shapley, 1986; Meister & Berry, 1999; Ocko et al., 2018). These are also referred to as brisk-sustained (or X or beta) and brisk-transient (or Y or alpha), respectively. All four types tile the retina and serve as parallel information streams, though in the mouse there are no ON-transient RGCs. There are also different streams for color and motion information, including classes of direction-selective ganglion cells (often referred to as W cells in the cat retina) (Mauss, Vlasits, Borst, &

Feller, 2017; Meister & Berry, 1999; Seabrook et al., 2017). With respect to adaptation, RGCs adapt their signaling to dramatic changes in overall luminance, thus adjusting their dynamic range on the fly to match the environment (Meister & Berry, 1999; Sterling & Laughlin, 2015; Weber, Krishnamurthy, & Fairhall, 2019; Webster, 2011). This computation cannot be described mathematically by the linear center-surround model, but it is predicted by efficient coding. RGCs also show an approximate invariance to contrast, again not predicted by the center-surround model.

Despite a clear role for efficient coding in the retina, this additional computational complexity and the overall variety of computation suggests that other principles may be relevant. For example, some researchers suggest crucial roles for species-specific computation (Yilmaz & Meister, 2013). Others argue that the retina performs feature extraction computations that are explicitly useful for subsequent species-specific processing, such as object recognition and motion detection (Schwartz, 2021). A compromise might be had between these perspectives by taking the approach of Simon Laughlin, Vijay Balasubramanian, and Peter Sterling, who argue “Given the information required for behavior, the retina minimizes its computational cost” (Balasubramanian & Sterling, 2009; Sterling & Laughlin, 2015). Their hypothesis is a somewhat more inclusive statement of efficient coding, making more explicit the trade-off between information and resources, whether those be energy or space or time (S. B. Laughlin, 2001; Niven & Laughlin, 2008). For example, the energy required to transmit on average  $N$  spikes per second is approximately quadratic in the number of spikes (Balasubramanian & Sterling,

2009; Simon B. Laughlin et al., 1998; Perge, Koch, Miller, Sterling, & Balasubramanian, 2009). Meanwhile, the amount of information transmitted by those spikes is proportional to  $\log_2(N)$  in the best possible scenario (Simon B. Laughlin et al., 1998; Strong, De Ruyter Van Steveninck, Bialek, & Koberle, 1998). Therefore, the energetic cost per bit of information shows a dramatic law of diminishing returns, such that each additional bit of information requires correspondingly more energy. This can be so severe that it makes sense to split one high firing rate channel into multiple parallel channels in order to reduce total energy consumption (Balasubramanian & Sterling, 2009; Simon B. Laughlin et al., 1998). Overall, this suggests that even when highly species-specific computations are performed downstream of the retina, it still behooves the system to efficiently allocate limited resources.

In conclusion, efficient coding explains a surprising number of empirical findings. It accurately predicts the observed spatial and temporal receptive field properties of retinal ganglion cells (Atick, Li, & Redlich, 1992; Atick & Redlich, 1990, 1992; Balasubramanian & Sterling, 2009; Doi et al., 2012; Meister & Berry, 1999; Ocko et al., 2018), including their extent in visual space, their spacing within the lattice of other RGCs, center-surround spatial structure, temporal band-pass filtering, and opponent color coding. It also correctly predicts the existence of different subtypes of ganglion cell, including ON/OFF and midget/parasol, from the statistical structure of natural scenes (Gjorgjieva, Sompolinsky, & Meister, 2014; Ocko et al., 2018). It provides precise quantitative predictions for the contrast sensitivity functions of monkey and human observers (Atick & Redlich, 1992;

Buchsbaum & Gottschalk, 1983; Russell L De Valois, Morgan, & Snodderly, 1974; Van Hateren, 1993). It therefore seems reasonable to believe that many other properties of the retina could ultimately be explained when more complicated natural image statistics and nonlinear retina models are included in the derivations.

## **Thalamus**

### *Structure*

The thalamus is often described as a relay between the sensory periphery and cortex. Leaving the retina, most of the axons in the optic nerve terminate in the thalamus. It is composed of a variety of sub-regions, most of which are specialized for a particular sensory modality, transmitting information from RGCs, and related cell types for other modalities, to cortex. Thalamus is surrounded by a region known as thalamic reticular nucleus (TRN) that receives input from cortex and sends inhibitory projections to the rest of thalamus (Crandall, Cruikshank, & Connors, 2015). There are also higher-order regions that receive no input from the sensory periphery, but rather communicate reciprocally with cortical and sub-cortical regions. The presence of these regions and significant direct feedback from cortex suggests a role for thalamus beyond simple relay. For example, cortical feedback might provide a gain control or gating mechanism in the thalamus, selectively suppressing or enhancing certain inputs (Crandall et al., 2015). Alternatively, feedback could alter the filtering properties of thalamus, making dynamic adjustments dependent on the signal-to-noise ratio of the inputs (D. Dong & Atick, 1995).

In the visual domain, the primary relay region is the dorsal lateral geniculate nucleus (dLGN). dLGN relay neurons receive direct inputs from RGCs, with between 1-20 different RGCs from one eye contacting each relay neuron (Chinfei Chen & Regehr, 2000; Hammer, Monavarfeshani, Lemon, Su, & Fox, 2015; Rompani et al., 2017). These comprise about 90% of thalamic neurons, while the other 10% are interneurons that also receive direct input from RGCs and then inhibit relay neurons. Inputs from each eye are segregated within dLGN, with a majority emerging from the contralateral eye. These inputs are also organized in a precise retinotopic manner. In primates, different types of RGC innervate specific layers of dLGN. Midget RGCs synapse onto P cells in parvocellular layers while parasol RGCs synapse onto M cells in magnocellular layers (Kandel et al., 2014). Between these are also koniocellular layers that receive input from a heterogeneous set of RGCs (Hendry & Reid, 2000).

In mice, there are no clear cytoarchitectonic layers, but there still exists anatomical segregation of the RGC input. dLGN is divided into *shell* and *core* regions, with different types of RGC innervating neurons in each region. In addition, inputs are retinotopically arranged in both shell and core. Primary inputs to V1 pass through the core, while presumptive modulatory inputs, including those from direction-selective RGCs, pass through the shell (Bickford, Zhou, Krahe, Govindaiah, & Guido, 2015; Kerschensteiner & Guido, 2017; Seabrook et al., 2017). The core contains so-called X-like and Y-like relay neurons, which likely receive inputs from “traditional” X (beta) and Y (alpha) RGCs (Bickford et al., 2015; Krahe, El-Danaf, Dilger, Henderson, & Guido, 2011). X- and Y-

like relay neurons receive *driving* inputs from relatively few RGCs, with large synapses that make contacts close to the soma. Feedback from TRN and cortex send *modulatory* inputs, with small synapses targeting distal dendrites (Bickford et al., 2015; Seabrook et al., 2017). The shell, by comparison, contains W-like relay neurons, receiving driving inputs from both direction-selective RGCs (W cells in the cat) and projection neurons in the superior colliculus (Bickford et al., 2015).

### *Function*

dLGN relay neurons largely inherit their functional properties from the RGCs that innervate them (DeAngelis, Ohzawa, & Freeman, 1995; Kandel et al., 2014). For this reason, receptive fields of dLGN neurons greatly resemble those found in RGCs. Neurons in dLGN core show center-surround spatial organization and biphasic temporal filters, while neurons in the shell show direction- and orientation-selective responses (Kerschensteiner & Guido, 2017; Meister & Berry, 1999). Despite these inherited functional properties, additional image processing occurs in dLGN, especially in the temporal domain.

The basic circuit design, with interneurons receiving inputs from RGCs and then inhibiting relay neurons that receive similar inputs, is suggestive of a predictive coding network (see Figure 2). Experimental evidence suggests that dLGN performs a temporal whitening/decorrelation operation on its inputs consistent with this hypothesis (Dan et al., 1996). The retina performs an imperfect predictive coding operation, such that RGC

outputs are largely spatially decorrelated but only partially temporally decorrelated. dLGN seems to continue the process of temporal decorrelation. A theoretical analysis akin to that elaborated in the Appendix on Efficient Coding reveals the need for two dLGN cell classes: lagged and non-lagged (D. Dong & Atick, 1995). Each class has a center-surround spatial filter and biphasic temporal filter, comparable to those found in retina. However, the temporal filter in the lagged case is shifted in time such that lagged neurons respond to a given stimulus with a delay relative to the non-lagged class. These lagged neurons do not exist in the retina, but there is considerable experimental evidence that they do exist in dLGN (DeAngelis et al., 1995; Hartveit, 1992; Saul & Humphrey, 1990).

Lagged and non-lagged cells seem to arise independent of class, occurring in both X-like and Y-like relay neurons (Saul & Humphrey, 1990). As mentioned above, in dLGN core X-like cells likely receive inputs from brisk-sustained RGCs while Y-like cells receive inputs from brisk-transient RGCs. These inputs dictate the response properties of the relay neurons, which match those of their respective input RGCs (Kerschensteiner & Guido, 2017). However, the match is most prominent in the spatial domain, while the lagged and non-lagged response types are indicative of additional intrageniculate temporal processing. In addition to these cell types, W-like relay cells are prominent in the shell. These show both direction- and orientation-selective responses to a greater extent than found in primates and cats.

Overall, the anatomical structure and functional properties of dLGN support its role as a relay from retina to cortex. Relatively few RGCs contact each relay neuron with large synapses placed close to the soma, ideally positioned to strongly influence relay output. This is verified functionally as dLGN relay neurons tend to have receptive field properties that closely match their input RGCs. However, some dLGN neurons integrate information from many types of RGC and have complex response properties not found in the retina. The system also contributes to efficient coding by whitening its inputs in the temporal domain through lagged and non-lagged cell classes not present in the retina. These findings, and anatomical evidence for significant inputs from TRN and cortex, argue for a more nuanced perspective.

## **Primary Visual Cortex**

### *Structure*

In the retina, all of the primary cell types have been identified, along with their connectivity to other cells, their synaptic structures, basic neurotransmitter types, and functional roles within the circuit. Knowledge of the retina is exquisitely detailed, often down to nanometer precision. Our knowledge of V1, by comparison, is limited. This is largely due to the extreme complexity of cortical circuits. Cortical pyramidal neurons often have thousands (sometimes tens of thousands) of synapses (Spruston, 2008). While the retina has a largely feedforward architecture (photoreceptors to bipolar cells to RGCs), the cortex has lateral recurrent connectivity and intense reciprocal feedback across the brain (K. D. Harris & Shepherd, 2015). In almost all cortical regions, there is no notion of the precise number of

different cell types nor even a consensus regarding how to define cell types (DeFelipe et al., 2013). Correspondingly little is known about the great diversity of neurotransmitters these cells likely exchange, nor the precise effects of different neuromodulators such as acetylcholine.

In humans and monkeys, V1 resides in the occipital lobe. It is area 17 in Brodmann's designation. In mice, V1 is in a near identical location relative to the bones and sutures of the skull, resting on the dorsal surface of the brain, lateral to lambda and just rostral to the lambdoid suture. Lambda designates the meeting point of the sagittal and lambdoid sutures, which divide the two parietal bones along the midline and the parietal from occipital bone, respectively. Visual information travels from the retina to thalamus through the optic nerve, and then from thalamus to V1 by way of the optic radiations.

Almost all neocortical areas, including V1, have a six-layer structure (Ángel García-Cabezas, Zikopoulos, & Barbas, 2019; Barbas, 2015; K. D. Harris & Shepherd, 2015).

Layer 1- The most superficial layer is sparsely occupied by interneurons, glia, and dendrites, and densely packed with axons. Most axons emerge distally from regions sending information to V1. For example, W-like dLGN relay neurons send inputs to layer 1 of V1 (Bickford et al., 2015; Seabrook et al., 2017), while other cortical and sub-cortical regions also project axons to layer 1. The dendrites of excitatory pyramidal neurons in deeper layers (2-5) extend into layer 1 to receive these inputs.

Layer 2/3- Layers 2 and 3 are often lumped together, and these contain many long-range excitatory pyramidal neurons that preferentially target other cortical regions. Visual cortical regions are arranged hierarchically and layer 2/3 neurons are thought to transmit information up the hierarchy (Ángel García-Cabezas et al., 2019; Felleman & Van Essen, 1991; J. A. Harris et al., 2019). In the primate visual system, layer 2/3 projection neurons mostly target secondary visual cortex (V2), but also V3, V4, and the middle temporal visual area (MT). In mice, layer 2/3 neurons project to a variety of higher-order visual regions, such as latero-medial area (LM) and rostro-lateral area (RL), and these reciprocally send feedback to V1 through layer 1 (Seabrook et al., 2017; Siegle et al., 2021). Many inhibitory interneuron types also reside in layer 2/3, such as vasoactive-intestinal-peptide (VIP) and somatostatin (SOM) expressing interneurons (Rudy, Fishell, Lee, & Hjerling-Leffler, 2011).

Layer 4- Layer 4 is the primary thalamic input layer, with thalamic relay neurons targeting excitatory pyramidal and spiny stellate neurons and also inhibitory interneurons, mostly of the parvalbumin (PV) class. X-like and Y-like dLGN relay neurons target layers 4 and 5 of V1. These thalamorecipient neurons in turn synapse onto local interneurons, neighboring layer 4 cells, and the pyramidal projection neurons of layer 2/3. In primates, layer 4 is further divided into sub-layers, each of which receives input from a different thalamic neuron type, but it is unclear whether this distinction applies to mice.

Layer 5- Layer 5 is a cortical and sub-cortical feedback layer, sending long-range excitatory projections to regions lower in the hierarchy. V1 layer 5 excitatory neurons project to superior colliculus, for example (Dhande et al., 2015).

Layer 6- The deepest layer seems to have at least two distinct excitatory cell populations, one which sends feedback to TRN and dLGN and another that projects to other cortical regions (Briggs, 2010; Seabrook et al., 2017).

Excitatory pyramidal neurons are thought to integrate and transmit information both locally and to other regions, while interneurons (mostly inhibitory, expressing gamma-Aminobutyric acid [GABA]) sculpt pyramidal neuron activity. Pyramidal neurons have very complex dendritic structures and many appear to act as coincidence detectors of their convergent inputs. Inputs from different cell types and regions are segregated to different dendritic domains (Spruston, 2008). Excitatory feedback projections and SOM interneurons, such as Martinotti cells, target apical dendrites (Kepecs & Fishell, 2014). PV interneurons, like the fast-spiking basket cells, make synaptic contacts close to the soma (Hu, Gan, & Jonas, 2014). This segregation of inputs, along with evidence for nonlinear dendritic processing, suggests individual pyramidal neurons are capable of complex computations beyond linear integration and thresholding (Gidon et al., 2020; Spruston, 2008).

In recent years, researchers have discovered a conserved circuit motif across cortical areas that includes layer 2/3/4/5 pyramidal neurons and PV, SOM, and VIP interneurons (Pfeffer, Xue, He, Josh Huang, & Scanziani, 2013; Rudy et al., 2011). The basic circuit consists of pyramidal-to-interneuron excitation, PV-to-SOM and PV-to-VIP and PV-to-pyramidal inhibition, SOM-to-PV and SOM-to-VIP and SOM-to-pyramidal inhibition, and VIP-to-

SOM and VIP-to-PV inhibition. Each class therefore has mechanisms to inhibit other classes, and the VIP class has a particularly unique disinhibitory effect on pyramidal neurons (Karnani, Jackson, Ayzenshtat, Hamzehei Sichani, et al., 2016). Interneuron classes have also been shown to exhibit within-class mutual co-activity, created by direct electrical coupling and disinhibition of opposing interneuron classes (Karnani, Jackson, Ayzenshtat, Tucciarone, et al., 2016).

### *Function*

The extreme structural complexity of cortex is accompanied by a corresponding increase in the variability of neural activity. Relatively simple mathematical models can achieve high accuracy in predicting how RGCs respond to visual inputs (Pillow et al., 2008). RGCs also exhibit extremely low trial-to-trial variability when exposed to the same visual stimuli (Berry, Warland, & Meister, 1997). By comparison, cortical neurons in awake-behaving animals exhibit very high variability under typical conditions (Stringer, Pachitariu, Steinmetz, Reddy, et al., 2019; T Goris, Anthony Movshon, & Simoncelli, 2014) and mathematical models of their activity make comparatively poor predictions (Cadena et al., 2019).

The most basic model of an excitatory V1 neuron's activity is known as the LNP model, for linear-nonlinear-Poisson (Hughes et al., 2006; I. M. Park & Pillow, 2011; Rust & Schwartz, 2005). For a single neuron,

$$N \sim \text{Poisson}(\lambda(\mathbf{x}))$$

$$\lambda(\mathbf{x}) \triangleq \exp(b + \mathbf{w}^T \mathbf{x})$$

Here,  $N$  is the number of spikes the neuron emits on a given trial (usually summed over a time window from about 40-100ms after the onset of a visual stimulus),  $\mathbf{x}$  is a vectorized version of the image shown to the animal,  $\mathbf{w}$  is known as the neuron's linear receptive field, and  $b$  is the baseline firing rate. The relationship between the visual stimulus,  $\mathbf{x}$ , and the neuron's firing rate,  $\lambda$ , is a linear function,  $\mathbf{w}^T \mathbf{x}$ , followed by a nonlinearity, in this case the exponential function. This is a generalized linear model (Truccolo, 2004), and note that the exponential nonlinearity can be exchanged for other functions, such as the sigmoid.

This version of the LNP model captures the activity of so-called simple cells, which were first described by Torsten Wiesel and David Hubel in the cat (DeAngelis et al., 1995; D H Hubel & Wiesel, 1959; David H Hubel & Wiesel, 1962). These cells are most prevalent in layer 4 of V1. They respond to oriented bars of light and have sub-regions within the receptive field that preferentially respond to either increments or decrements of light intensity, but never both. These are called excitatory and inhibitory zones, and the overall receptive field structure resembles a Gabor filter with alternating dark and light zones. Hubel and Wiesel did not use the LNP model, but it provides an accurate description of what they observed. Looking at the model, we see that the intensity of the visual stimulus at each location in space is encoded in the elements of the vector  $\mathbf{x}$ , and each element is simply multiplied by a corresponding element in  $\mathbf{w}$  to give the firing rate. If a given element of  $\mathbf{w}$  is positive, then increments in light intensity will tend to increase neuronal firing while decrements will tend to decrease it. Thus, increments and decrements of light

have opposing effects on neural activity, and this behavior is often referred to as linear spatial summation.

Hubel and Wiesel also observed cells that they described as complex (R L De Valois, Yund, & Hepler, 1982; David H Hubel & Wiesel, 1962). Complex cells show spatial invariance, such that they do not have identifiable excitatory and inhibitory zones. Light increments *and* decrements within a given spatial region can increase the cell's firing. A simple adjustment to the basic LNP model is able to capture this behavior:

$$N \sim \text{Poisson}(\lambda(\mathbf{x}))$$

$$\lambda(\mathbf{x}) \triangleq \exp\left(b + \frac{1}{2} \mathbf{x}^T \mathbf{C} \mathbf{x}\right)$$

This is a spike-triggered covariance model (I. M. Park & Pillow, 2011). Now, a neuron's receptive field(s) are given by the eigenvectors of the matrix  $\mathbf{C}$ . Different eigenvectors can have inhibitory and excitatory zones that overlap, thus establishing the complex behavior. These cells were found more often in the superficial and deep layers of V1, avoiding layer 4.

When  $\mathbf{x}$  is taken to be a brief video segment, instead of a static image, LNP models learn a full spatiotemporal receptive field. These adequately capture many properties of neurons in V1, such as position selectivity (neurons only respond to stimuli in certain regions of visual space), orientation selectivity (neurons are responsive to edges with a precise and limited range of orientations), temporal frequency selectivity, and direction selectivity (for stimuli moving only in one direction) (DeAngelis et al., 1995; C. M. Niell & Stryker, 2008). In macaque monkeys, the vast majority of excitatory cells exhibit orientation selectivity,

while about 25-35% show direction selectivity (R L De Valois et al., 1982). Despite much lower visual acuity, mouse V1 has comparable proportions of orientation and direction selective cells (C. M. Niell & Stryker, 2008). Inhibitory neurons, especially in the mouse, tend to be much less selective, likely due to a convergence of excitatory inputs with unmatched tuning properties (Hofer et al., 2011; C. M. Niell & Stryker, 2008).

LNP models are often referred to as phenomenological because they capture the phenomena of neural firing in response to visual stimuli. There are also mechanistic models that attempt to explain *how* these responses arise in actual neural circuits, and normative models that offer explanations as to *why* these types of responses might be present. Mechanistically, Hubel and Wiesel proposed a straightforward neural architecture to model the behavior of V1 simple cells. In their model, small groups of dLGN neurons innervate a given V1 neuron. If dLGN relays have center-surround receptive fields aligned in partially-overlapping regions of visual space, then a V1 cell integrating those inputs could exhibit an elongated receptive field, inhibitory and excitatory zones, and orientation selectivity. This model seems to be accurate, at least in cats and mice (Lien & Scanziani, 2013; Reid & Alonso, 1995). One can also imagine a complex cell emerging from the integration of several simple cells, which would explain the prevalence of complex cells in layer 2/3, presumably receiving inputs from layer 4 simple cells.

Regarding normative models, a number of researchers have offered efficient coding as a plausible explanation of V1 neural activity (Atick & Redlich, 1993; Bell & Sejnowski,

1997; Olshausen & Field, 1996; Simoncelli, 2003; Van Vreeswijk, 2001; Zylberberg et al., 2011) (see also Appendix on *Efficient Coding*). As proven in the Appendix, to efficiently encode high-dimensional information one can build an autoencoder. With that in mind, many researchers have shown that the receptive fields of artificial neurons from autoencoder networks trained on whitened natural scenes come to resemble those found in V1 simple cells (Atick & Redlich, 1993; Olshausen & Field, 1996; Zylberberg et al., 2011). These results agree with the hypothesis explored here, that the visual system is an unsupervised learning device. If the retina and dLGN have largely decorrelated/whitened the visual input in space and time, then V1 would serve to remove higher-order correlations and further compress information. Under this hypothesis, edges and Gabor filters must be examples of higher-order correlations in the distribution of natural scenes (Bell & Sejnowski, 1997).

To summarize thus far, V1 neurons exhibit increasing complexity in their responses to visual inputs relative to retinal and thalamic neurons. RGCs and dLGN relay neurons mostly show center-surround receptive fields, while V1 receptive fields are often elongated Gabor filters (simple cells) or not captured at all by basic linear receptive field models (complex cells). A variety of unsupervised autoencoder models capture the behavior of simple cells (Olshausen & Field, 1996, 1997) and suggest that Barlow's efficient coding hypothesis may apply beyond the retina and dLGN. It is also worth noting that more complex unsupervised learning models accurately predict behavior in V1 and across the visual system (Cadena et al., 2019; Zhuang et al., 2021).

Despite the success of the simple/complex distinction and the LNP model in predicting the activity of cortical neurons to arbitrary stimuli, these fall well short of comparable models for RGCs and relay neurons. For one, they do not account for phenomena such as light and contrast adaptation, known to occur in the retina. Nor can they account for the behavior of cells called hypercomplex by Hubel and Wiesel, and now referred to as end-stopped cells (Gilbert, 1977). These cells respond to oriented bars of light flashed on their receptive fields, like simple cells, but prefer bars of a specific length. Responses increase up to the preferred length and then begin to decrease. This type of behavior could be explained by the linear model, if an elongated and oriented excitatory zone were completely surrounded by inhibitory zones. However, the responses of these cells have been shown to exhibit nonlinear spatial summation, so the LNP model is not sufficient to explain the end-stopped behavior.

There is also evidence for *divisive normalization* along the visual pathways and especially within V1 (Carandini & Heeger, 1994, 2012). Divisive normalization explains a number of V1 response properties not captured by traditional models. We will describe the computational principle by way of a mathematical model. Take the simplified example of a set of orientation tuned cells, i.e., each cell in a population of  $M$  prefers a different orientation from  $0 - 180^\circ$ . If the tuning of a given neuron to orientation is given by  $f_m(\theta)$ , where  $f$  is typically some circular-Gaussian-like function of the orientation, then an LNP-like model would be:

$$N_m \sim \text{Poisson}(f_m(\theta))$$

The basic normalization model is, instead:

$$N_m \sim \text{Poisson}(R_m(\theta))$$

$$R_m(\theta) = \gamma \frac{f_m(\theta)^d}{\sigma + \sum_{m=1}^M f_m(\theta)^d}$$

$\gamma$  is a scaling factor, the summation is over a local pool of neurons, and  $\sigma$  and  $d$  are free parameters that control the overall shape of the tuning curves.

Divisive normalization is a very accurate model for light and contrast adaptation in the retina and for a variety of nonlinear behaviors observed in V1 (Carandini & Heeger, 2012). One example of the latter is cross-orientation suppression. When a grating stimulus with an orientation orthogonal to a V1 neuron's preferred orientation is presented alone, the neuron does not respond. However, when an orthogonal orientation is presented together with the preferred orientation, firing decreases as a function of the orthogonal grating's contrast. The LNP model cannot capture this behavior, but the divisive normalization model can (the denominator increases when other neurons are stimulated). Though the precise mechanism controlling this behavior is unknown, most models posit some form of lateral inhibition (Carandini & Heeger, 2012). Similar models also describe other nonlinear behaviors such as surround suppression and contrast-invariant orientation tuning (Carandini & Heeger, 2012), but see Priebe & Ferster 2012 for alternative explanations of these (Priebe & Ferster, 2012). In addition, theoretical work shows that divisive normalization may be a requisite computation for efficiently encoding information with power-law-like distributions (Bucher & Brandenburger, 2020; Lyu, 2010).

Returning to mechanistic models of V1, the picture that emerges is one in which structure dictates function. dLGN precisely segregates inputs from each eye and from different RGC types, and this pattern of connectivity controls the functional properties of dLGN relay neurons. In a similar manner, cortex segregates inputs. dLGN inputs to V1 are arranged retinotopically across the cortical surface; their precise arrangement in visual space governs the orientation selectivity of thalamorecipient neurons; and, in macaques, ferrets, and cats (likely all predatory mammals), inputs from each eye are segregated into *ocular dominance columns* (David H Hubel & Wiesel, 1969; Law, Zaks, & Stryker, 1988; Christopher M. Niell, 2015). Within a column, individual layer 4 pyramidal neurons receive thalamic inputs predominantly emanating from one eye. Columns preferring each eye then alternate across V1, creating a mosaic (note, V1 is the first visual region to integrate inputs from both eyes). Similar columns are found for orientation selectivity. In mice, the situation is somewhat different: inputs are still heavily segregated, but there are no ocular dominance or orientation columns (Seabrook et al., 2017). Instead, mouse V1 has a distinct binocular region that receives inputs from both eyes, though the contralateral eye is favored, while orientation tuning is random across the cortical surface, a “salt-and-pepper” organization (Espinosa & Stryker, 2012a; Ohki, Chung, Ch, Kara, & Clay Reid, 2005; Dario L. Ringach et al., 2016).

Two recent articles reveal the extent to which the precise structural organization of the cortex governs its function. One study found that cortical direction selectivity in V1 layer

4 was not inherited from thalamic direction-selective relay neurons, but rather computed *de novo* by way of retinotopically precise wiring (Lien & Scanziani, 2018). Spatially offset transient and sustained relay neurons were found to synapse onto individual layer 4 neurons. As a moving stimulus crossed through the thalamic receptive fields, it activated each neuron class sequentially. A stimulus moving first into the sustained cell's receptive field and then into the transient cell's was sufficient to evoke firing in the cortical neuron. In the opposite direction, the transient cell would fire quickly and its excitatory post-synaptic potential would decay prior to the arrival of the sustained cell's inputs, preventing firing. Thus, the cortical neuron would only fire for a stimulus moving in the sustained-to-transient direction, but not in reverse. Another study showed that direction selectivity in layer 2/3 of V1 was again computed *de novo*, rather than inherited from layer 4 neurons (Rossi, Harris, & Carandini, 2020). In this case, the precise spatial organization of pre-synaptic inputs within cortex endowed layer 2/3 neurons with direction selectivity. Inputs were again spatially offset, with excitatory inputs tending to come from one location in space and inhibitory inputs tending to come from a nearby location. A moving stimulus in the preferred direction would first activate the excitatory input group and then the inhibitory. In the opposite direction, the inhibitory group would be activated first, thus preventing the excitatory group from triggering firing in the post-synaptic neuron. Both results reveal the incredible wiring precision of V1 circuitry and indicate the extent to which structure informs function.

To conclude, we consider potential future directions for vision research. Recent work using convolutional neural networks (CNNs) suggest that shallow neural networks approximate V1 activity (Burgid et al., 2021; Cadena et al., 2019; Zhuang et al., 2021). These phenomenological models capture a wide variety of the non-canonical (and nonlinear) response properties of V1 neurons. Often criticized for being black boxes, CNNs provide a metric against which other models can be compared. They also have potential for dissection, interpretation, and transformation into mechanistic models; a CNN, though complex, is much less so than the brain (Tanaka et al., 2019). The success of these models may also argue for a new approach to studying cortical computation. Traditional efforts, like those of Hubel and Wiesel, have emphasized the computations performed by cortical neurons as functions of either sensory inputs or motor outputs. In the visual domain, this approach has led to a general consensus among researchers that the goal of visual neuroscience is to predict neural activity to arbitrary and naturalistic visual stimuli using mathematical models that are human interpretable (Carandini et al., 2005). An alternative approach is to attempt to understand the learning rules that operate throughout the circuit. With a sophisticated understanding of the rules, one may be able to instantiate them into a deep (perhaps recurrent) neural network and accurately model the brain. The actual computations performed within the network would almost certainly be uninterpretable to the human mind, but we would still have proven sufficient understanding to create a model that accurately represents the biological system. With that in mind, the next section reviews experiments on neural plasticity and learning in the visual system.

### Visual Cortical Plasticity

Since the discovery of simple and complex cells by Hubel and Wiesel in the 1960s, the visual cortex has been one of the most thoroughly studied regions of the neocortex (Carandini et al., 2005; Espinosa & Stryker, 2012a; Christopher M. Niell, 2015). Experiments span a wide range of species and an incredible diversity of techniques. One common thread has been to use V1 as a model system for the study of cortical development and plasticity. Because V1 neurons exhibit selectivity to certain stimuli over others, and their selectivity in adulthood is well characterized, an obvious question is how selectivity emerges during development (Espinosa & Stryker, 2012a; Hensch, 2005). Some of the most salient functional properties of V1 neurons are their selectivity for stimulus position in visual space (retinotopy), for orientation, and for direction. Are these hard-wired in the genetic code? Are they learned through visual experience? How do they change under developmental perturbations? Furthermore, researchers have asked how visual experience can alter these properties in adulthood, either through task-based perceptual learning or through unsupervised exposure to novel visual environments (Frenkel et al., 2006; Karmarkar & Dan, 2006; Sato & Stryker, 2008).

In order to better understand some of the key findings and outstanding problems in the visual cortical plasticity field, this section will review two forms of plasticity: developmental ocular dominance plasticity through monocular deprivation and *sequence learning*. The former is perhaps the single most thoroughly studied form of experience-dependent plasticity in all of neuroscience. The latter is a novel form of unsupervised

perceptual learning that has been shown to occur in mice through passive, repetitive visual experience (Gavornik & Bear, 2014). In the process, we will highlight the relevance of efficient coding to cortical plasticity and attempt to grapple with the extreme degree of biological complexity that most mathematical models of plasticity fail to capture.

### *Ocular Dominance Plasticity & Monocular Deprivation*

As described before, the cortical surface forms a retinotopic map of visual space, and in cats, ferrets, and primates, orientation selectivity is clustered across V1 in a mosaic. In addition, the retinotopic map is divided into ocular dominance columns within which inputs from a given eye predominate. The presence of these ocular dominance columns afforded Hubel and Wiesel a model system to study cortical development and plasticity (David H. Hubel & Wiesel, 1964; David H Hubel & Wiesel, 1965; David H Hubel, Wiesel, & LeVay, 1977; Torsten N Wiesel & Hubel, 1963b). Inspired by the prevalence of strabismus and amblyopia in humans, they either surgically removed or sutured one eye during early development in kittens and macaque monkeys. After a period of this *monocular deprivation* (MD), they observed profound changes to the structure and function of cortex in the region contralateral to the deprived eye. In particular, they observed *ocular dominance plasticity* (ODP), characterized by a shift in ocular dominance toward the non-deprived eye: the entire ocular dominance mosaic was dramatically altered in V1, but they found effectively no change in dLGN (Torsten N Wiesel & Hubel, 1963a). Subsequent research has revealed similar effects in many other species, including ferrets and mice (Gordon & Stryker, 1996; Issa, Trachtenberg, Chapman, Zahs, & Stryker, 1999) (mice do

not have ocular dominance columns but a shift in ocular dominance toward the non-deprived eye still occurs in individual neurons within the binocular region of V1).

Since its initial discovery, there has been an explosion of research on ODP, particularly in mice due to the ability to perform genetic manipulations (Espinosa & Stryker, 2012b; Hofer, Mrsic-Flogel, Bonhoeffer, & Hübener, 2006). One of the most interesting early findings was that there exists a *critical period*, during which ODP occurs rapidly and, when prolonged, leads to permanent changes in the structure of V1 (Gordon & Stryker, 1996; Hensch, 2005; Issa et al., 1999; Shatz & Stryker, 1978). In mice, the critical period for ODP occurs from age 21-35 days, peaking at about 28 days (Gordon & Stryker, 1996). Researchers have converged on a general timeline for ODP during the critical period and observed that there exist three essential phases: deprived-eye depression, open-eye potentiation, and recovery. Each phase is characterized by different functional and structural changes, and different plasticity mechanisms.

The first stage of ODP is a period of Hebbian plasticity, during which functional responses to the deprived eye decrease (Cooper & Bear, 2012; Espinosa & Stryker, 2012b). This lasts for about 3 days (Frenkel & Bear, 2004). In this period, responses to the open, non-deprived eye, remain constant. Reduced responses to the deprived eye occur predominantly in layer 4 of V1 (at first) and are consistent with LTD of thalamocortical synapses. Blocking N-methyl-D-aspartate receptors (NMDARs) inhibits this effect (Bear, Kleinschmidt, Gu, & Singer, 1990), as signaling through the NMDARs during MD causes an internalization of

$\alpha$ -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors (AMPARs) (Crozier, Wang, Liu, & Bear, 2007). This reduction in the number of post-synaptic AMPARs appears to be the proximal cause of synaptic depression, though structural plasticity mechanisms may also be important, for example dendritic spine retraction (Sun, Sebastian Espinosa, Hoseini, & Stryker, 2019). There is also evidence that disruption of the immediate early gene *Arc* prevents AMPAR internalization, such that mice lacking the *Arc* gene do not show deprived-eye depression (McCurry et al., 2010). Overall, this ODP stage is very well understood. Remaining questions mostly seem to involve the precise signaling mechanisms through which NMDARs and *Arc* interact to internalize AMPARs, and the relative contributions of functional and structural plasticity.

ODP's second stage starts on day 3 and ends on day 5. During this period, functional responses in layer 4 of V1 potentiate to stimuli targeting the open eye, almost perfectly compensating for the loss of visual responsiveness through the deprived eye (Frenkel & Bear, 2004). The mechanistic underpinnings of this potentiation continue to be a matter of some debate (Cooper & Bear, 2012; Espinosa & Stryker, 2012b). There is evidence for both traditional LTP and homeostatic synaptic scaling. Open-eye potentiation requires NMDARs, consistent with LTP (Cho et al., 2009; Sato & Stryker, 2008). Meanwhile, blocking tumor necrotic factor  $\alpha$  (TNF $\alpha$ ) receptors eliminates open-eye potentiation but has no effect on deprived-eye depression (Kaneko, Stellwagen, Malenka, & Stryker, 2008). Mutation of the TNF $\alpha$  gene in glial cells had been shown to prevent synaptic scaling (Stellwagen & Malenka, 2006), implicating this mechanism as well. It seems that some

combination of the two mechanisms may be responsible for the effect (Mrsic-Flogel et al., 2007). Importantly, mice exposed to MD for 3 days and then placed in complete darkness did not show open-eye potentiation (Blais et al., 2008). This suggests that visual experience through the open eye is necessary for potentiation, though this would not necessarily be expected if homeostatic synaptic scaling were the primary mechanism.

The second stage, and ODP in general, has also been associated with the BCM theory of synaptic plasticity and its sliding modification threshold (Cooper & Bear, 2012). Recall that the threshold governs a shift from LTD to LTP: synaptic inputs that lead to below-threshold postsynaptic activation experience LTD or no change, while inputs driving strong postsynaptic activation undergo LTP. At eye closure, on day 1 of MD, open eye inputs to contralateral V1 are weak and thus insufficient to drive LTP. However, as time passes, reduced activity in the postsynaptic neuron leads to a reduction in the threshold, subsequently allowing LTP to occur for the same weak inputs. There is evidence that such a sliding threshold between LTD and LTP occurs in V1 after light deprivation (Kirkwood, Rioult, & Bear, 1996), so this is a plausible explanation for the second stage of ODP.

The final stage of ODP is recovery. During this phase, the suture on the deprived eye is removed and responses to both eyes return to baseline levels, as long as the duration of MD is brief. Two months of MD during the critical period in kittens leads to irreversible changes (Torsten N Wiesel & Hubel, 1963b), and in mice this timeline may be only a few weeks. Recovery requires neurotrophic growth factor signaling (Kaneko, Hanover,

England, Stryker, & Keck, 2008), suggesting that axon growth and/or dendritic spine remodeling may occur. In particular, inhibition of TrkB kinase activity prevents recovery. Brain-derived neurotrophic factor (BDNF) is a ligand for TrkB receptors and has been implicated in a wide variety of structural plasticity processes (Gómez-Palacio-Schjetnan & Escobar, 2013; Huberman & McAllister, 2002). TrkB kinase inhibition has no effect on the earlier stages of ODP, so all three stages seem to occur through different mechanisms.

Overall, the stages and mechanisms of ODP reveal incredible complexity in the nervous system as it changes with experience. Theoretical models, *a priori*, could never have predicted the timing of these stages nor their molecular mechanisms. And, there are further mechanisms that we have not reviewed, including a crucial role for inhibitory interneurons in regulating ODP and the timing of the critical period (Hensch, 2005; Hensch & Quinlan, 2018; Hooks & Chen, 2020; Sale, Berardi, Spolidoro, Baroncelli, & Maffei, 2010). Still, mathematical models of plasticity and efficient coding provide a scaffold to understand these experimental findings and to ask further questions. Intuitively, it is sensible to redistribute synaptic resources toward the non-deprived eye, and this is generally consistent with an efficient encoding of information. However, we might ask: how does the nervous system *know* that deprived-eye synapses no longer transmit behaviorally-relevant information? These synapses continue to show spontaneous activity during MD, and the difference between random noise and true visual information would be hard to distinguish at the level of individual neurons (in a strict sense, noise is also information). Hebbian plasticity and BCM theory provide a potential explanation, as random noise across many

synapses is unlikely to temporally align and cooperatively activate the postsynaptic cell (Cooper & Bear, 2012). Thus, pre-synaptic activity will fail to reliably correlate with post-synaptic activity, promoting synaptic depression.

An alternative possibility is that pre-synaptic firing no longer correlates with relevant behavioral signals, thus targeting those synapses for degradation. For example, as an animal moves, corollary discharges are sent throughout the brain, potentially signaling predictions regarding what kinds of sensory feedback ought to be produced (Crapse & Sommer, 2008; Guitchounts et al., 2020; Keller, Bonhoeffer, & Hübener, 2012; Stringer, Pachitariu, Steinmetz, Reddy, et al., 2019; Zmarz & Keller, 2016). When the true sensory input does not match the expected input, this may induce plasticity mechanisms that attempt to minimize the mismatch / prediction error (Keller & Mrsic-Flogel, 2018; Rao & Ballard, 1999; Michael W Spratling, 2010). As we explored in the section on *Unsupervised Learning*, generating and verifying predictions creates a form of self-supervised learning expected to aid the system in learning the probability distribution of its inputs (or perhaps the joint distribution of inputs and outputs), which are dramatically altered with MD. Overall, ODP, which occurs in an unsupervised manner, is also consistent with the notion of the visual system as an unsupervised learning device.

### *Sequence Learning*

In the context of this writing, *sequence learning* has a very limited and precise meaning. It is a form of experience-dependent plasticity found to occur in mouse V1 (Gavornik & Bear,

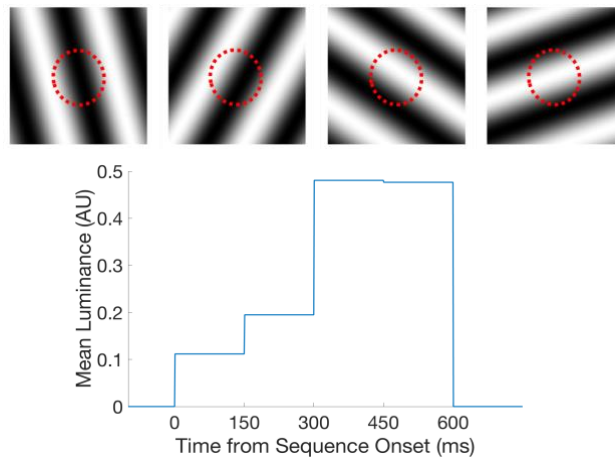
2014). Though it is potentially related to other forms of learning that one might describe in similar terms, and though it likely occurs in other species, we restrict our focus to this particular form of plasticity. Sequence learning occurs when mice sit passively and view repeated presentations of a sequence of high-contrast sinusoidal gratings of differing orientations (Figure 4). It occurs across several days, only requiring a few minutes of passive exposure each day. It is therefore a form of unsupervised learning, though one could potentially argue that there is a reinforcement learning component as it probably feels like punishment to the mice.

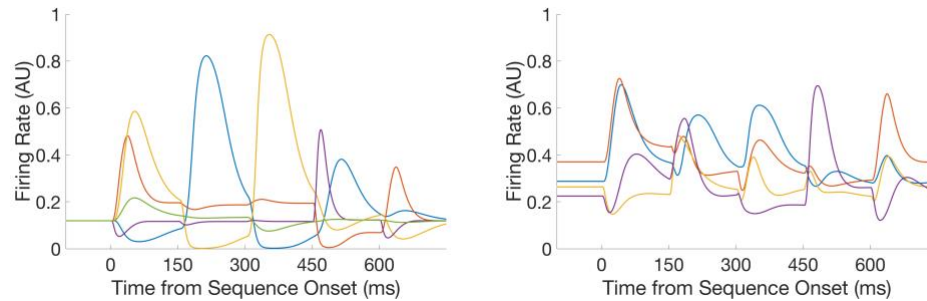
Sequence learning was discovered in the context of another form of experience-dependent plasticity known as stimulus-selective response potentiation (SRP). SRP occurs when mice sit passively, head-fixed, viewing a high-contrast, phase-reversing sinusoidal grating that oscillates at low frequency (typically 0.5 – 2 Hz) (Cooke & Bear, 2010; Frenkel et al., 2006; Kaneko, Fu, & Stryker, 2017). After several days of passive exposure, for only a few minutes each day, visually-evoked local field potentials (VEPs) potentiate only in response to the precise stimulus viewed during training (the now-familiar stimulus). The degree to which this increase is selective for the familiar grating is remarkable. Changes in orientation of as little as 5 degrees from the trained orientation significantly reduce VEP magnitudes from ~200% of baseline to ~130% (Cooke & Bear, 2010, 2014; Frenkel et al., 2006). Changing the spatial frequency from 0.05 cycles per degree to 0.1 reduces responses to baseline. Training with a stimulus at 50% contrast causes an ~150% potentiation in VEP magnitude after 5 days, but showing the same stimulus at 100% contrast shows no

differentiation between familiar and novel. Given that neurons in retina and dLGN are relatively invariant to contrast, this kind of contrast-sensitive tuning hints at additional processing beyond the first thalamocortical synapse: a simple Hebbian modification of excitatory thalamocortical synapses is insufficient to explain SRP (Cooke & Bear, 2014; Montgomery, Hayden, Chaloner, Cooke, & Bear, 2022). The effect does depend on NMDARs and is blocked when NMDAR antagonist 3-(2-carboxypiperazin-4yl)propyl-1-phosphonic acid (CPP) is either locally or systemically present during the training phase, known as induction (Cooke, Komorowski, Kaplan, Gavornik, & Bear, 2015). However, knocking out NMDARs in thalamorecipient excitatory neurons in layer 4 of V1 does not impact SRP (Fong et al., 2020). Furthermore, complex network interactions, involving the activity of both PV and SOM interneurons, seem to be critically important (Hayden, Montgomery, Cooke, & Bear, 2021; E. S. Kaplan et al., 2016).

Sequence learning is a variation on SRP. Rather than alternate between two phases of the same orientation, the sequence stimulus shows four gratings with different orientations, ABCD. It adds a sub-second timing component to SRP and a complex set of spatiotemporal transitions (Figure 4). Each grating is held on screen for a brief period and then gives way to the next with no interleaving gray screen. In the original experiments, the four gratings were presented for exactly 150ms each (Gavornik & Bear, 2014). The physiological effects of this exposure were measured using VEPs and multi-unit activity, both recorded in layer 4 of binocular V1. They found that several days of training led to dramatic increases in VEP magnitude, and also to significant changes in multi-unit firing. Within a single day of

training, no changes in VEP magnitude were observed, so the underlying plasticity mechanism requires at least a few hours for expression and perhaps a period of sleep. After five days of training, playing the same stimulus in reverse, DCBA, revealed mild potentiation relative to naïve baseline and significantly less than the trained ABCD. Changing the timing of the sequence from the trained 150ms to 300ms also dramatically reduced VEP magnitudes. Control mice who viewed randomized sequences (elements still held for 150ms each but with a variety of orientations) showed no potentiation or specificity for stimulus timing. Thus, whatever form of neural plasticity underlies this effect, it is very selective for the order and timing of the trained sequence, ABCD. Finally, subsequent research has shown that similar effects occur in anterior cingulate cortex (Sidorov et al., 2020) and that an intact hippocampus is required for sequence learning but not for SRP (Finnie, Komorowski, & Bear, 2021).





**Figure 4: Sequence Stimulus Representation in a Model of dLGN and V1**

**Top:** Visual representation of the sequence stimulus, with four sinusoidal gratings representing ABCD. In dotted red is the outline of a realistically-sized dLGN receptive field. Thus, the stimulus inside the circle shows what a dLGN neuron would see as the sequence is presented. **Middle:** Average luminance during a 150ms per element sequence, as seen by the dLGN neuron represented above. Luminance values were calculated as the mean inside the receptive field. **Bottom Left:** Example traces of hypothetical dLGN activity, computed from a simple model with center-surround spatial and biphasic temporal receptive fields. Different dLGN neurons can respond very differently to the sequence, depending on their receptive field location in visual space, and whether they are ON, OFF, or ON/OFF types. **Bottom Right:** Example traces of hypothetical V1 simple cells that average the activity of a small number of dLGN inputs. Even though these neurons were designed to be orientation selective, they still tend to respond to most elements in the sequence. This is due to the salient steps in luminance that occur at element transitions and the location of the receptive field relative to the phase of the gratings.

Mechanistically, little is known about the changes that occur during learning. VEPs are generally difficult to interpret, as they tend to reflect dendritic conductances across several layers of cortex rather than local spiking activity (Katzner et al., 2009). However, both systemic and local injection of scopolamine, an antagonist for muscarinic acetylcholine receptors (mAChRs), during training prevented learning-related changes in the VEPs. In addition, systemic injection of CPP, an NMDAR antagonist, had no effect. Because NMDARs are implicated in numerous forms of synaptic plasticity, including LTP, LTD, and STDP, sequence learning is likely unrelated to these more traditional Hebbian learning paradigms. This in itself is interesting because the sequence stimulus drives high firing rates in layer 4 of V1, as measured by multi-unit and single-unit activity. Based on the

properties of retina and dLGN, the stark transitions between gratings ought to drive high activity in thalamocortical neurons (Figure 4). Therefore, thalamocortical synapses are likely the primary source of excitation creating the initial barrage of spikes in V1 about 50ms after the onset of each grating. According to simple Hebbian plasticity rules, this ought to induce strengthening of those synapses. That this does not happen is certainly worthy of further reflection.

The relevance of acetylcholine to sequence learning may provide crucial clues and context for understanding both its mechanism(s) and its purpose. Acetylcholine is a critical neuromodulator for learning and memory (Hasselmo & Stern, 2006; Saar, Grossman, & Barkai, 2001). It has also been implicated in visual cortical coding, attention, and perceptual learning (Goard & Dan, 2009; Herrero et al., 2008; Mincses, Alexander, Datlow, Alfonso, & Chiba, 2013). Acetylcholine is often found to facilitate activity at more traditional synapses, suggesting one of its roles may be to promote heterosynaptic plasticity (Kuo, Rasmusson, & Dringenberg, 2009; Mincses et al., 2013; Pinto et al., 2013; Rasmusson, 2000). V1 is innervated by cholinergic projections from the basal forebrain, specifically the horizontal limb of the diagonal band of Broca (Huppé-Gourgues, Jegouic, & Vaucher, 2018). Acetylcholine is believed to be released through bulk transmission, meaning cholinergic axons do not tend to terminate in individual synapses (Bruno et al., 2006; Sarter, Hasselmo, Bruno, & Givens, 2005), though acetylcholine release can rapidly alter cortical activity (Pinto et al., 2013; Rajan Dasgupta, Seibt, & Beierlein, 2018).

In V1, there are a variety of acetylcholine receptors, including nicotinic and muscarinic subtypes (Disney, Alasady, & Reynolds, 2014; Disney & Aoki, 2008; Sadahiro, Sajo, & Morishita, 2016). mAChRs come in five subtypes, M1-M5, that are differentially expressed across layers and cell types, though M1 and M2 are the predominate subtypes in rodents. M1 receptors are strongly expressed across layers in rodents, mostly at postsynaptic sites on pyramidal neuron dendrites (Groleau, Kang, Huppe-Gourgues, & Vaucher, 2015; Gullledge, Bucci, Zhang, Matsui, & Yeh, 2009). Their activation leads to an influx of calcium ions ( $\text{Ca}^{2+}$ ) and they have been shown to heterosynaptically enhance LTP at hippocampal CA1 synapses (Dennis et al., 2015). M2 receptors are found most often in layer 4 of cortex, but they are also expressed across layers (W. Zhang et al., 2002). M2 receptors are mostly inhibitory *autoreceptors* on cholinergic axon terminals, where they block release of acetylcholine through negative feedback (C. L. Douglas, Baghdoyan, Lydic, & Baghdoyan, 2002). They are also expressed on pre-synaptic inhibitory (mostly SOM) terminals and serve to inhibit the release of GABA (Salgado et al., 2007). Interestingly, M2 receptors in V1 have been found to cover the cortical surface in a mosaic of low- and high-expression patches (Ji et al., 2015; Q. Wang & Burkhalter, 2007; Q. Wang, Gao, & Burkhalter, 2011). M2 dense patches correspond to functional regions of high spatial acuity, while M2 poor regions exhibit high temporal acuity. This has led to the hypothesis that M2 dense and poor patches correspond to regions giving way to ventral- and dorsal-like streams in the mouse (M2 dense to ventral/object, M2 poor to dorsal/motion). Evidence from our lab suggests that M2 receptors mediate sequence learning (Sarkar, Reyes, Jensen, & Gavornik, 2022).

Given the weight of evidence, we can speculate on sequence learning's potential functional role in visual processing and its mechanism of action. With respect to efficient coding, the most general rule is that unexpected and rare events ought to be encoded with more symbols than expected and common events. In the nervous system, this means more spikes for unexpected events [in a linear system, firing rates would directly encode information/surprise,  $-\log(P(x))$ ]. Thus, repeated exposure to a behaviorally irrelevant and predictable stimulus ought to steadily suppress responses in cortex. This has been shown to occur across the visual system (Freedman, Riesenhuber, Poggio, & Miller, 2006; Montgomery et al., 2022; Woloszyn & Sheinberg, 2012). If B, C, & D always follow A, then there is no need to transmit BCD: A contains all of the information encoded in the sequence. Complete suppression of the responses to BCD would require either an accurate timing mechanism or a prolonged period of inhibition persisting for the duration of the sequence. Now, there are competing concerns because the same synapses that transmit ABCD also transmit other information; the mouse only watches the stimulus for about two minutes per day. We might therefore expect the system to adapt to the novel input statistics, for example by selectively and modestly suppressing temporal information at the sequence frequency, akin to a notch filter. The issue with testing this hypothesis is that we cannot be certain whether this suppression would generalize to other stimuli or be restricted to the exact neurons and synapses involved. Despite this concern, in the experiments described below, we look for evidence that the visual system has learned the sequence by showing mice variations on ABCD. Though we could not predict *a priori* what exactly we would

observe, we expected 1) ABCD to itself be a novel and unexpected stimulus in naïve mice, thus evoking significant activity and 2) novel sequences to violate learned expectation in trained mice and show a differential effect relative to controls.

Regarding the mechanism, one clear hypothesis is that M2 receptors expressed on SOM interneurons are at least partially responsible for sequence learning. It would be difficult to speculate the exact mode of action, but perhaps acetylcholine gates heterosynaptic plasticity at GABAergic synapses, either strengthening or weakening them. In the novel experimental data presented here, we find evidence that evoked multi-unit responses to A (and maybe to B, C, & D) are suppressed in a late window (100-150ms) after sequence onset. Since about 80% of cortical neurons are excitatory, decreased firing in this late window may be the result of increased inhibition. The extended time course of late-window suppression suggests that one possible explanation is potentiation of inhibition through modification of pre-synaptic GABA release or post-synaptic GABA-B receptor density. Comparable forms of inhibitory plasticity, or LTP-i, have been found experimentally (Gandolfi, Bigiani, Porro, & Mapelli, 2020; Hennequin et al., 2017). However, recurrent connectivity within V1, or modifications to the temporal tuning curves of dLGN axons, could also be responsible for the observed effect of suppression in the late window.

A somewhat different perspective is to consider recent findings that the visual system may “straighten” visually-evoked neural trajectories (Henaff et al., 2021; Hénaff, Goris, & Simoncelli, 2019). The basic idea is that many behaviors, like catching a baseball, require

us to extrapolate and predict the future. In doing so, the visual system must make its predictions using only neural activity (initially, RGC spikes), so in essence it must predict future neural states based on recent neural trajectories. Hénaff et al. argue that these predictions will be easier to make if the neural trajectories are closer to linear. The overall paradigm is still in its infancy, and it is not entirely clear that the concept of straightening is well-posed and mathematically precise, but it is clearly related in some way to unsupervised learning and efficient coding. In our experimental data, we do not quantify the effect, but we see something resembling straightening, such that neural trajectories to ABCD in naïve mice are more complex compared to those in trained mice.

## **EXPECTATION VIOLATIONS PRODUCE ERROR SIGNALS IN MOUSE V1**

This section contains the text of our recent paper, currently published in pre-print form on bioRxiv and under review. The authors are Byron Price, Cambria Jensen, Anthony Khoudary, and Jeff Gavornik (Price, Jensen, Khoudary, & Gavornik, 2022).

### **Introduction**

Sensory inputs contain statistical regularities that convey information about the external environment. Numerous lines of experimental and theoretical evidence suggest that the brain exploits these regularities to understand and predict the structure of the world around us (H. B. Barlow, 1961; Chalk, Marre, & Tkačik, 2018; Friston, 2005; Hosoya et al., 2005; Keller & Mrsic-Flogel, 2018; Palmer et al., 2015; Rao & Ballard, 1999; M.W. Spratling, 2017; Zmarz & Keller, 2016). Such predictions are useful for a wide variety of behavioral and neural tasks including efficient movement coordination (Körding & Wolpert, 2004; McNamee & Wolpert, 2019; Wolpert, Ghahramani, & Jordan, 1995), trajectory extrapolation (Montague & Sejnowski, 1994; Palmer et al., 2015), and information compression (H. Barlow, 2001a, 2001b; Dan et al., 1996; Srinivasan et al., 1982) among others. Exactly how this is done remains an open question, but it is becoming increasingly clear that early visual areas encode efficient representations of their inputs both in space and across time (H. Barlow, 2001b; H. B. Barlow, 1961; Collewijn et al., 2008; Dan et al., 1996; Elias, 1955; Hosoya et al., 2005; Kelly, 1985; Kuang et al., 2012; Lappe, Bremmer, & Van Den Berg, 1999; Leinweber et al., 2017; Palmer et al., 2015; Rucci, 2008; Srinivasan et al., 1982; Zmarz & Keller, 2016). The predictive coding model posits that this is accomplished by comparing predictions against incoming sensory data, then

transmitting prediction error signals (Elias, 1955; Keller & Mrsic-Flogel, 2018; Rao & Sejnowski, 2001; M.W. Spratling, 2017; Srinivasan et al., 1982; Zmarz & Keller, 2016). This model has many attractive elements, and may help explain phenomena ranging from extra-classical receptive field properties in V1 (Atick & Redlich, 1993; Rao & Ballard, 1999) to the activity of dopamine neurons in the ventral tegmental area (VTA) (W. Schultz, Dayan, & Montague, 1997; Wolfram Schultz, 2016). However, many biological implementation details remain vague and specific elements of the model have not been validated in the brain. Particularly lacking is a clear explanation of how experience-dependent plasticity encodes the temporal relationships required for predictions into neural circuits.

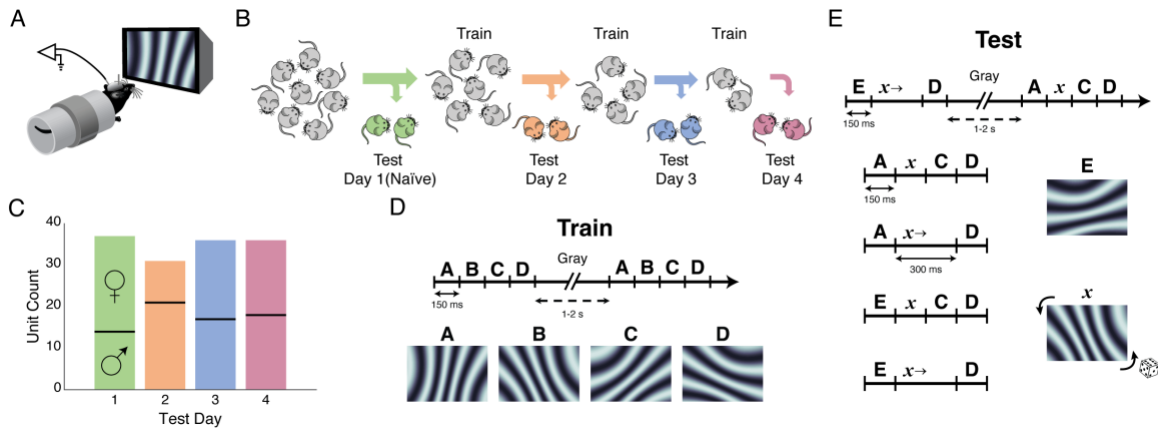
Experimental evidence originating in multiple labs over the last decade has demonstrated that visual experience shapes V1 circuits to encode and predict temporal relationships (Finnie et al., 2021; Garrett et al., 2020; Gavornik & Bear, 2014; Gillon et al., 2021; Homann, Koay, Glidden, Tank, & Berry, 2017; Orbán, Berkes, Fiser, & Lengyel, 2016; Shuler & Bear, 2006; Weliky, Fiser, Hunt, & Wagner, 2003; Zmarz & Keller, 2016). Insights gained from V1 may apply broadly to other regions as well (R. J. Douglas & Martin, 2004; R. J. Douglas, Martin, & Whitteridge, 1989; Edelman & Mountcastle, 1978; Hawkins & Ahmad, 2016), making this an experimentally accessible area in which to study the cortical basis of predictive processing. In the present study, we investigated predictive processing in the context of a specific form of sequence learning previously described as in V1 (Finnie et al., 2021; Gavornik & Bear, 2014) and anterior cingulate cortex (ACC) (Sidorov et al., 2020). Specifically, we passively exposed mice to rapidly-flashed

sequences of sinusoidal gratings, whose spatiotemporal structure differs dramatically from the approximate  $1/f$  spectrum of their natural environment (Carandini et al., 2005; Ocko et al., 2018; Olshausen & Field, 1997). In doing so, we sought to determine whether prediction errors were generated in V1 and the extent to which those prediction errors were tuned to the spatiotemporal structure of the learned sequence.

## Results

### *Experimental Design and Sequence Stimulus*

Our goal in designing the experiment was to characterize how neural responses change with experience, under the assumption that we could not track individual neurons across days. To do so we created a novel randomized Train-Test experimental protocol (Figure 5) that presented a set of Test stimuli to mice after zero to three days of training. During training, each mouse was presented with a single training sequence, ABCD (Figure 5d), in order to condition them to the spatiotemporal structure of that sequence. After training, mice viewed randomized test sequences to determine how different types of expectation violation change neural responses in V1 (Figure 5e). Animals were randomly assigned a Test day (1-4). Animals that saw the Test set on Day 1 (e.g., after zero days of training) served as naïve controls for the “trained” mice who saw the Test stimulus after at least one day of training. Each mouse saw the Test stimulus set only once and then were removed from the experiment. The randomization procedure insured that the amount of data collected from each Test day group was approximately balanced (Figure 5b-c).



**Figure 5: Experimental Design**

**A.** Awake head-fixed mice viewed stimuli from a distance of 25 cm simultaneous with extracellular recording via chronically implanted wire bundles. **B.** Mice were randomly assigned to see Test stimuli on day 1 (25%, naïve controls, green), day 2 (33% of remaining mice, orange, 1 day of training), day 3 (50% of remaining, blue, 2 days of training), or day 4 (100% of remaining, magenta, 3 days of training). **C.** The number of units recorded on each day was approximately uniform and evenly distributed between female and male mice (top & bottom of each bar). In total, we recorded 140 unique multi-unit channels from 56 mice. **D.** Each training session consisted of 200 presentations of the sequence ABCD. Each element had a unique orientation and was held on screen for 150ms and the full sequence lasted 600ms. Sequences were separated from each other by a gray screen, held for a uniform random interval between 1 and 2 seconds. **E.** During Test sessions mice were exposed to 600 presentations of randomly selected novel sequences:  $AxCD$ ,  $ExCD$ ,  $Ax \rightarrow D$ , and  $Ex \rightarrow D$  where  $x$  indicates a random orientation (uniformly distributed  $\pm 60^\circ$  around B) and  $\rightarrow$  indicates an omitted third element (second element on screen for 300ms, 50% of trials).

All mice were awake and head fixed while viewing sequence stimuli during both *Train* and *Test* sessions. Each element in the sequence lasted 150ms and a complete 4-element sequence lasted 600ms (Figure 5d,e). Element transitions within a sequence were continuous and a gray screen was used to separate individual sequence presentations. To eliminate the possibility that the animal could anticipate when the next sequence would begin, the gray screen was displayed for a random interval drawn from a uniform distribution between 1 and 2 seconds. In all sessions (*Train* and *Test*), we recorded spiking data from Layer 4 neurons within the binocular region of V1 (Supplemental Figure 1),

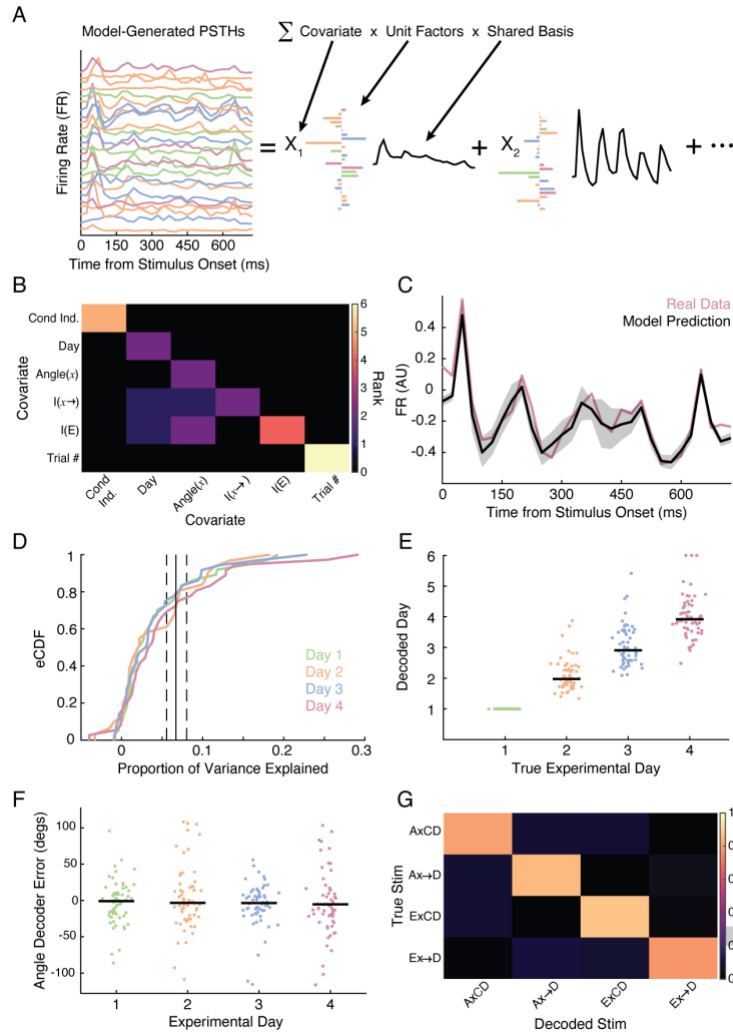
along with 50-Hz infrared video of the face and whisker pad (Supplemental Figure 2). In total, we identified 140 distinct, visually-responsive multi-unit channels from 56 mice (29 females, 27 males; P66.8  $\pm$  0.5 days at experimental start [mean  $\pm$  SEM]; see Methods for unit inclusion criteria).

During *Test* sessions, each sequence was initiated with the familiar element A or a novel element E that was not contained in the training sequence. The second *Test* sequence element,  $x$ , was a randomly oriented grating held on the screen for either 150ms ( $AxCD$  and  $ExCD$ ) or, in 50% of trials, 300ms ( $Ax\rightarrow D$  and  $Ex\rightarrow D$ ). This design allowed us to test the importance of familiarity (A vs E), orientation ( $x$ ), temporal expectation ( $\rightarrow$ ), and days of training, as well as to look for evidence of error signals associated with positive (unexpected inclusion) and negative (unexpected omission) prediction errors as described in various predictive coding models (Keller & Mrsic-Flogel, 2018; M.W. Spratling, 2017).

#### *A Statistical Model (MbTDR) Captures Stimulus-Dependent Neural Variability*

We used a statistical modeling tool known as model-based targeted dimensionality reduction (MbTDR) (Aoi, Mante, & Pillow, 2020; Aoi & Pillow, 2018) to analyze our high dimensional, heterogeneous data. We hoped the model would reveal learning-related changes in neural activity that are not evident from examination of the PSTHs. MbTDR is a supervised probabilistic model that projects high-dimensional data onto low-rank subspaces spanned by different regression covariates (Figure 6a). Each subspace is a set of shared basis functions, low-dimensional neural trajectories, along with “unit factors” that specify how much each basis contributes to the trial-by-trial firing rate of a given unit. We

created covariates related to the structure of the sequence stimulus and our experimental design, including condition independent (baseline PSTHs), experimental day, first element (A or E, encoded as an indicator function,  $[I(E)]$ ), the orientation of the randomized second element ( $Angle(x)$  &  $Angle(x)^2$ ), second element duration  $[I(x\rightarrow)]$ , trial number, and combinations/interaction-terms thereof. Using a greedy forward stepwise algorithm that minimizes the Akaike Information Criterion, the model selects the optimal rank for each covariate and interaction term. The algorithm allows the optimal rank for a given covariate to be zero and finding a rank greater than zero is comparable to discovering a significant effect for that covariate in a regression model such as ANOVA. The optimal rank is the number of unique PSTH templates that are necessary to explain the variability in PSTHs across the population. If the responses for all units were completely independent, and no low-dimensional representation were possible, then the optimal rank for each covariate would equal the total number of recorded units.



**Figure 6: MbTDR Captures Stimulus-Dependent Neural Variability**

**A.** Model-based targeted dimensionality reduction (MbTDR, after Aoi & Pillow 2018) models PSTHs as a linear combination of covariates, unit factors, and shared basis functions. MbTDR discovers a low-rank representation, consisting of a small set of shared bases that are weighted differently for each unit (color code as in Figure 1b). **B.** Rank of optimal model, fit by maximum likelihood and a greedy rank estimation algorithm. Covariates were chosen to represent the experimental design and Test stimulus (see main text for details). Diagonal elements are the rank of each covariate, and off-diagonal elements represent the rank of interaction terms. **C.** An example day-4-unit PSTH (binned at 25ms with no smoothing, z-scored to zero spontaneous baseline firing and unit variance, as in all subsequent figures and analyses). The MbTDR prediction (black) matches the neural data (magenta, 95% confidence interval computed from observed Fisher information in gray). This unit had 4.41% held-out explained variance. **D.** Empirical cumulative distribution functions (eCDFs) of explained variance across the 140 units. Each line represents the eCDF computed using held-out data from one Test day (60 trials / day). The vertical black line is an estimate of explainable variance with a bootstrap 95% confidence interval. **E.** Test days can be accurately decoded (each dot represents a single held-out pseudo-trial, x-axis jitter added for

visibility). Due to our experiment and model design, Day 1 trials are automatically known. **F.** Decoding error for  $x$ , the randomized angle of the second element, on the same held-out data (each dot is one pseudo-trial). The standard deviation on the error is 38.7 degrees. **G.** Confusion matrix for decoding the four primary stimulus types. Overall accuracy was 82.5%, while chance for 60 trials is  $25 \pm 2.7\%$  (mean  $\pm$  SEM, gray box on color bar shows 95% confidence interval of this estimate). The identity matrix would constitute perfect accuracy.

We fit a single model for all recorded units on 90% of trials using maximum likelihood and the greedy algorithm mentioned above; 10% of the data was held-out of the model fitting in order to estimate explained variance and to use for decoding. This was done only once, reserving the held-out 10% throughout training. The final total rank of the model was 26, with ranks for each covariate and interaction term displayed in Figure 6b (model bases are depicted in Supplemental Figure 5). All of the fundamental covariates were significant, but the only significant interaction terms were:  $I(x \rightarrow) * Day$ ,  $I(E) * Day$ ,  $I(x \rightarrow) * Angle(x)$ ,  $I(E) * Angle(x)$ ,  $I(E) * Angle(x)^2$ . The relevance of these terms is discussed below (for more details on the model and covariates, see Methods and Supplemental Table 1).

The final model achieved 6.2% explained variance on the held-out data, which was comparable to our estimate of  $6.7 \pm 0.7\%$  *explainable variance* (mean  $\pm$  SEM) (Figure 6c-d). In order to estimate explainable variance, we fit PSTHs to repeated presentations of the same sequence (ABCD on Training days). In this case, the PSTH captures predictable variability due to the stimulus, and all unexplained variance must be due to factors that are inconsistent across trials (movement, neural variability, etc.). There was a strong positive correlation between evoked firing rate and held-out explained variance, despite the fact that units were normalized to unit variance prior to model fitting:  $\rho = 0.45$  ( $p < 1e-6$ ,

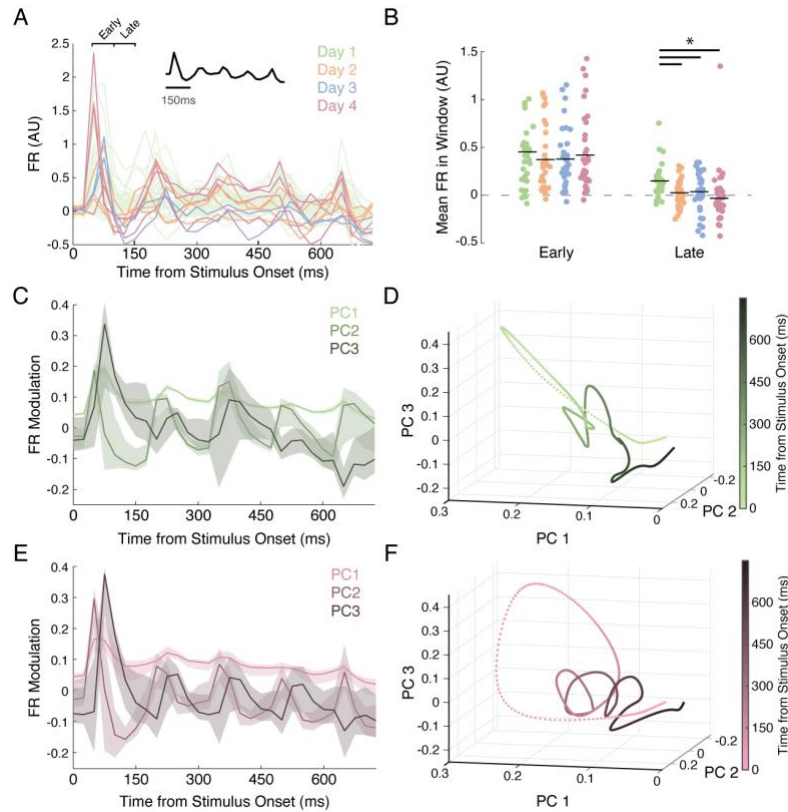
Spearman's rho permutation test,  $N=140$  units). However, there was no such correlation between explained variance and experimental day:  $\rho = 0.053$  ( $p=0.54$ ), indicating that model fit was comparable regardless of the number of training days. Finally, 125/140 units had greater than 0% held-out explained variance, and 48/140 had greater than 5% (see Figure 6c for an example PSTH from a unit with about 5% explained variance), indicating the model accurately captured stimulus-dependent, trial-by-trial fluctuations in neural firing.

To further convince ourselves the model fit the data well, we used Bayes' theorem to decode stimulus covariates on held-out data (Figure 6e-g). For each *Test* day, we created 60 pseudo-trials by combining data from all recorded units, about 35 per day. Stimulus features could be accurately decoded on individual pseudo-trials, revealing strong correlations with ground truth:  $\rho = 0.78$  for experimental day ( $p<1e-6$ , Spearman's rho permutation test, excluding Day 1, Figure 6e);  $\rho = 0.50$  for angle of  $x$  ( $p<1e-6$ , Figure 6f);  $\rho = 0.59$  for trial number ( $p<1e-6$ );  $\rho = 0.80$  for the third element omitted ( $I(x\rightarrow)$ ,  $p<1e-6$ ); and  $\rho = 0.81$  for E starting the sequence ( $I(E)$ ,  $p<1e-6$ ). When asking the decoder to determine which of the four primary Test stimulus types had been displayed ( $AxCD$ ,  $ExCD$ ,  $Ax\rightarrow D$ , or  $Ex\rightarrow D$ ), the decoder achieved 82.5% accuracy on 240 held-out pseudo-trials, compared to 25 +/- 2.8% by chance (Figure 6g). Together, these results demonstrate the efficacy of the model in capturing stimulus-dependent neural variability.

*MbTDR Reveals Coordinated Training-Dependent Changes in Neural Activity*

We found a modest increase in evoked firing rate across Test days. The global median evoked firing rate was 7.8 Hz, while for each Test day it was: 1- 6.4 Hz; 2- 11.0 Hz; 3- 10.6 Hz; 4- 14.2 Hz (two-sided permutation test for difference in median firing rate:  $p = 0.026$  Day 1 vs. Day 2;  $p = 0.034$  for Day 1 vs. Day 3;  $p = 0.006$  for Day 1 vs. Day 4). Besides this simple measure of learning-dependent change, the MbTDR model-selection process also chose *Day* as a significant covariate with a rank of 2. Units that were strongly modulated by these bases are shown in Figure 7a. Note that in this and subsequent figures firing rates are z-scored to zero spontaneous baseline firing and unit variance.

Comparing evoked responses across days, it was visually evident that the overall increase in firing rate was not uniform over the duration of the sequence. The initial response following element transitions tended to increase in magnitude, while firing rate during a later sustained response decreased. To quantify this effect, we calculated the average firing rates in early (51-100ms) and late (101-150ms) windows after the onset of element A across test days. While increases in firing during the early window were not significant, there was significantly less firing in the late window for units from trained mice (Figure 7b). This dip in firing was also captured by the MbTDR *Day* covariate, as seen in Figure 7a.



**Figure 7: MbTDR Reveals Coordinated Training-Dependent Change in Neural Activity**

**A.** Example PSTHs to AxCD (computed from the raw data, not the model fit). Units from trained mice (Days 2,3,4) that were strongly modulated by the Day covariate (inset) are overlaid above example PSTHs from Day 1 units (green). Note the dip in firing after the onset of A in trained units. Some trained units drop as low as -0.5 on this z-scale, while no unit from Day 1 (the naïve group) drops below -0.15 in that same window. Early / Late designations indicate the time windows used in **B**. **B.** Comparison of mean normalized firing rate across Test days (again computed from the raw data, not model fit). Left shows mean firing rate in an early window (51-100ms after onset of A), right during a late window (101-150ms). There was no significant difference between trained and naïve groups in the early window (two-sided permutation test for difference in mean from Day 1, with a threshold set to  $p < 0.05/6 = 0.0083$  for multiple comparisons: Day 2  $p = 0.441$ ; Day 3  $p = 0.437$ ; Day 4  $p = 0.668$ ). However, evoked firing in the late window was significantly lower in all trained groups (Day 2  $p = 6.92e-4$ ; Day 3  $p = 0.0052$ ; Day 4  $p = 5.02e-4$ ). Black dotted line marks the spontaneous baseline firing rate. **C.** First, second, and third principal components of evoked neural activity to ABCD in naïve mice, computed from the MbTDR fit (see Methods) with 95% confidence intervals. **D.** Dynamic latent trajectory representation of the data in **C** in principle component space. Data was projected from 25-ms bins as in **C** to 1-ms bins using a Gaussian radial basis with 12.5-ms standard deviation. **E.** Same as **C**, but for the set of Day 4 units. The first component remains unchanged across days, while the second and third differ dramatically from Day 1. Note the dip in the second component from Day 4 around 100ms, which mirrors decreased firing in the late window observed in **B**. **F.** Same as **D**, but for Day 4 units. The second and third dimensions show rotational dynamics which, along with the first component, create a spiraling latent trajectory.

Next, we used MbTDR to visualize low-dimensional dynamic neural trajectories across Test days by performing a singular-value decomposition of the model fits for all possible test sequences (see Methods). The first principal component (Figure 7c,e in light shade) shows an initial excursion at the onset of the stimulus and then a slow relaxation back to baseline. This was completely unchanged across days and accounted for about 91.7% of the variance in the PSTHs (i.e., signal variance) on Day 1 and 78.1% on Day 4. However, the second and third components, accounting for 6.2% of the variance on Day 1 and 16.5% on Day 4, changed dramatically between naïve and trained mice (Figure 7c-f). Figure 7c shows the first three components on Day 1, while Figure 7d provides a dynamic representation of those components plotted against one another. The dynamic representation from naïve mice crosses over itself and exhibits sharp direction changes, which might be described as highly tangled (Russo et al., 2018) or highly curved (Henaff et al., 2021; Hénaff et al., 2019). Fully trained mice, by comparison, evidenced stronger rotational dynamics in the second and third dimensions, and a spiraling trajectory in the first three components (Figure 7e-f).

#### *Orientation Tuning Does Not Shift Significantly with Training*

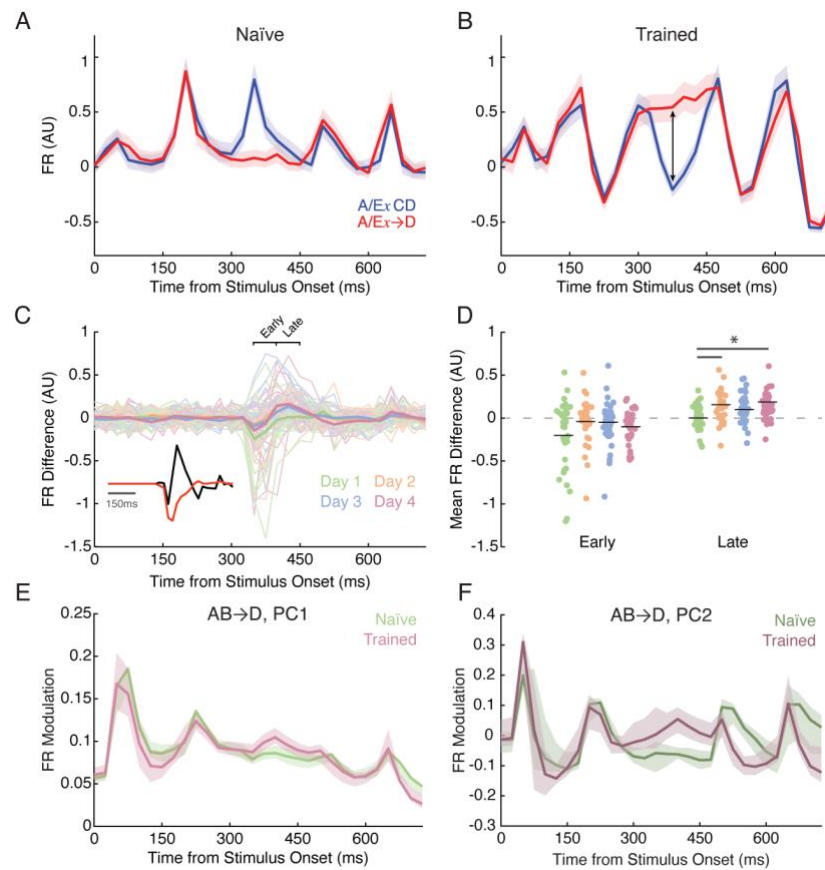
We hypothesized that sequence learning would cause a shift in population orientation tuning from an approximately uniform distribution (all angles equally represented with some over-representation for the cardinal angles) toward a distribution with more mass on the orientations of the trained elements, ABCD. This kind of unsupervised shift in orientation tuning has been observed in other experimental protocols, specifically SRP

(Kaneko, Fu, & Stryker, 2017). Randomizing the angle of the second element allowed us to estimate orientation tuning curves for each unit and compare against the trained B. The MbTDR model-selection process picked out  $Angle(x)$ ,  $Angle(x)^2$ ,  $I(x \rightarrow) * Angle(x)$ ,  $I(E) * Angle(x)$ , and  $I(E) * Angle(x)^2$  as significant covariates. This suggests individual units were not only tuned to the orientation of the second element, but that their tuning depended on the identity of the first element (stimulus-history-dependent tuning). Supplemental Figure 3a shows the shared bases for the  $Angle(x)$  and  $Angle(x)^2$  covariates. From the MbTDR output, we can obtain a complete orientation tuning curve for each unit, along with a peak tuning angle (Suppl. Figure 3b). With these peak tuning values, we then constructed an empirical cumulative distribution function (populating tuning curve) for each Test day (Suppl. Figure 3c). The peak-tuning cumulative distribution functions were *not* significantly different between naïve and trained mice (two-sided KS test, naïve [Day 1] vs. trained [Days 3 & 4]:  $D=0.106$ ,  $p=0.932$ ). This result suggests that training does not shift second-element orientation tuning curves towards the trained orientation of B on a large scale, though we cannot rule out the possibility that some neurons shift. This question could be addressed with a longitudinal dataset tracking individual units across days.

#### *Unexpected Omissions Cause Negative Prediction Errors*

Test-day stimuli  $Ax \rightarrow D$  and  $Ex \rightarrow D$  deviate from expectation by omitting the third element, C, from the sequence. This omission of an expected stimulus element was designed to elicit a negative prediction error in trained mice, indicating the absence of an expected stimulus. In units from naïve mice, neural firing during an omission reliably

decayed back to baseline (red trace in Figure 8a for one example unit). In trained mice, by comparison, about 25% of units showed a response consistent with a negative prediction error (red trace in Figure 8b for example unit). In those units, firing rates ramped in anticipation of the expected onset of element C, reaching their peak at about 300-325ms after the onset of the sequence. When C was shown, its onset precipitated a rapid decrease in firing (blue trace in Figure 8b). When C was omitted, however, elevated firing persisted until the onset of the next sequence element, D (red trace in Figure 8b), consistent with a negative prediction error.



**Figure 8: Unexpected Omissions Cause Negative Prediction Errors**

**A.** Example PSTHs from a Day 1 / naïve unit comparing A/ExCD (blue) and A/Ex→D (red) trials (mean with 95% confidence intervals). **B.** Example PSTH from Day 3. Omission of element C (in

red) drives increased and sustained firing in a late window after its expected onset. **C.** Difference PSTHs for all recorded units (red minus blue from previous example, marked by arrow in B). Inset shows shared basis functions from MbTDR fit for the  $I(x \rightarrow)$  covariate (red) and interaction term  $I(x \rightarrow) * \text{Day}$  (black). **D.** Comparison of difference PSTHs in early (left) and late (right) windows (marked in C) after expected onset of element C. The early window is 51-100ms after expected time of C onset, while late is 101-150ms. There was no significant difference in mean from trained and naïve groups in the early window (two-sided permutation test for difference in mean from Day 1, with a threshold set to  $p < 0.05/6 = 0.0083$  for multiple comparisons: Day 2  $p = 0.061$ ; Day 3  $p = 0.091$ ; Day 4  $p = 0.194$ ). In the late window, however, firing was significantly higher on Days 2 and 4 (Day 2  $p = 0.0012$ ; Day 3  $p = 0.016$ ; Day 4  $p = 5.02e-4$ ). **E-F.** The first and second principal components of neural activity for the  $AB \rightarrow D$  condition, in naïve (green) and fully trained (magenta) mice with 95% confidence intervals.

To further investigate this phenomenon, we created a set of difference PSTHs for all recorded units (Figure 8c) by subtracting the blue traces in Figure 8a-b from the red for all units ( $A/Ex \rightarrow D - A/ExCD$ ). The difference PSTHs clearly show a tendency for units from trained mice to fire at an elevated rate when C is omitted (Figure 8c, thick lines depict the mean across units, positive values indicate firing is greater when C is omitted). To quantify this effect, we calculated the mean firing rate difference in early (51-100ms) and late (101-150ms) windows after the expected onset of C (Figure 8d). While there was no significant difference in the early window, trained units from Days 2 and 4 had significantly higher firing-rate differences in the late window than those differences computed from naïve units. Defining all those units as prediction-error units whose late-window mean firing rate difference exceeded three standard deviations of the baseline, we found the following proportions on each day: Day 1- 3/37 (8.1%); Day 2- 7/31 (22.6%); Day 3- 7/36 (19.4%), Day 4- 13/36 (36.1%).

Furthermore, the MbTDR model-selection process chose  $I(x \rightarrow)$  and  $I(x \rightarrow) * \text{Day}$  as significant covariates (Figure 8c inset, also Figure 6b). The shared basis functions

recapitulate the difference PSTHs, showing that only in trained units (black trace, Figure 8c inset) the omission of C led to increased firing in the late window. We also used the model fit to visualize the first and second principal components of neural activity (Figure 8e-f). Comparing Day 1 and Day 4, these principal components clearly provide evidence for a negative prediction error. On Day 4, for example, PC2 shows a sustained increase in firing after the expected onset of C, while on Day 1 PC2 recapitulates the example unit from Figure 8a by decaying to baseline.

We next sought to determine the spatiotemporal specificity of the negative prediction errors, i.e., the extent to which they are tuned to the precise sequence observed during training (ABCD). We hypothesized that angles similar to B would elicit larger negative prediction error responses than angles far from B, but only in trained mice. The model selection procedure chose  $I(x \rightarrow) * Angle(x)$  as a significant interaction covariate, though it did not choose other covariates that might govern the orientation tuning of the negative prediction error response, such as  $I(x \rightarrow) * I(E)$  or  $I(x \rightarrow) * Angle(x) * Day$ . Therefore, the negative prediction error evident during the omitted third element likely depends on the angle of the second element,  $x$ , but not the angle of the first element (A versus E). However, because of the design of the model, the  $I(x \rightarrow) * Angle(x)$  term can have a complex relationship with other covariates. For example, the significant  $Angle(x)^2$  covariate implies that neurons are tuned quadratically to the orientation of the second element, as expected in V1. However, the underlying shared basis function for  $Angle(x)^2$  extends in time to at least 500ms into the sequence (see Supplemental Figures 3a, 4a). This implies

that orientation tuning curves for the second element,  $x$ , may be different from negative-prediction-error tuning curves computed during the omitted third element. To investigate this further, we used the model fit to determine the orientation tuning of the negative prediction errors during the early and late windows (51-100ms and 101-150ms after the expected onset of C).

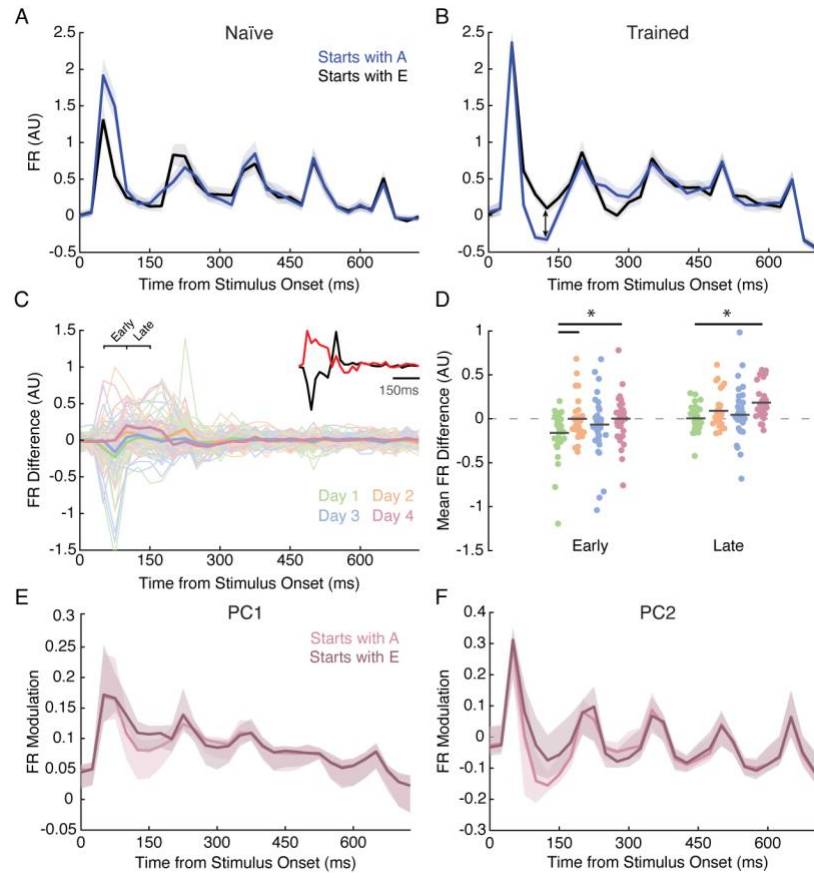
Supplemental Figure 4 depicts several example prediction error orientation tuning curves, from which we extracted peak tuning angles for each unit, as before. Across the population of recorded units, there was no appreciable difference between peak tuning angles in naïve and trained mice. We therefore conclude that there is limited spatiotemporal specificity to the negative prediction errors, i.e., they seem to occur even when the preceding visual sequence elements do not match those presented during training. This may reflect a limitation in the ability of layer 4 cells in V1 to adapt to the novel spatiotemporal statistics of the trained sequence, or perhaps that the step-function-like transitions between sequence elements are sufficiently salient to elicit the predictive response themselves.

#### *Unexpected Substitutions Cause Positive Prediction Errors*

The *Test* day stimulus set was designed to induce not only negative prediction errors, but also positive prediction errors that occur when an expected stimulus is replaced by an unexpected one. Given that the mice always saw ABCD during Training, replacing the expected A with a novel E violates expectation and should produce positive prediction errors. The MbTDR model selection process chose  $I(E)$  and  $I(E) * Day$  as significant covariates. The  $I(E)$  covariate suggests that the neural responses to A and E are different,

as expected due to their differing orientations. The  $I(E) * Day$  covariate suggests that this difference between A and E evolves with training (see Figure 9c inset for shared basis functions corresponding to these covariates).

In the PSTHs from trained mice, sequences starting with the familiar A showed decreased firing in a late window after stimulus onset (Figure 9b, also see Figure 7a,b). By comparison, sequences starting with E looked much more like the responses in naïve units (Figure 9b black trace, compared to Figure 9a black or blue). To quantify this effect, we looked at the firing rate differences between sequences starting with E and those starting with A (Figure 9c); these are the black traces from Figure 9a-b minus the blue. The firing rate differences in trained mice were significantly different from Day 1 in both the early and late windows (Figure 9d). During the late window in naïve mice, responses to A and E were approximately equal. In fully trained mice, however, the responses to A in the late window were significantly lower than those to E. This effect, of decreased firing in the late window for A but not E, was also picked up by MbTDR (Figure 9e,f). Training therefore created a differential neural response between familiar and novel stimuli. More precisely, the data suggests that training initiates a late-evolving inhibitory process that is only activated by expected stimuli. Unexpected stimuli, whether in naïve mice seeing the stimulus for the first time or in trained mice seeing a novel stimulus, do not activate the same process and show elevated firing in the late window.

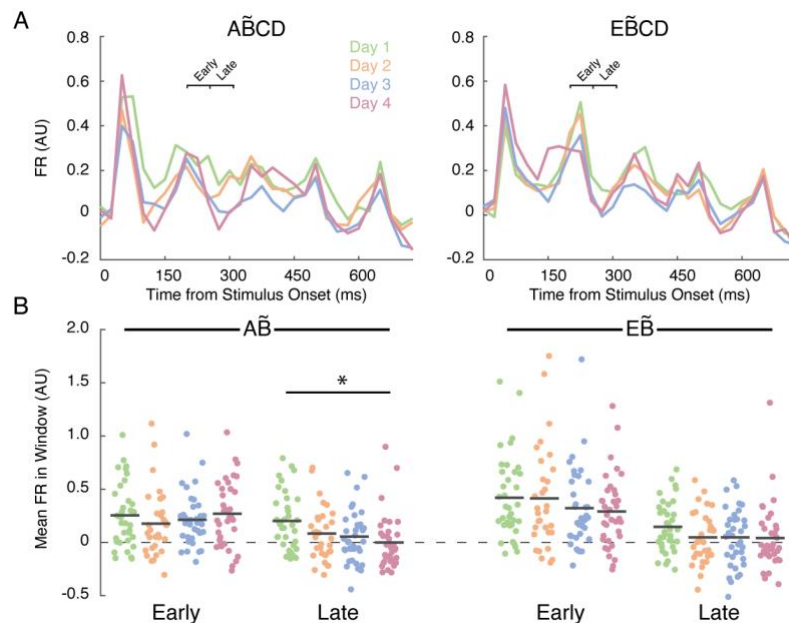


**Figure 9: Unexpected Substitutions Cause Positive Prediction Errors**

**A.** Example PSTHs from a Day 1 / naïve unit comparing all trials starting with A (blue) and E (black) with 95% confidence intervals. **B.** As in A but for a Day 4 unit. Presentation of the trained A drives a sustained decrease in firing during a late window after its onset that is not present following E. **C.** Difference PSTHs from all recorded units (black minus blue, illustrated with arrow in B). Inset shows shared basis functions from MbTDR fit for the  $I(E)$  covariate (black) and interaction term  $I(E)*Day$  (red). **D.** Comparison of difference PSTHs in early (left, 51-100ms) and late (right, 101-150ms) windows after onset of the first sequence element (A or E). There was a significant difference between naïve and trained groups in the early window (two-sided permutation test for difference in mean from Day 1, with a threshold set to  $p < 0.05/6 = 0.0083$  for multiple comparisons: Day 2  $p = 0.0062$ ; Day 3  $p = 0.23$ ; Day 4  $p = 0.0071$ ). In the late window, the firing rate difference was significantly greater only on Day 4 (Day 2  $p = 0.029$ ; Day 3  $p = 0.34$ ; Day 4  $p = 4.0e-6$ ). **E-F.** The first and second principal component of neural activity for the ABCD (light shade) and EBCD (dark shade) conditions with 95% confidence intervals. Both the first and second components show a significant dip following A, but not E, in the same late window.

During training, element A reliably predicted B. Assuming that learning encodes expected sequence order, the response to element B should look different if it is preceded by A (expected) versus E (unexpected). The model basis for  $I(E)$  modulates neural responses

for about 400ms (supplemental Figure 5) supporting the idea that individual elements influence the responses to subsequent visual stimuli. To investigate further, we identified sequences on Test days where the angle of the second element  $x$  was within  $\pm 5$  degrees of the trained element B, dubbed  $A\tilde{B}CD$  and  $E\tilde{B}CD$ . As shown in Figure 10a-b, the response to  $\tilde{B}$  was different between naïve and trained mice only for the sequence  $A\tilde{B}$ . In the late window after the onset of  $\tilde{B}$ , units from trained mice showed reduced firing relative to the naïve group, comparable to the positive prediction error effect observed between A and E. There was no significant difference between naïve and trained mice when E initiated the sequence. Though the statistics suggest that the response to the second element in the sequence depends on training and is differentiated by whether or not it is predicted by the preceding element, the data is not conclusive on this point. For example, the late-window  $\tilde{B}$  response following E shows a dip that looks similar to that seen when preceded by A. The current experimental design makes it difficult to interpret this result.



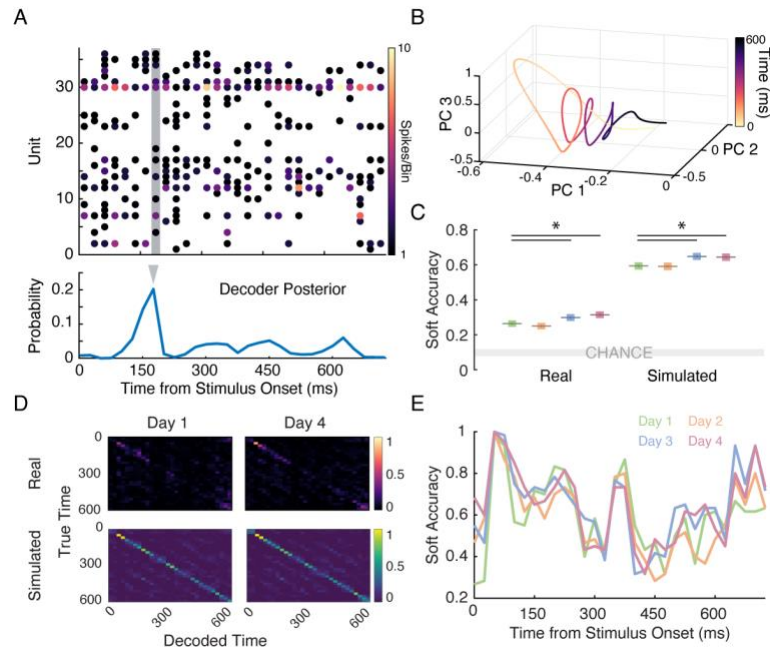
### Figure 10: Predictions Span Element Transitions

**A.** Left: Average PSTHs across all recorded units to the sequence  $A\tilde{B}CD$ , where  $\tilde{B}$  indicates that  $x$  was within  $\pm 5$  degrees of the trained  $B$  ( $B$  was 120 degrees, so  $\tilde{B}$  includes all angles from 115 to 125 degrees). Right: Same as Left, but for sequences  $E\tilde{B}CD$ . **B.** Mean firing rate in early (51-100ms) and late (101-150ms, marked in A) windows following the onset of the second element,  $\tilde{B}$ , when preceded by  $A$  (left) or  $E$  (right). Each dot represents one unit. Only the late window for  $A\tilde{B}$  was significantly different between naïve and trained mice (two-sided permutation test for difference in mean from Day 1, with a threshold set to  $p < 0.05/12 = 0.0042$  for multiple comparisons:  $A\tilde{B}$  early Day 2  $p = 0.293$ ; Day 3  $p = 0.520$ ; Day 4  $p = 0.837$ .  $A\tilde{B}$  late Day 2  $p = 0.065$ ; Day 3  $p = 0.016$ ; Day 4  $p = 0.001$ .  $E\tilde{B}$  early Day 2  $p = 0.951$ ; Day 3  $p = 0.280$ ; Day 4  $p = 0.137$ .  $E\tilde{B}$  late Day 2  $p = 0.095$ ; Day 3  $p = 0.100$ ; Day 4  $p = 0.105$ ).

#### *Reliable Temporal Information is Contained in the Neural Code*

As described earlier, a prominent hypothesis in visual neuroscience proposes two distinct pathways for “what” and “where” information (Flindall & Gonzalez, 2020; Ungerleider & Mishkin, 1982). Goodale & Milner later proposed a “how” pathway, that does not simply locate objects in the environment but coordinates the sensorimotor transformations necessary to grab and manipulate them (Goodale & Milner, 1992; Milner & Goodale, 2008). In addition to these classical pathways, the early visual system may also be a prominent source of “when” information (Rauschecker, 2018). Such information could be useful to establish temporal context, predict the timing of future events, or coordinate sensorimotor behavior (Goel & Buonomano, 2014; Mauk & Buonomano, 2004). “When” information may be measured in relation to a saccade, the onset of a movement, or relative to co-occurring events. In the case of the sequence stimulus, the onset of the sequence marks a strong departure from ordinary visual experience and could therefore be used to initialize a clock. Based on these ideas, we hypothesized that training would improve our ability to decode elapsed time from the evoked neural population response.

To test our hypothesis, we used the MbTDR fit to perform instantaneous time decoding, attempting to use the model and held-out neural data to predict elapsed time on individual pseudo-trials and time bins (Figure 11a). The decoder takes in a neural population response vector from a single time bin and uses the model to predict which bin, of 30 possible, the data came from. If the neural code in V1 were a poor source of temporal information the decoder would fail to produce accurate time estimates. We expected the decoder to perform better with training and that is what we discovered (Figure 11c-d). On Day 1, the decoder achieved a soft accuracy of 26.0%, compared to 9.8 +/- 1% by chance (soft accuracy allows for the decoder to be off by one time bin in either direction). On Day 2, soft accuracy was 24.9%; on Day 3, 29.8%; and on Day 4, 32.6%. We performed a two-sided permutation test on the difference in accuracy between Day 1 and all subsequent days, finding the accuracy to be significantly greater on Days 3 and 4 (Figure 11c). Thus, the ability to decode time improves with training.



**Figure 11: Temporal Information in the V1 Neural Code**

**A.** Top: Example pseudo-trial used for instantaneous time decoding. Each row represents a unit (all from the same Test day), while each column represents one 25-ms time bin. Bottom: Posterior probability distribution over possible time bins, calculated for the bin indicated by the shaded box above. In this case, the posterior correctly classifies the time bin based on a maximum a posteriori (MAP) estimate (marked with the gray triangle). **B.** Dynamic latent trajectory representation of ABCD from Day 3-4 units demonstrate spiraling dynamics (as Figure 3) that create unique locations in PC space for each time point. **C.** Accuracy of temporal decoding across 60 held-out pseudo-trials and 30 time bins / trial for real (left, ~35 units per day) and simulated (right, 250 units per day, see Methods) data. "Soft accuracy" allows the decoder to be off by one time bin in either direction. Accuracy improved significantly with training (two-sided permutation test for difference in accuracy from Day 1, threshold set to  $0.05/3=0.0167$  to control for multiple comparisons. Real data: Day 2  $p=0.26$ ; Day 3  $p=0.0059$ ; Day 4  $p<1e-6$ . Simulated data: Day 2  $p=0.802$ ; Day 3  $p=3.25e-4$ ; Day 4  $p=6.29e-4$ ). **D.** Time decoding confusion matrices for naïve (left) and trained (right) units for real (top) and simulated (bottom) data. **E.** Decoding soft accuracy in each time bin for simulated data. Training increases accuracy primarily at the beginning and ending of the sequence, which might be explained by the slight increase in evoked firing rate across days of training. Decoding is most accurate around element transitions (0, 150, 300, 450, 600 ms).

As described above, Day 1 has the highest N and model explained variance is no better in trained than naïve mice, so this result cannot be explained by differential fit to the data. However, firing rates did increase across days, which may affect decoder accuracy. Therefore, we removed the highest firing channels, destroying the correlation between

firing rate and day ( $\rho = 0.06$ ,  $p=0.54$ ), then re-ran the decoder. Temporal decoding accuracy still improved significantly between Days 1 and 4: Day 1- 22.1%; Day 2- 24.1% ( $p = 0.14$ ); Day 3- 21.9% ( $p = 0.92$ ); Day 4- 27.5% ( $p = 4.6e-5$ ). When we looked more closely at the individual time bins, however, we found that most of the improvement with training occurred during the evoked response to A and the offset at the end of the sequence. We reasoned that temporal decoding performance might be limited by the relatively small number of neurons we recorded from. To investigate further, used MbTDR to simulate a population of 1000 units (250 per day) whose dynamic activity matched the recorded data (Figure 7c-e). The temporal decoder was more accurate with the simulated data (Figure 7c-d) and this accuracy also increased with training, but similar to the real data improved accuracy from Day 1 to 4 was largely restricted to sequence onset and offset (Figure 7d-e). We conclude that evoked neural dynamics in V1 can serve as a reliable source of “when” information for downstream regions and that this information may increase with training. However, our data shows that improvements in decoding accuracy are not uniform across the temporal extent of the sequence, raising the question of whether increases in decoding accuracy provide evidence that the circuits explicitly encode temporal representations. Answering this question will likely require recording cells in other cortical layers and, potentially, other brain regions as well.

## Discussion

We have provided experimental evidence to support several basic principles of predictive processing. We used a statistical model to capture complex and heterogeneous changes in neural population activity across an unsupervised learning process. The model revealed

coordinated changes in evoked neural firing that depended strongly on training, including predictive ramping in advance of the expected onset of visual stimuli and spiraling 3D latent trajectories, which might be used as a basis for keeping time. In addition, the model uncovered a significant reduction in firing in a late window after stimulus onset, from about 100-150ms after both A and B, which progressively decreased with training. Due to this reduction, the unexpected substitution of the trained element A with the novel E caused elevated firing in that same late window. This differential response, consistent with a positive prediction error, was never present in naïve mice. Similarly, the omission of trained element C, which was done by extending the duration of the second element in the sequence, drove sustained, elevated firing in a late window after the expected time of C onset. This result is consistent with a negative prediction error. We note, however, that this was not restricted to sequences obeying the precise trained sequence order, i.e., those starting with AB. Indeed, the negative prediction error was only weakly tuned to the orientation of the preceding elements.

Finally, we used the model to perform instantaneous time decoding, showing that we could reliably decode elapsed time. This result is consistent with V1 being a source of “when” information, as the neural data at different timepoints throughout the sequence was uniquely identifiable. We also note that this ability to decode time is closely related to the concept of temporal redundancy reduction (H. Barlow, 2001a; H. B. Barlow, 1989). Redundancies in a sensory neural code arise when the same sensory information is transmitted at multiple points in space and time. Such redundancies arise from spatial and

temporal autocorrelations in the incoming visual signals, which are then propagated by neurons in the visual system. There is significant experimental evidence that the retina and lateral geniculate nucleus (LGN) work to reduce spatial and temporal redundancies in the neural code transmitted to V1 (Dan et al., 1996; D. Dong & Atick, 1995; Hosoya et al., 2005; Srinivasan et al., 1982). Given the complexity of the incoming signals, however, retinal and LGN processing are likely insufficient to fully remove the autocorrelations and compress the data. V1 might therefore continue the process of redundancy reduction, removing higher-order spatial and temporal correlations, before transmitting to downstream regions. Were this the case, we might expect unsupervised training with a stimulus like ours to reduce temporal redundancies and autocorrelations in the transmitted neural code as the system adapts to the novel stimulus statistics. Though indirect, temporal decoding accuracy might be useful as a surrogate measure for a lack of temporal redundancies.

Three notable recent studies investigated complementary ideas. In the first study, Marina Garrett and colleagues trained mice to perform a visual change detection task in which familiar and novel images were sequentially presented (Garrett et al., 2020). In general, their results closely matched the predictions of efficient coding and predictive processing. Excitatory neuron activity was reduced and more sparsely coded for familiar versus novel images, consistent with the idea that the visual system changed over time to compress its representation of the familiar images. They also showed that L2/3 VIP interneuron activity became predictive of the temporal structure of the sequence. In particular, VIP firing rates

ramped in advance of expected stimulus presentations, and ramping activity persisted when an expected stimulus was omitted. In our study, after training, a comparable ramp begins before the onset of each element. In trials where the third sequence element was omitted, the ramp plateaued and sustained for about 100ms until the appearance of the next sequence element. The difference between the continued ramping seen in Garrett et al. and our “ramp-then-sustain” phenotype may be due to recording location and neuron type. In particular, L2/3 VIP neurons that ramp may disinhibit L4 excitatory neurons, whose activity eventually saturates and sustains.

In another study, Peter Finnie et al. analyzed two forms of experience-dependent plasticity in V1 after bilateral lesion of the hippocampus (Finnie et al., 2021). They induced plasticity through sequence learning and also through stimulus-specific response potentiation (SRP). In the SRP protocol, mice passively view a sequence of phase-reversing sinusoidal gratings at 2 Hz (Cooke & Bear, 2010; Cooke et al., 2015; Frenkel et al., 2006). These two forms of plasticity have previously been shown to rely on different mechanisms, and this most recent study further confirmed those mechanistic differences: sequence learning was eliminated in the lesioned animals while SRP was not affected. This suggests that some forms of visual plasticity may require an intact hippocampus and is particularly important given recent investigations revealing strong interactions between the hippocampus and V1 (Diamanti et al., 2021; Fournier et al., 2020). Their study also used visually-evoked potentials to detect the presence of a negative prediction error when the second element

was omitted by extending the duration of the first sequence element. Our results replicate theirs with multi-unit data and provide additional insights into potential mechanisms.

In a final study, Colleen Gillon et al. analyzed learning-related changes associated with passive exposure to image sequences (Gillon et al., 2021). They found significant changes in neural activity due to experience, including differences between feedback and feedforward layers of V1, and positive prediction errors in response to unexpected stimuli. In addition, those neurons that responded most strongly to an unexpected stimulus in one imaging session were found to have the largest learning-related changes in subsequent sessions. Thus, unexpected stimuli may preferentially support learning. These results are also consistent with theories of efficient coding and predictive processing. However, it is important to note their conclusions hinged on evidence for a positive prediction error. These can be generated in simple feedforward circuits, for example ones that perform principal components analysis, and are consistent with a wide variety of models of cortical function (Keller & Mrsic-Flogel, 2018). Negative prediction errors, especially those generated in the complete absence of sensory stimulation, are much more difficult to explain and direct evidence for them is minimal [see (Fiser et al., 2016; Keller et al., 2012)].

To conclude, we would like to emphasize several key points. If the visual system does indeed perform some kind of predictive processing, then its predictions must be based on the environmental statistics that the brain encounters. Therefore, responses to the sequence stimulus in naïve mice, on Day 1, can themselves be thought of as prediction errors to an

unexpected stimulus. The sequence stimulus violates known natural environmental statistics and ought to drive plasticity processes in the visual system, which ultimately alter those expected statistics. The evidence collected thus far suggests the mouse visual system does not learn a complete, abstract representation of the sequence, but rather slowly adapts to the new environmental statistics.

In addition, though prediction errors are often emphasized as crucial experimental indicators of predictive processing, they can also be thought of as one example of a more fundamental computational goal, namely information compression (Creutzig & Sprekeler, 2008; Palmer et al., 2015; Srinivasan et al., 1982; Tishby et al., 2000). In this view, incoming sensory data is passed through a bottleneck that reduces the entropy of transmitted data, across both space and time. This is the express purpose of the original predictive coder from telecommunications engineering, which reduces spatial and temporal autocorrelations in the transmitted code (Elias, 1955). Thus, a variety of experimental results, including a reduction in temporal autocorrelations in LGN (Dan et al., 1996) and increasing sparsity in V1 (Failor et al., 2021; Olshausen & Field, 1996; Van Vreeswijk, 2001), all point toward some form of efficient coding or predictive processing (Chalk et al., 2018; Niven & Laughlin, 2008; Pitkow & Meister, 2012). A truly efficient code ought to compress information in both space and time. Given our findings, and the convergence of overlapping results in other recent studies, future research might utilize these and other types of sequential stimuli, in conjunction with theoretical tools such as MbTDR or the

information bottleneck (Palmer et al., 2015; Tishby et al., 2000), as a reliable way to probe unsupervised learning and sensory processing in the nervous system.

## **Methods**

### *Animals*

Male and female C57BL/6 mice (Charles River Laboratories) were housed with same-sex littermates (four mice per cage) on a standard 12-hour light/dark cycle and provided food and water *ad libitum*. All experiments were performed during the mouse's light cycle at approximately the same time of day (~10am-2pm). Mice were 66.8 +/- 0.5 postnatal days old at experimental start (minimum of 61 days and maximum 73 days). Of 56 mice that were considered for further analysis (see Criteria for Multi-Unit Inclusion below), 27 were male and 29 female. There were no significant differences in spontaneous or evoked firing rates between units from male and female mice (Mann-Whitney U test:  $p=0.17$  and  $p=0.63$ , respectively). There were also no clear qualitative differences in stimulus-evoked responses, so all analyses were performed ignoring the male/female distinction. All procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of Boston University.

### *Experimental Design*

Our goal in designing the experiment was to access neural data across the learning process, under the restriction that we could not expect to record from the same units across days. To do so, we created a novel randomized Train/Test protocol (see Figure 5), where mice were selected to see either a Training stimulus or a Test stimulus on any of four

experimental days. The randomized selection process acted as a filter, whereby a mouse who saw the Test stimulus was removed from the process and done with the experiment. The overall result of this filter was that ~25% of mice saw the Test stimulus on Day 1 (the naïve group), ~25% on Day 2, ~25% on Day 3, and ~25% on Day 4. A mouse who saw the Test stimulus on Day 4 would have seen the Training stimulus on Days 1-3, while a mouse who saw the Test stimulus on Day 2 only saw the Training stimulus on Day 1. In order to achieve the desired uniform distribution across days, the following probabilities were used to select mice for Train/Test: Day 1, 75% see the Training stimulus / 25% see the Test stimulus; Day 2, 66.667% Training / 33.333% Test; Day 3, 50% Training / 50% Test; Day 4, 100% Test. Thus, the population of mice remaining in the experiment decreases across days, until they have all seen the Test stimulus. All data analyses were then performed only on neural responses to the Test day stimuli.

### *Electrode Implantation*

Mice were anesthetized with an intraperitoneal injection of 50 mg per kg ketamine and 10 mg per kg xylazine and prepared for chronic recording as described previously (Cooke & Bear, 2010; Frenkel et al., 2006; Gavornik & Bear, 2014). To facilitate head restraint, a steel headpost was affixed to the skull anterior to bregma using cyanoacrylate glue. Small (<0.5 mm) burr holes were drilled over binocular primary visual cortex (3.1 mm lateral from lambda) and a custom-made recording bundle (20- $\mu$ m outer diameter tungsten H-Formvar wire, California Fine Wire Company) with 6 wires tightly wound together was placed 450 $\mu$ m below the cortical surface (Layer 4 / Layer 4/5 border). All recordings were performed in the left hemisphere, as our setup included an infrared camera directed at the

right eye. A reference electrode (silver wire, A-M systems) was placed below the dura at ~1mm lateral from bregma in the same hemisphere as the bundle in V1. All electrodes were rigidly secured to the skull using cyanoacrylate glue. Dental cement was used to enclose exposed skull and electrodes in a protective head cap. Buprenex (0.1 mg per kg) was injected subcutaneously for postoperative pain amelioration. All surgeries were performed around postnatal day 60. Mice were monitored for signs of infection and allowed at least 48 hours of recovery before habituation to the recording and restraint apparatus.

#### *Data Recording*

Each multi-unit bundle contained 4-6 channels per mouse, depending on the custom build process. Each channel was electroplated using a Nano-Z system to yield a final impedance of ~210+/-10 kOhms at 1000 Hz. Data from each channel was amplified and digitized using an Open Ephys recording system. Spiking activity was digitized at 30-kHz and bandpass filtered from 300-6000 Hz using a causal 2<sup>nd</sup> order Butterworth filter (implemented in Open Ephys). This bandpass data was extracted from a binary storage format and analyzed using custom software written in MathWorks MATLAB (see below). Mice were head-fixed and awake during all recordings.

#### *Criteria for Multi-Unit Inclusion*

The bandpass data from a single mouse and session,  $\mathbf{Z} \in \mathbb{R}^{(\text{recording time}) \times (\text{channels})}$ , was first transformed to its singular value decomposition ( $\mathbf{Z} = \mathbf{USV}^T$ ), yielding an orthogonal set,  $\mathbf{U}$ , and placing movement artifacts into a single dimension (invariably the first dimension, with the most variance). Each column of the matrix  $\mathbf{U}$  was then reduced to a

set of timestamps representing multi-unit spike times. This was done by identifying all those times the voltage trace crossed a negative-going threshold of 4 robust standard deviations from the mean (the robust standard deviation was estimated as 1.4826 times the median absolute deviation). A given effective channel (column of  $\mathbf{U}$ ) was then selected for further analysis if it met the following criteria:

- 1) Had a mean evoked firing rate greater than 1 Hz (with a sequence of 4 elements and 150ms / element, 1 Hz is less than 1 spike per trial)
- 2) Passed a statistical test for visual responsiveness
- 3) Had no more than  $R=0.258$  stimulus-evoked correlation with all other simultaneously recorded channels ( $R=0.258$  is equivalent to  $R^2=0.067$ , which was our estimate of the proportion of explainable variance in our experimental setup)
- 4) Showed no signs of artifacts in spike raster

The test for visual responsiveness consisted of two statistical tests, and each effective channel had to pass at least one of the tests with a Bonferroni-corrected p-value  $\leq 0.01/2$ .

The visual stimulus had four visual elements and a fifth “element” driven by the offset response as the screen switched back to gray. The first test was a two-sided KS test, which compared the evoked distribution of inter-spike intervals (ISIs) in a window spanning all five visual elements to a null distribution of comparable size drawn from periods with no visual stimulation. This test captures multi-unit channels with evoked distributions unique from their spontaneous distributions. The second test was a chi-square difference of deviance test. For each effective channel, we fit two Poisson GLMs to model the evoked peri-stimulus time histogram (PSTH). The first model, null, had 2 parameters such that the evoked PSTH was restricted to be a linear function of the time elapsed in the sequence. The second model, full, had  $1+T$  parameters for 1 baseline parameter and  $T$  time bins ( $T=30$

for 25ms bins and  $150\text{ms} \times 5 = 750\text{ms}$  total time in a sequence). We then compared the difference of their deviances to a chi-square CDF with degrees of freedom given by the difference between the number of parameters in the two models (T-1). This test captures multi-unit channels with “interesting” PSTHs (those with more wiggles than a linear function). The final result of this process preserved 140 distinct multi-unit channels (of a total 347 potential channels) from 56 mice (of 67 mice implanted).

### *Visual Stimulus Presentation*

Visual stimuli were generated using custom software written in MATLAB with the PsychToolbox extension (<http://psychtoolbox.org/>) to control stimulus drawing, timing, and synchronization with the Open Ephys. Stimuli were presented on a screen placed 25cm from the mouse, directly in front to target stimulation of binocular V1. The stimulus consisted of a sequence of 4 visual elements, followed by an inter-sequence gray period. Each visual element persisted on screen for 150ms, while the gray period was randomized from a uniform distribution between 1 and 2 seconds. The total duration of a sequence was  $4 \times 150\text{ms} = 600\text{ms}$ . Each visual element was a full-screen oriented sinusoidal grating (0.05 cycles per degree), shown at 75% contrast. For consistency, we used the same orientations across mice: A-  $75^\circ$ , B- $120^\circ$ , C- $35^\circ$ , D-  $160^\circ$ , E-  $10^\circ$ ,  $x$ - Uniform ( $60^\circ, 180^\circ$ ). Grating stimuli were gamma corrected to insure a linear gradient and constant total luminance. Stimuli were also corrected to ensure the gratings had a constant spatial frequency in retinotopic coordinates. This consisted of a spherical transformation, as described in (Marshall et al., 2011), supplemental information. During experiments, animal handling

consisted of placing each mouse into the head-fixed presentation apparatus. The handler was unaware of whether the mouse would see the Test or Training stimulus.

*Training Stimulus* – If a mouse was randomly selected to see the Training stimulus, it saw 200 presentations of the sequence ABCD. Each sequence was presented in four blocks of 50 presentations, with each presentation separated by a gray screen drawn from a randomly-selected uniform interval of 1 to 2 seconds, and each block separated by 60 seconds. The total time of visual stimulation was  $200 \times 600\text{ms} = 2$  minutes and the total time in the apparatus  $\sim 10$  minutes.

*Test Stimulus* – If a mouse was randomly selected to see the Test stimulus, it saw 600 presentations of sequences from one of four conditions:  $AxCD$ ,  $Ax \rightarrow D$ ,  $ExCD$ ,  $Ex \rightarrow D$ .  $x$  is a randomly oriented element and  $x \rightarrow$  means the second element was held on screen for 300ms, such that nothing changed at the B-C transition time observed during Training. As before, each sequence was presented in blocks of 50 presentations, but now for a total of 12 blocks. The order of the presentations and the orientation of the second element,  $x$ , was randomized for each Test Day, i.e., all mice seeing the Test stimulus on Day 1 saw the same set of sequences in the same order, while mice on Day 2 saw a different set in a different order (this allowed for the creation of pseudo-trials with data grouped across all mice recorded on the same Test Day). The total time of visual stimulation was  $600 \times 600\text{ms} = 6$  minutes and the total time in the apparatus  $\sim 30$  minutes.

### *Data Analysis*

All analyses aside from the explainable variance estimation were performed only on Test day data, so there were no repeated measures. In general, we had  $N=140$  multi-unit channels from the Test days, with 37 units on Day 1, 31 on Day 2, 36 on Day 3, and 36 on Day 4. Therefore, if we performed a statistical test comparing Day 1 and Day 4, there would be  $37+36=73$  data points, assumed independent and identically distributed within a test day group as most units came from different mice and those that came from the same mouse were required to have a very low pairwise evoked correlation (see Criteria for Multi-Unit Inclusion above). Any figure depicting a bootstrap confidence interval was computed as a bootstrap pivotal interval (Wasserman, 2004).

#### *Pre-Processing*

Firing rate was first computed as a spike count in 25-ms bins, with no smoothing. Each unit's firing rate was then normalized by subtracting out the spontaneous baseline firing rate and dividing by the evoked standard deviation to yield zero baseline mean and unit variance. This z-scored firing rate was used in all analyses and is displayed in all figures (the notation *FR (AU)* used in multiple figures refers to this z-scored firing rate, while the notation *FR Modulation* is reserved for shared bases computed from MbTDR, see below). Dot plots (as in Figure 7b) were computed directly from the z-scored firing rates, averaged in 50-ms windows as described in the main text. In the case of the firing rate difference plots (Figure 8c-d and Figure 9c-d), we first computed for each unit separately z-scored PSTHs for the two stimuli being compared (e.g., A/ExCD and A/Ex→D). Next, we subtracted one PSTH from the other to create a difference PSTH. Finally, we averaged the

difference PSTH in a 50-ms window to yield a *Mean FR Difference*. Statistical tests were then performed on the FR differences across Test days (see below).

### *Explained Variance Estimation*

In order to estimate the expected proportion of variance explained for our setup and stimulus, we utilized data recorded on the Training days. On those days, each animal saw 200 presentations of the same stimulus, ABCD. Therefore, the PSTH provides an accurate estimate of the neural variability explainable by the stimulus. Using the same criteria for multi-unit inclusion as used on the Test days, we compiled a set of 156 unique multi-unit channels. For each unit, we fit a Poisson GLM to model the PSTH (as in the Criteria for Multi-Unit Inclusion section, with 1+T parameters for 25-ms time bins and T=30). We then calculated the proportion of neural variance explained by the model:

$$\text{Variance Explained} = 1 - \text{var}(\text{model fit} - \text{data}) / \text{var}(\text{data})$$

The average variance explained from these 156 channels provided our estimate of explainable variance: 6.7 +/- 0.7% (mean +/- SEM). Thus, a substantial proportion of the neural variability cannot be captured by a model that simply accounts for the visual stimulus, as anticipated by other research on the mouse visual system (Stringer, Pachitariu, Steinmetz, Carandini, & Harris, 2019).

### *Model-Based Targeted Dimensionality Reduction (MbTDR)*

Many of the data analyses were performed by fitting a single supervised statistical model to all recorded multi-unit data across the four Test days. The model is known as model-based targeted dimensionality reduction (Aoi et al., 2020; Aoi & Pillow, 2018), which

reduces dimensionality by projecting the data onto low-dimensional subspaces corresponding to different covariates. The covariates in this case are those related to the structure of the experimental design and the Test day stimulus set: condition-independent (a constant value of 1), experimental day, element A or E to start the sequence, the orientation of the second element ( $Angle(x)$ ), whether the second element was held on screen [ $x \rightarrow$ ], trial number, and interaction terms between these. We also included several triplet interaction terms:  $Angle(x) * I(E) * Day$  ,  $Angle(x) * I(x \rightarrow) * Day$  ,  $Angle(x)^2 * I(x \rightarrow) * Day$ ,  $Angle(x)^2 * I(E) * Day$  , none of which were ultimately chosen in the model selection process. Note that all covariates were designed so that the baseline corresponds to the sequence ABCD on Day 1. For example, day is encoded by  $\log(Day)$  so that the covariate for Day 1 is zero, A or E to start the sequence is encoded as an indicator function  $I(E)$ , and the orientation of the second element,  $x$ , is centered so that the orientation of B used during training ( $120^\circ$ ) is equal to zero. For more information on these covariates and how they were encoded, see Supplemental Table 1. The output of the model is comparable to de-mixed principal components analysis (dPCA) (Kobak et al., 2016).

We will now describe the MbTDR model and how it derives from a standard linear regression (for a more complete description, see (Aoi & Pillow, 2018)). For the present dataset, we recorded from  $N = 140$  multi-unit channels and  $M = 600$  trials/unit. The data was binned at 25ms, with  $T = 30$  time bins per trial for a total trial time of 750ms

(4\*150ms for each of the visual elements and 150ms for the offset response as the screen turns back to gray). Each unit's firing rate was first z-scored as described above.

Assuming the covariates vary by trial only (not within a trial) and we wish to infer how neural activity evolves over time on individual trials, we might create a linear model for one unit,  $n$ , on one trial,  $m$ :

$$\mathbf{y}_m^{(n)} \sim \text{Normal}(\mathbf{y}_m^{(n)} \mid b_n \mathbf{1}_T + x_{m,1} \boldsymbol{\beta}_1^{(n)} + \dots + x_{m,p} \boldsymbol{\beta}_p^{(n)}, \lambda_n \mathbf{I}_T)$$

where  $\mathbf{y}_m^{(n)} \in \mathbb{R}^T$  is the firing rate of unit  $n$  on all  $T$  timepoints of trial  $m$ , each  $\boldsymbol{\beta}_1^{(n)}, \dots, \boldsymbol{\beta}_p^{(n)} \in \mathbb{R}^T$  is a regression coefficient vector for unit  $n$  that covers the  $T$  timepoints of a single trial,  $b_n$  is a unit-specific baseline firing rate,  $\lambda_n$  is a unit-specific variance, and  $x_{m,p} \in \mathbb{R}$  is the value of the  $p$ -th covariate on trial  $m$ . Each column of  $\boldsymbol{\beta}_p^{(n)}$  is therefore the PSTH of the neuron projected onto the subspace spanned by the  $p$ -th covariate. If we assume the baseline firing,  $b_n$ , variance,  $\lambda_n$ , and covariate-specific bases,  $\boldsymbol{\beta}_p^{(n)}$ , are constant across trials, then we could use a least-squares approach to learn the parameters.

Recording from  $N$  units, with  $T$  timepoints per trial, and using this linear model, we would find a set of parameters,  $\mathbf{B}^{(n)} = [\boldsymbol{\beta}_1^{(n)}, \dots, \boldsymbol{\beta}_p^{(n)}]$ , for each unit, plus the baseline firing rates and noise variances. For the entire dataset, that is  $2N + TPN$  parameters ( $\sim 140,000$  in the present case). With the limitations on recording time in animal experiments, and the variability of neural data, collecting enough trials to reliably estimate so many parameters would be quite difficult for more than a few covariates.

The key idea of MbTDR is to reduce the total number of parameters by attempting to discover a low-rank representation for the regression coefficients corresponding to each covariate. That is, find  $\mathbf{S}^{(p)} \in \mathbb{R}^{T \times R_p}$  and  $\mathbf{W}^{(p)} \in \mathbb{R}^{R_p \times N}$  such that:

$$[\boldsymbol{\beta}_p^{(1)}, \dots, \boldsymbol{\beta}_p^{(N)}] = \mathbf{S}^{(p)} \mathbf{W}^{(p)}$$

where  $R_p$  is the rank of the representation for covariate  $p$ . This low-rank representation effectively discovers a unique set of shared time-varying basis functions,  $\mathbf{S}^{(p)}$ , for each covariate, which are modulated independently for each unit by the latent factors,  $\mathbf{w}_n^{(p)} \in \mathbb{R}^{R_p}$ . The shared basis functions are a low-dimensional representation of high-dimensional neural trajectories. With this adjustment, the total number of parameters can be dramatically reduced to  $2N + P\bar{R}(N + T)$ , where  $\bar{R}$  is the average rank ( $\sim 4,500$  in the present case).

Under MbTDR, the full likelihood of all recorded data is:

$$P(\mathbf{Y}|\mathbf{X}, \mathbf{S}, \mathbf{W}, \mathbf{b}, \boldsymbol{\lambda}) = \prod_{n=1}^N \prod_{m \in \{\mathbf{M}_n\}} \text{Normal}(\mathbf{y}_m^{(n)} | b_n + \sum_{p=1}^P x_{m,p}^{(n)} \mathbf{S}^{(p)} \mathbf{w}_n^{(p)}, \lambda_n \mathbf{I}_T)$$

where now  $\mathbf{Y}$  is the set of all  $\mathbf{y}_m^{(n)}$ ,  $\mathbf{X}$  is the set of all trial covariates,  $\mathbf{S}$  and  $\mathbf{W}$  are the sets of all  $\mathbf{S}^{(p)}$  and  $\mathbf{W}^{(p)}$ ,  $\mathbf{b}$  is the vector of all  $b_n$ ,  $\boldsymbol{\lambda}$  is the vector of all  $\lambda_n$ , and  $\mathbf{M}_n$  is the set of all trials unit  $n$  participated in.

To fit this model, the original authors proposed an expectation conditional maximization either algorithm (ECME), so the optimization is actually over a marginal likelihood:  $P(\mathbf{Y}|\mathbf{X}, \mathbf{S}, \mathbf{b}, \boldsymbol{\lambda})$ . We wrote custom scripts to perform ECME for initialization and then

passed the result to a trust-region algorithm (MATLAB *fminunc*) to directly maximize the marginal likelihood. To compute the marginal likelihood, all of the “unit factors” are given a  $Normal(w_{r,n}^{(p)}|0, 1)$  prior and then integrated out. This conjugate Gaussian prior, along with the independence of the unit factors, dramatically decreases the complexity of both the derivation and the optimization problem. The optimal set of parameters are then maximum marginal likelihood estimates  $\hat{\mathcal{S}}, \hat{\mathbf{b}},$  and  $\hat{\lambda}$ , along with the posterior mean for  $\mathbf{W}$ , computed from  $P(\mathbf{W}|\mathbf{X}, \mathbf{Y}, \hat{\mathcal{S}}, \hat{\mathbf{b}}, \hat{\lambda})$ .

To determine the optimal rank of the coefficient matrix for each covariate, we performed a greedy forward stepwise algorithm that sought to minimize the Akaike Information Criterion (AIC) (again following (Aoi & Pillow, 2018)). To begin, we fit the model independently  $P$  times (with the  $p$ -th covariate set to rank 1 and all others to 0), and then chose the covariate for which the AIC was minimized,  $p_{min}[1]$ . With the rank of covariate  $p_{min}[1]$  now equal to 1, and all others still set to zero, we repeated the process by again fitting the model  $P$  times, iteratively adding 1 to the rank of each covariate, and finding the minimum AIC across the  $P$  iterations,  $p_{min}[2]$ . Now, with  $p_{min}[1]$  set to 1 and  $p_{min}[2]$  set to 1, the process continued (if  $p_{min}[1]=p_{min}[2]$ , then that covariate would be at 2 with all others still equal to 0). This was repeated until the global minimum AIC was achieved. 90% of the data (540 trials / unit) was used for this fitting procedure. The remaining 10% (60 trials / unit) was excluded from all preliminary analysis and model fitting, and reserved for decoding and for estimation of the held-out proportion of variance explained by the model.

With the optimal model parameter estimates, we can evaluate the model fit for each unit and stimulus type:

$$\mathbf{Model} = \mathbf{1}_T \mathbf{b}^T + \sum_{p=1}^P x_p \widehat{\mathcal{S}}^{(p)} \widehat{\mathcal{W}}^{(p)}$$

Where  $\mathbf{1}_T$  is a vector of  $T$  ones,  $x_p$  is a scalar representing the value of the  $p$ -th covariate for the desired trial type, and  $\mathbf{Model} \in \mathbb{R}^{TxN}$  are the model fits. To evaluate the model response to the training sequence ABCD, we simply set the covariates appropriately (i.e.,  $I(x \rightarrow) = 0, I(E) = 0, Angle(x) = 0$ ). To determine the orientation tuning of each unit to the second element when A starts the sequence, we can set  $I(x \rightarrow) = 0$  and  $I(E) = 0$ , and then change the value of  $Angle(x)$  across its range from 60 to 180 degrees. This yields a PSTH for each unit and each angle, from which we then extract the maximal response in a window from 50-150ms after the onset of the second element (200-300ms after the onset of the entire sequence). The set of maximal responses across possible angles is the orientation tuning curve for one unit. Taking the angle for which the orientation tuning curve is a maximum gives the peak tuning angle for that unit.

For visualization purposes and to compare the low-rank representation across conditions, as in Figures 7-9, we performed the following singular-value decomposition:

$$\mathbf{UTV}^T = [\mathbf{Model}_{ABCD}^T, \mathbf{Model}_{AB \rightarrow D}^T, \mathbf{Model}_{EBCD}^T, \mathbf{Model}_{EB \rightarrow D}^T]^T$$

Thus,  $\mathbf{U} \in \mathbb{R}^{4Tx(\text{total rank})}$ , and every column of  $\mathbf{U}$  represents a set of low-dimensional neural trajectories that can be reliably compared across conditions (the notation FR Modulation, used on the y-axis of multiple figures, refers to the columns of  $\mathbf{U}$ , which are

normalized by the SVD procedure). The first  $T$  elements of the first column of  $\mathbf{U}$  are the response to  $ABCD$  in the dimension with the greatest singular value, while the next  $T$  are the response to  $AB \rightarrow D$  in that same dimension. To make comparisons across days (as in Figure 7c-f), we restricted the set of units used in the model fit such that, for example,  $\mathbf{Model} \in \mathbb{R}^{T \times N_{Day 1}}$ . Next, we computed the singular-value decomposition with that restricted set of model PSTHs, and then repeated for each Test day.

The variance accounted for by each component,  $i$ , is computed with  $\mathbf{T}$ :

$$VarExp_i = \frac{\mathbf{T}_{ii}^2}{\sum_{j=1}^{Rank} \mathbf{T}_{jj}^2}$$

This is the variance explained by each component in the singular-value decomposition of the model fits: it is variability in the PSTHs rather than the neural data, so it represents signal rather than noise variance.

### *Decoding Pseudo-Trial Covariates*

In Figure 6, we showed several examples of stimulus decoding. MbTDR allows for decoding by invoking Bayes' rule:

$$P(\mathbf{X}|\mathbf{Y}, \widehat{\mathbf{S}}, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}}) \propto P(\mathbf{Y}|\mathbf{X}, \widehat{\mathbf{S}}, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}})P(\mathbf{X})$$

In our experiment, all covariates were randomly generated from uniform distributions, so  $P(\mathbf{X})$  is a constant that can be ignored. Even though the base model is linear in its covariates, the optimal decoder is non-linear, so we again used a trust-region algorithm to maximize  $P(\mathbf{X}|\mathbf{Y}, \widehat{\mathbf{S}}, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}})$  with respect to the covariates  $\mathbf{X}$  on individual pseudo-trials. In particular, the set of covariates are created from a ‘‘fundamental’’ set, e.g.,  $Angle(x)^2$ ,

and  $I(x \rightarrow) * Day$  come from the more fundamental  $Angle(x)$ ,  $Day$ , and  $I(x \rightarrow)$ . A linear decoder would provide different estimates for each covariate, e.g.,  $\hat{x}$  and  $\widehat{x^2}$  such that  $(\hat{x})^2 \neq \widehat{x^2}$ . The non-linear decoder, however, outputs point estimates of the fundamental set of covariates, given the neural data and the optimal model. Those estimates include: experimental day, angle of the second sequence element ( $Angle(x)$ ), trial number, a posterior probability for the indicator of E starting the sequence ( $\mathbb{E}[I(E)]$ ), and a probability for the indicator of the third element having been omitted ( $\mathbb{E}[I(x \rightarrow)]$ ).

Pseudo-trials are not simultaneously recorded but are created by grouping together single-trial data across all neurons recorded on the same Test day. For example, 37 units from 15 mice were recorded on Day 1. Those mice all saw the same stimuli in the same order, so each trial is comparable across mice. To create a Day 1 pseudo-trial, we selected one of the held-out trials and created a joint log likelihood during that trial by summing over all 37 units. We then found the set of fundamental covariates,  $\widehat{\mathbf{x}}_m \in \mathbb{R}^5$ , that maximized the joint log likelihood across those units for a given held-out trial,  $m$ :

$$\widehat{\mathbf{x}}_m = \mathbf{arg\,max}_x \sum_n \log \{P(\mathbf{y}_m^{(n)} | \mathbf{x}, \widehat{\mathbf{S}}, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n)\}$$

*Decoding Time*

In Figure 11, we performed time decoding. Data from each 750-ms trial was binned at 25ms, creating a total of 30 bins per trial. The problem of decoding time on a single trial is equivalent to asking which bin from the model fit is most consistent with the data. We can

create a scalar normalized firing rate for one unit, timepoint, and trial:  $y_{t,m}^{(n)}$ . Then, the likelihood of the data in that one time bin is:

$$P\left(y_{t,m}^{(n)} \mid \mathbf{x}_m^{(n)}, \widehat{\mathbf{S}}_t, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right)$$

We take the model fit,  $\widehat{\mathbf{S}}_t$ , at the same bin and have assumed that the trial covariates,  $\mathbf{x}_m^{(n)} \in \mathbb{R}^5$ , are their true values, i.e., those observed by the mouse on that trial.

If we were to decode time with the data from one unit, the posterior distribution over potential model time bins,  $\tau$ , for a specific bin of neural data,  $t$ , would be:

$$P\left(\tau \mid y_{t,m}^{(n)}, \mathbf{x}_m^{(n)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right) = \frac{P\left(y_{t,m}^{(n)} \mid \mathbf{x}_m^{(n)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right)P(\tau)}{\sum_{l=1}^T P\left(y_{t,m}^{(n)} \mid \mathbf{x}_m^{(n)}, \widehat{\mathbf{S}}_l, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right)P(l)}$$

Because the prior distribution over time bins,  $P(\tau)$ , is uniform, the optimal decoded bin considering all units recorded on a given Test day, trial, and true time bin ( $t$ ) is:

$$\hat{t}_{t,m} = \mathbf{arg\,max}_{\tau} \sum_n \log \{P\left(y_{t,m}^{(n)} \mid \mathbf{x}_m^{(n)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right)\}$$

where we vary  $\tau$  from 1 to  $T$ , selecting the optimal time bin as the one that maximizes this joint log likelihood. This is the time decoder we used in the Results section, which might be called a “conditional” decoder because the time-bin posterior is conditional on the true values of the covariates,  $\mathbf{x}_m^{(n)}$ . To quantify the accuracy of this decoder, we computed a “soft accuracy” on the decoded time bin for each day:

$$\text{Soft Accuracy} = \frac{100\%}{M * T} \sum_{m=1}^M \sum_{\tau=1}^T I(|\hat{t}_{\tau,m} - \tau| \leq 1)$$

The soft accuracy allows the decoder to be off by one time bin in either direction, so there is a 3/30 chance of guessing correctly for the central bins (bins 2-29) and a 2/30 chance on the ends (bin 1 and bin 30), which averages to 9.77%. A 95% confidence interval on chance soft accuracy (with  $M = 60$  trials and  $T = 30$  time bins on a given Test day) is [8.4, 11.3]%. Therefore, on a given day, a soft accuracy greater than about 12% performs significantly better than chance. In order to determine whether decoding accuracy improved significantly with Training, we performed a permutation test described in the next section. Assuming only one time bin, the 95% confidence interval is [2.1, 17.5]% for 60 trials.

To test the robustness of the result, and in order to more accurately capture the uncertainty faced by the nervous system, we also tried a “marginal” decoder that integrates out the trial covariates. For one unit:

$$P\left(y_{t,m}^{(n)} \mid \widehat{\mathbf{S}}_t, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right) = \int P\left(y_{t,m}^{(n)} \mid \mathbf{x}_m^{(n)}, \widehat{\mathbf{S}}_t, \widehat{\mathbf{W}}, \hat{b}_n, \hat{\lambda}_n\right) P\left(\mathbf{x}_m^{(n)}\right) d\mathbf{x}_m^{(n)}$$

To compute the integral across all units, we used a Monte Carlo sampling procedure. First, we randomly drew a set of trial covariates,  $\mathbf{x}^{(iter)}$ , from their respective *a priori* uniform distributions:

$$\begin{aligned} Day &\sim \text{Discrete Uniform}(1,4) \\ Angle &\sim \text{Uniform}(60,180) \\ Trial &\sim \text{Discrete Uniform}(1,600) \\ I(E) &\sim \text{Bernoulli}(0.5) \\ I(x \rightarrow) &\sim \text{Bernoulli}(0.5) \end{aligned}$$

Next, we computed the joint log likelihood of the data from one time bin given those sampled covariates:

$$\log \{P(\mathbf{y}_{t,m} | \mathbf{x}^{(iter)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}})\} \triangleq \sum_n \log \{P(y_{t,m}^{(n)} | \mathbf{x}^{(iter)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}_n, \widehat{\boldsymbol{\lambda}}_n)\}$$

Then, we repeated this process for each of the  $\tau = 1, \dots, T$  bins of the model, ultimately computing a conditional posterior distribution over bins:

$$P(\tau | \mathbf{y}_{t,m}, \mathbf{x}^{(iter)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}}) = \frac{P(\mathbf{y}_{t,m} | \mathbf{x}^{(iter)}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}})P(\tau)}{\sum_{l=1}^T P(\mathbf{y}_{t,m} | \mathbf{x}^{(iter)}, \widehat{\mathbf{S}}_l, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}})P(l)}$$

Finally, we sampled from this distribution (which is multinomial), yielding one sample from the marginal posterior:

$$P(\tau | \mathbf{y}_{t,m}, \widehat{\mathbf{S}}_\tau, \widehat{\mathbf{W}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\lambda}})$$

Repeating this process for 1000 iterations, the optimal time bin is then the  $\tau$  with the greatest number of posterior samples. The results were comparable to those with the conditional decoder, though accuracy dropped somewhat. Bootstrap 95% confidence intervals on the soft accuracy were: Day 1- [22.0%,25.6%]; Day 2- [21.6%,25.1%] ( $p = 0.72$ , two-sided difference in soft accuracy from Day 1 permutation test); Day 3- [24.8%,28.4%] ( $p = 0.037$ ); Day 4- [26.1%,29.8%] ( $p = 2.5e-3$ ).

Using MbTDR, we also simulated data from 1000 neurons (250 per day), matching Test days for noise variance. To simulate data, we discovered kernel-density estimates for the distributions of unit factors ( $\mathbf{W}$ ) unique to each Test day (see, for example, Supplemental Figure 3a), and for the noise variance ( $\boldsymbol{\lambda}$ ) irrespective of Test day. Next, we randomly drew from those distributions, creating new units for each Test day whose values for the unit factors matched the corresponding Test day's actual neural data. With 60 trials of simulated data, we then performed time decoding exactly as we did for the real data. With more units,

decoding soft accuracy improved significantly to ~60% (up from ~27% in the neural data, where we had ~35 neurons per Test day): Day 1- 59.4%; Day 2- 59.1%; Day 3- 64.7%; Day 4- 64.4%.

### *Non-Parametric Statistical Tests*

In the Results section, we performed several non-parametric permutation tests to determine the significance of median differences, mean differences, and correlations between evoked firing rates, firing rates & held-out explained variance, decoded covariates & ground-truth covariates, etc. In each case, the null distribution of our test statistic may not follow a known distribution, so we resorted to non-parametric tests (controlling for multiple comparisons when applicable). For the correlation, we performed an approximate two-sided permutation test using Spearman's rho as the test statistic. For the example of firing rate and explained variance, we have a dataset:  $\{FR_n, expVar_n\}_{n=1}^{140}$  for the complete set of multi-unit channels. We first calculated the observed test statistic by Spearman's rank correlation:  $\rho_{obs}$ . Next, we permuted the dataset so as to break dependence between the two variables, e.g., now the firing rate from the 1<sup>st</sup> unit is associated with the explained variance of the 17<sup>th</sup> unit. Then, we calculated a permuted correlation:  $\rho_{perm}$ . We repeated this process  $L$  times by randomly permuting the dataset and calculating new values for  $\rho_{perm}^{(l)}$ . The p-value is then:

$$p = 1 - \frac{1}{L} \sum_{l=1}^L I(|\rho_{perm}^{(l)}| \leq |\rho_{obs}|)$$

In every case, we performed  $L = 1e6$  permutations, such that an estimated p-value of zero implies the true p-value is less than  $1e-6$ .

The permutation test on the decoder soft accuracies (Figure 11c) was performed as follows. For each Test day, we had 60 trials and 30 time bins per trial to decode. Thus, the decoder output on a given Test day can be written as a matrix,  $\hat{\mathbf{t}}_{Day} \in \mathbb{R}^{30 \times 60}$ . After computing the soft accuracy from this matrix, we calculated a difference in soft accuracy from Day 1 statistic:

$$DSA_{obs}(x, 1) = \text{Soft Accuracy}(\hat{\mathbf{t}}_{Day\ x}) - \text{Soft Accuracy}(\hat{\mathbf{t}}_{Day\ 1})$$

Next, we computed a null distribution for this statistic using random permutations of the data. The null distribution is the distribution of soft accuracy differences, were the soft accuracy the same on both days. For example, if we have decoder outputs from Day 1 and Day 3,  $\hat{\mathbf{t}}_1$  and  $\hat{\mathbf{t}}_3$ , then the full dataset can be written:  $\hat{\mathbf{t}}_{13} \triangleq [\hat{\mathbf{t}}_1, \hat{\mathbf{t}}_3] \in \mathbb{R}^{30 \times 120}$ . The permuted dataset randomly shuffles values in each row of this matrix (preserving the time bin information), and then re-isolates decoder output matrices to compute the permuted soft accuracy difference:

$$[\hat{\mathbf{t}}_1^{(perm)}, \hat{\mathbf{t}}_3^{(perm)}] \leftarrow \hat{\mathbf{t}}_{13}^{(perm)}$$

$$DSA_{perm}(3,1) = \text{Soft Accuracy}(\hat{\mathbf{t}}_3^{(perm)}) - \text{Soft Accuracy}(\hat{\mathbf{t}}_1^{(perm)})$$

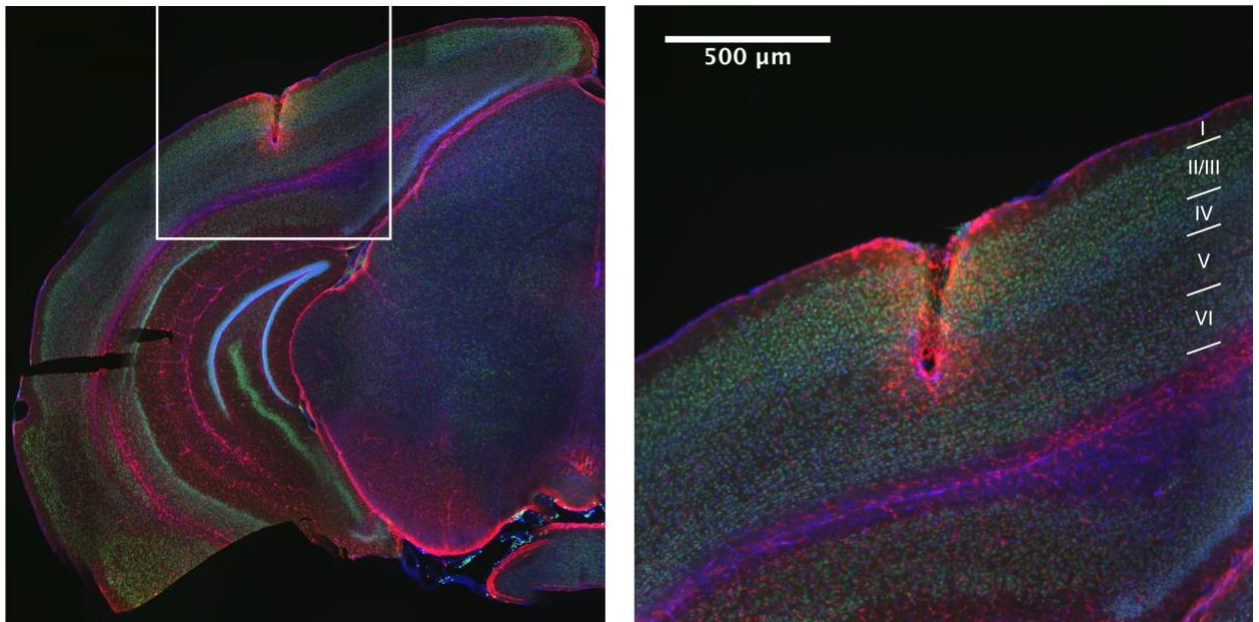
The test's p-value is then:

$$p = 1 - \frac{1}{L} \sum_{l=1}^L I(|DSA_{perm}^{(l)}(x, 1)| \leq |DSA_{obs}(x, 1)|)$$

A similar procedure was used in all of the permutation tests from Figures 7-9, with  $L = 1e6$  permutations.

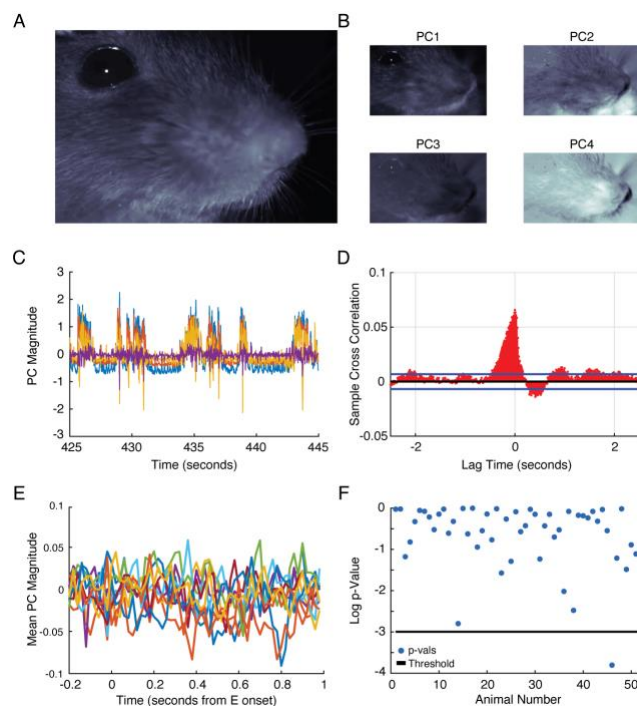
See Supplemental Table 2 for a summary of all the statistical tests used, along with summary statistics of the data.

### Supplemental Information and Figures



#### Supplemental Figure 1: Post-Mortem Histology

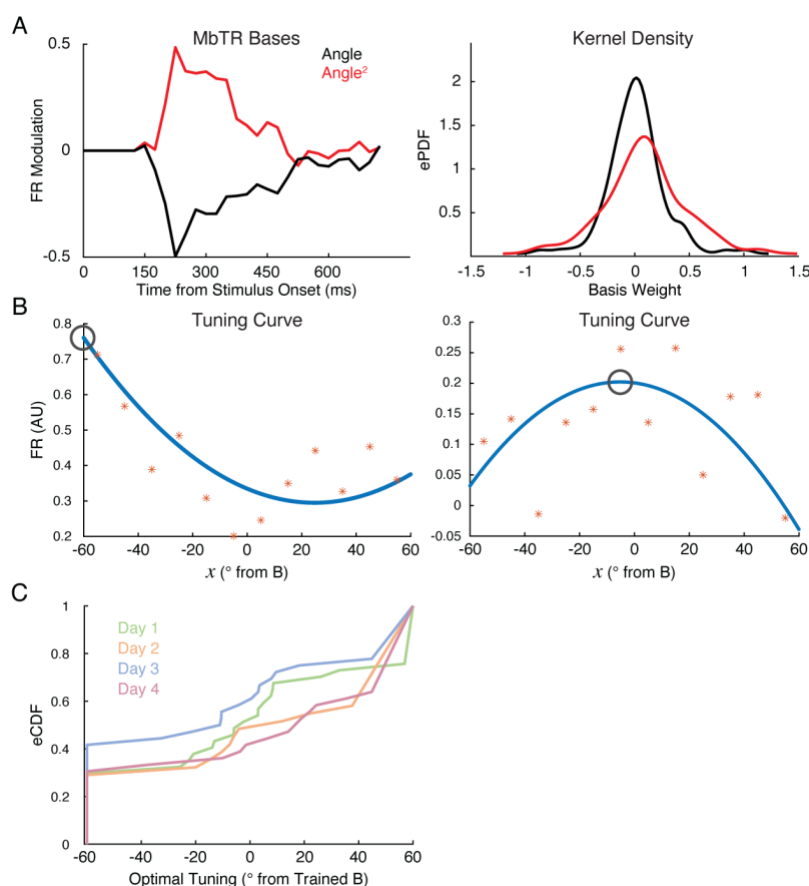
Immunofluorescence image of 50-micron thick coronal section through V1 showing representative electrode placement in binocular Layer 4, near Layer 4/5 border. Blue labeling represents nuclei stained by Hoechst, green represents neuronal nuclei stained by NeuN, and red represents astrocytes stained by GFAP (glial fibrillary acidic protein).



**Supplemental Figure 2: Absence of Stimulus-Aligned Movement**

**A.** Example image of the right eye and whisker pad of a mouse in our apparatus captured at 50Hz using an infrared camera. The cheek and nose move in and out of focus depending on the mouse's facial position. **B.** First four principal component (PC) eigenvectors of a motion energy movie of the mouse's face. These were produced by performing an online Expectation Maximization (EM) algorithm on the absolute value of the difference between adjacent movie frames. Video was acquired during every recording session for each mouse. The first PC represents global movement of the snout and whisker pad, while the second seems to capture snout movements alone. The third and fourth seem to capture more precise movements of the whisker pad, in opposing directions. Other PCs are less identifiable. **C.** Example traces of the first four principal components from a randomly selected 20-second time window. Periods of general quiescence are punctuated by robust movement epochs. Rhythmic activity during quiescent periods is due to a periodic sniffing behavior. **D.** Sample cross-correlation between the first motion energy PC and the recorded neural data from one visually-responsive multi-unit channel. Blue lines show a 95% confidence interval of the null cross-correlation. Negative lags indicate neural data leads the movement signal. **E.** The mean of each of the first ten PCs, from one mouse, aligned to the onset of the deviant stimuli  $E_x \rightarrow D$  or  $E_x CD$ . Note the PCs were normalized to have unit variance (see y-axis scale in C), so fluctuations from zero are quite small here. **F.** p-values from likelihood ratio tests for deviant-stimulus-aligned movement in the first PC. For each mouse that had movement and co-recorded neural data from visually-responsive units (52 out of 56 animals), we fit two statistical models that captured the deviant-stimulus-aligned movement (modeling the dark blue trace in E) from the first PC. The full model used a set of Gaussian radial basis functions to capture the stimulus-evoked movement assuming normal residuals, while the null model assumed there was no stimulus-evoked movement. 51/52 (98.1%) of the tests failed to reject the null (all those dots above the black p-value threshold line of 0.05). As a control comparison, we randomly shifted the movement data (preserving its statistical structure but decoupling it from the timing of the stimulus), and found

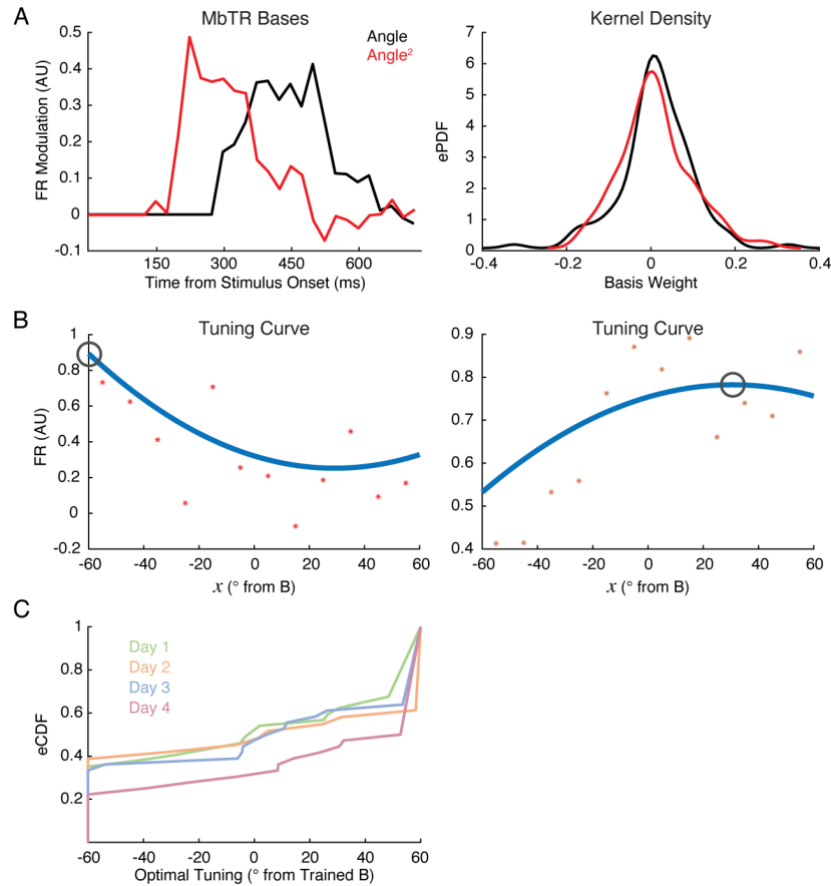
52/52 tests failed to reject the null. Thus, there was no evidence for deviant-stimulus-aligned movement. In the one case with a p-value less than 0.05, the effect size was small (maximum average stimulus-aligned movement about 1/10 of 1 standard deviation). We found a comparable effect when looking at all stimulus-aligned movement (rather than restricting the analysis to deviant stimuli). While movement provides a meaningful explanation of neural variability generally, the weight of the evidence indicated that the mouse's movement was not aligned to the timing of visual stimulation and therefore its effect would average out in any analysis that was locked to stimulus timing. We therefore did not include movement data in our analyses.



**Supplemental Figure 3: Orientation Tuning of the Second Sequence Element**

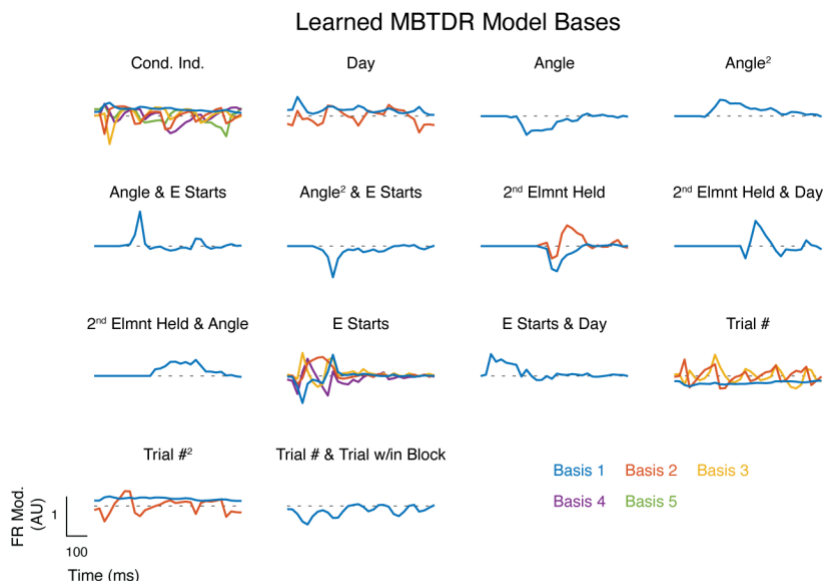
**A.** Left: shared basis functions for the  $\text{Angle}(x)$  (black) and  $\text{Angle}(x)^2$  covariates (red). Right: kernel density estimates of the unit factor distribution for the same covariates. These bases, along with the rest of the model, can be used to create a tuning curve for each unit at different timepoints after stimulus onset. **B.** Example tuning curves (blue) for two units, taken at the time bin with maximal evoked firing (50-75ms after the onset of the second element, i.e., the peak in A). Red stars are the raw data, averaged in bins of 10 degrees. The optimal tuning for each unit is the maximum of these curves (black circle). **C.** Empirical cumulative distribution functions for optimal

tuning on each Test day. These eCDFs were *not* significantly different between naïve and trained groups (two-sided KS test, naïve [Day 1] vs. trained [Days 3 & 4]:  $D=0.106$ ,  $p=0.932$ ).



#### Supplemental Figure 4: Orientation Tuning of the Negative Prediction Error

**A.** Left: shared basis functions for the  $I(x \rightarrow) * \mathbf{Angle}(x)$  (black) and  $\mathbf{Angle}(x)^2$  covariates (red). Note their significant overlap in time. Right: distribution of unit factors for the same covariates. Comparable to Supplemental Figure 3, we created negative-prediction-error orientation tuning curves from these bases for trials when the third element was omitted:  $I(x \rightarrow) = \mathbf{1}$ . **B.** Example tuning curves (blue) for two example units, taken in the late window after the expected onset of C (401-450ms after sequence onset). Red stars are the raw data, averaged in bins of 10 degrees. The optimal tuning for each unit is the maximum of these curves (black circle). **C.** Empirical cumulative distribution functions for optimal tuning on each Test day. These eCDFs were *not* significantly different between naïve and trained groups (two-sided KS test, naïve [Day 1] vs. trained [Days 3 & 4]:  $D=0.152$ ,  $p=0.592$ ). A very similar result holds for the early window ( $D=0.139$ ,  $p=0.697$ ).



**Supplemental Figure 5: MbTDR Bases**

Visualization of the learned bases ( $\mathbf{S}$ ) from the MbTDR fit. All covariates are represented in the model by a matrix  $\mathbf{SW}^T$ , from which we extract a singular-value decomposition  $\mathbf{UTV}^T = \mathbf{SW}^T$ . Depicted are the  $\mathbf{U}$  for each covariate. The bases are all aligned so that  $\text{median}(V_i) > \mathbf{0}$  for each column ( $i$ ) of the basis; upward deflections of the depicted basis therefore imply a positive change in firing rate for the majority of units. From top to bottom and left to right, the bases are: 1,1: condition independent; 1,2: Day; 1,3: Angle; 1,4: Angle Squared; 2,1: Angle & E starts; 2,2: Angle Squared & E starts; 2,3: Second Element Held; 2,4: Second Element Held & Day; 3,1: Second Element Held & Angle; 3,2: E Starts; 3,3: E Starts & Day; 3,4: Trial Number; 4,1: Trial Number Squared; 4,2: Trial Number & Trial Within Block Number.

Fundamental Covariate	Encoding	Start Time	Final Rank
Condition Independent	1	0ms (beginning of the sequence)	5
Day	$\log(\text{Day})$	0ms	2
Angle	$\text{Angle}(x) - \text{Angle}(B) \text{ radians}$	150ms (onset of second element, $x$ )	1
Second Element Held	$I(x \rightarrow)$	300ms (expected onset of third element, $C$ )	2
E starts	$I(E)$	0ms	4
Trial Number	$-\log(\text{trial \#})$	0ms	3
Trial within Block Number	$-\log(\text{trial \# in block})$	0ms	0

### Supplemental Table 1: MbTDR Information

Note that  $I(*)$  is the indicator function. Each basis spanned the time of the trial (750ms, or 30 time bins), except for the angle and second-element-held bases. For those, we forced the basis to be zero at the beginning of the trial to avoid overfitting (the second element, for example, does not come on screen until 150ms into the trial, so its angle could not possibly have an effect until that moment). **Day** was simply the test day, from 1 to 4. The logarithm causes this covariate to be zero on Day 1 and to grow sub-linearly beyond that (as we have observed in previous studies with this sequence protocol). Note that because the same units were never recorded across days, this **Day** covariate acts like an indicator function for training. If some underlying mode of neural activity was common across both naïve and trained mice, it would be picked up by the condition independent component. Alternatively, since the **Day** covariate is 0 for naïve mice, this covariate captures neural activity modes (bases) that are unique to trained mice. On each Test day, there were 600 trials, so trial number could range from 1 to 600. In addition, stimuli were presented in blocks of 50 trials, so the trial within block number could range from 1 to 50. The negative logarithm captures our intuition that adaptation and/or synaptic depression will generally have the effect of decreasing firing rates, though the unit factors allow for any given unit to be modulated in either direction. Significant interactions were: angle squared with a final rank of 1, E by angle with a rank of 1, E by angle squared with a rank of 1, second element held by day with a rank of 1, second element held by angle with a rank of 1, E by day with a rank of 1, trial number squared with a rank of 2, and trial number by trial within block with a rank of 1. We included all other possible interaction terms in the model fitting procedure, but all had a final rank of 0. We also included one triplet interaction term: E by second element held by day, which had a final rank of 0.

Figure	Time Window	Data Mean	Data Standard Deviation	Approximate 95% Confidence Interval	Test Type	Test Statistic
7b	Early	Day 1 FR· 0.453 2· 0.377 3· 0.384 4· 0.410	1· 0.434 2· 0.346 3· 0.305 4· 0.403	1· [0.31,0.60] 2· [0.25,0.50] 3· [0.28,0.49] 4· [0.27,0.55]	Two-sided permutation test	Difference in mean FR from Day 1 (naïve)
7b	Late	1 FR· 0.149 2· 0.023 3· 0.028 4· -0.027	1· 0.160 2· 0.139 3· 0.202 4· 0.281	1· [0.09,0.20] 2· [-0.03,0.07] 3· [-0.04,0.10] 4· [-0.12,0.07]	Two-sided permutation test	Difference in mean FR from Day 1
8d	Early	1 FR Difference· -0.200 2· -0.040 3· -0.066 4· -0.104	1· 0.392 2· 0.276 3· 0.262 4· 0.198	1· [-0.33,-0.07] 2· [-0.14,0.06] 3· [-0.16,0.02] 4· [-0.17,-0.04]	Two-sided permutation test	Difference in mean FR Difference from Day 1
8d	Late	1 FR Difference· -0.005	1· 0.189 2· 0.201 3· 0.200	1· [-0.07,0.06] 2· [0.08,0.23] 3· [0.04,0.17]	Two-sided permutation test	Difference in mean FR

		2· 0.153 3· 0.106 4· 0.155	4· 0.223	4· [0.08,0.23]		Difference from Day 1
9d	Early	1 FR Difference· -0.169 2· -0.007 3· -0.084 4· -0.009	1· 0.254 2· 0.234 3· 0.346 4· 0.254	1· [-0.25,-0.84] 2· [-0.09,0.08] 3· [-0.20,0.03] 4· [-0.10,0.08]	Two-sided permutation test	Difference in mean FR Difference from Day 1
9d	Late	1 FR Difference· 0.006 2· 0.095 3· 0.057 4· 0.182	1· 0.136 2· 0.195 3· 0.288 4· 0.183	1· [-0.04,0.05] 2· [0.02,0.17] 3· [-0.04,0.15] 4· [0.12,0.24]	Two-sided permutation test	Difference in mean FR Difference from Day 1
10b	Early (A $\tilde{B}$ )	1 FR· 0.252 2· 0.174 3· 0.211 4· 0.267	1· 0.294 2· 0.315 3· 0.249 4· 0.310	1· [0.15,0.35] 2· [0.06,0.29] 3· [0.13,0.30] 4· [0.16,0.37]	Two-sided permutation test	Difference in mean FR from Day 1
10b	Late (A $\tilde{B}$ )	1 FR· 0.200 2· 0.082 3· 0.052 4· -0.002	1· 0.264 2· 0.253 3· 0.245 4· 0.255	1· [0.11,0.29] 2· [-0.01,0.17] 3· [-0.03,0.14] 4· [-0.09,0.8]	Two-sided permutation test	Difference in mean FR from Day 1
10b	Early (E $\tilde{B}$ )	1 FR· 0.418 2· 0.411 3· 0.320 4· 0.289	1· 0.388 2· 0.483 3· 0.375 4· 0.342	1· [0.29,0.55] 2· [0.23,0.59] 3· [0.19,0.45] 4· [0.17,0.41]	Two-sided permutation test	Difference in mean FR from Day 1
10b	Late (E $\tilde{B}$ )	1 FR· 0.144 2· 0.045 3· 0.045 4· 0.039	1· 0.237 2· 0.245 3· 0.269 4· 0.307	1· [0.06,0.22] 2· [-0.04,0.13] 3· [-0.05,0.14] 4· [-0.07,0.14]	Two-sided permutation test	Difference in mean FR from Day 1
11c – Real Data	n/a	1 Soft Accuracy· 0.260 2· 0.245 3· 0.298 4· 0.326	n/a	1· [0.241,0.278] 2· [0.227,0.263] 3· [0.279,0.317] 4· [0.306,0.346]	Two-sided permutation test	Difference in soft accuracy from Day 1
11c – Simulated Data	n/a	1 Soft Accuracy· 0.594 2· 0.591 3· 0.648 4· 0.644	n/a	1· [0.574,0.615] 2· [0.570,0.612] 3· [0.627,0.668] 4· [0.623,0.664]	Two-sided permutation test	Difference in soft accuracy from Day 1

**Supplemental Table 2: Summary of Statistical Tests**

n/a: not applicable

## CONCLUSIONS

The hypothesis proposed at the beginning, that the visual system is an unsupervised learning device, and that time is somehow relevant, has been made somewhat more precise. We might think of the visual system as learning a latent-variable model of its inputs, and in particular one that explicitly accounts for dependences in space and time. The process of learning such a model is an example of information compression, as the latent variables provide a compact representation of the data in precisely Barlow's sense of efficient coding: to discover such a set of latent variables and represent data through them is efficient. A model like that proposed by Bakhtiari et al. provides an interesting starting point, as its internal components closely match data from the mouse visual system (Bakhtiari et al., 2021). In essence, their model is quite like a mixture of PCA and predictive coding. They propose a predictive learning scheme, in which the system uses a low-dimensional set of latent variables to predict the future, and then verifies its predictions, making corrections to improve subsequent predictions. In doing so, the model learns a lot about the structure of the environment.

Though we have focused on unsupervised learning, there is likely an important role in this story for active perception, sensorimotor control, supervised and reinforcement learning, and the general idea of learning about the environment through direct interaction with it. It seems plausible to imagine the brain actively seeking information from the environment, adjusting its sensors intentionally to maximize information (Chen Chen, Murphey, & MacIver, 2020; Klyubin, Polani, & Nehaniv, 2008). However, given the continuous flow

of visual data through the system, whenever we are walking or crawling or driving or eating, it would seem a waste not to take advantage. Indeed, there is ample evidence to support the notion that unsupervised learning is a crucial aspect of neural computation. It is therefore likely some combination of hard-wired circuitry and different learning schemes that define how the entire organism learns and navigates its environment.

The experiments we performed in mouse V1 provide evidence for unsupervised learning, as passive exposure to the sequence stimulus caused dramatic changes in V1 neural activity. Expectation violations were signaled with higher firing rates in the experimental group compared to controls, as expected of an efficient code. We did find that firing rates to ABCD *increased* across days on average, contrary to what would be expected of an efficient code. However, it appears that most, if not all, of the increase occurred only for the early, transient response to A. This is reasonable because A occurs randomly in time, so it is never expected, and also carries all of the information for the entire sequence. We also found an interesting dissociation between early and late windows after the onset of each sequence element. Most of visual neuroscience has focused on neural responses to static images averaged within our early window, so it is somewhat difficult to interpret this result. However, the timing of the suppression approximately matches RGC and dLGN temporal tuning curves; one can imagine straightforward changes in retinal or dLGN temporal tuning leading to changes in V1 like those we observed. Other explanations would have to involve local V1 inhibitory interneurons, or recurrent circuitry, either within V1, or across different brain regions.

It is important to note that our data shows no evidence that V1 learns the entire sequence, but rather at most three elements and more likely only two neighboring ones. When we substituted E for A, differences persisted for about 300-400ms and then vanished. Thus, D in ABCD and D in EBCD seem to have the exact same neural representation, despite being members of different sequences. This is consistent with data in humans showing that it is fairly easy to learn a one-back transition structure (a simple Markov chain), but difficult to learn a two-back structure (Maheu, Dehaene, & Meyniel, 2019; Meyniel, Maheu, & Dehaene, 2016; Modirshanechi, Kiani, & Aghajan, 2019). In general, the visual system deals with spatiotemporal information in an effectively continuous manner, but “sequence learning” in most contexts refers to a discrete process. Reconciling these two perspectives will be difficult, though the involvement of the hippocampus in sequence learning may allow for some cross-pollination with that literature.

The rest of this section will not summarize where we have been but rather consider where we ought to go. In an effort to imagine potential experiments, we’ll start with a series of thought experiments. What experiment would we do if we had infinite time and infinite resources? Even if we could record from every neuron in the brain simultaneously, unravel the intricate connectivity between all of them, reveal every cell type, what would be a good experiment? What about if we could record from a single brain, engaged in everyday activities, for extremely long periods of time? The goal is to let the imagination run wild. Then, hopefully with some notion of what types of experiments might be interesting, we

will consider what is actually possible today. What can we do with current technology that promises to make meaningful progress toward the goal of understanding unsupervised learning in the visual system?

### **Infinite Time, Infinite Resources**

Given infinite time and infinite resources, one almost trivial observation is that receptive fields are not a particularly good way to think about visual neurons. It certainly makes sense to consider how retinal ganglion cells compute functions of their visual inputs, but at all subsequent processing stages visual neurons compute functions of their synaptic inputs not the image impinging on the retina. We might therefore simultaneously record from visual cortical neurons and all of their pre-synaptic inputs. What kinds of dLGN relay neurons synapse onto a given layer 4 V1 neuron? How does that cortical neuron integrate its inputs? How does synapse location on the dendritic shaft influence firing? To what extent does feedback from higher visual areas influence the activity of that neuron? All of these questions could plausibly be answered if one could identify and record from a single neuron and all of its pre-synaptic inputs (knowing where those input neurons reside). This would constitute a major step toward understanding what types of computations V1 neurons perform.

Of particular interest would be to observe a cortical neuron and all of its pre-synaptic connections over time, in order to understand what kinds of plasticity rules operate locally at each synapse, or globally across the cell. Most neuroscientists believe that plasticity rules are functions only of pre- and post-synaptic neural activity. Plasticity acts locally, and is

directly related to action potentials and firing rates. Of course, there are homeostatic plasticity mechanisms and heterosynaptic plasticity, but these too are thought to be regulated by local neuronal activity. Recent experimental and theoretical work, however, suggests that there are other possibilities (Akhlaghpour, 2022; Pastuzyn et al., 2018). Neurons are able to package mRNA into viral capsids and transmit messages across the synapse. In addition, RNA molecules are in theory capable of universal computation (they are Turing complete), so the nature of such messages could be far more sophisticated than we currently imagine. To observe a neuron and all of its pre-synaptic inputs over time would provide convincing evidence either for or against the traditional view of local synaptic plasticity.

With respect to the fundamental hypothesis presented here, a key corollary hypothesis is that information at each stage of the visual hierarchy should be encoded with increasing efficiency. That is, the system ought to compress visual information, not just in the retina but throughout the processing stages. Given that visual information distributes to many areas throughout the brain and circuitry is highly recurrent, this would likely be difficult to test in the most general case. We could simplify, however, and look exclusively at V1. To conclusively measure such an effect, one would need to record neural activity at the optic nerve and the outputs of V1 (in theory recording from every axon in the optic nerve and every output axon from V1, or at least those transmitting to higher visual regions). The hypothesis is that the sum of the entropies of the optic nerve axons is greater than the sum of the entropies of the V1 axons, but that mutual information is approximately equal to the entropy of the joint distribution of optic nerve axons (Appendix on Efficient Coding). Thus,

information is preserved at the V1 output, but encoded more efficiently. Making such a measurement would require extremely long recording times under naturalistic conditions (Paninski, 2003). Alternatives include experiments similar to the one we performed here, training an animal to expect something and then violating that expectation and looking for evidence that unexpected events are encoded with higher firing rates than expected ones. In general, however, we cannot know exactly what the animal has learned nor the ways in which new information is stored in the context of its previous knowledge, so there are definitely confounding factors.

The predictive coding network architecture shown in Figure 2 is recapitulated in the neural circuitry of the retina, with inhibition from amacrine cells onto RGCs likely responsible for passing on the prediction. One staunch prediction of a predictive coding model like this is that an inhibitory cell and its direct excitatory output ought to be decorrelated on very fast timescales. Recall that the excitatory neurons output a prediction error,  $r(t) = x(t) - \alpha x(t - 1)$  in the simplest case, while the concurrent inhibitory output is  $x(t - 1)$ . Under the model,  $r(t)$  is pure noise and so is guaranteed to be perfectly uncorrelated with  $x(t - 1)$ . In practice, decorrelation is likely far from perfect, but one would expect a fair degree of decorrelation. In V1, the circuit architecture is very similar to the predictive coding network: thalamocortical inputs arrive simultaneously to both PV and excitatory layer 4 neurons. The PV cells then inhibit the excitatory cells. This raises the possibility that the first step of cortical processing is another predictive coding step. Most *in vivo* studies have shown that PV and excitatory cells in layer 4 are actually highly correlated (Dipoppa et al., 2018; Hu et al., 2014; Karnani, Jackson, Ayzenshtat, Tucciarone, et al., 2016), seemingly

in contradiction to this hypothesis. However, most of these studies have either used two-photon calcium imaging, binned their data to at least 100ms, or measured correlations in response to non-natural visual environments. It is possible that decorrelation between PV and excitatory occurs only on faster timescales in natural environments, and that at longer timescales or in non-natural environments spurious correlations emerge. Given infinite time and resources, it would be very easy to look for local PV/excitatory decorrelation.

With respect to sequence learning, there are a number of interesting potential experiments. The most obvious, hinted at before, would be to look for evidence of heterosynaptic plasticity at GABAergic synapses in V1. If we could record from and manipulate every cell, then it would be straightforward to measure the strength of inhibition onto different cells both pre- and post-synaptically. We could measure how the number and size and release rate of pre-synaptic vesicles change with learning, or do the same for the density of post-synaptic GABA receptors (and then compare to controls). Inhibitory plasticity, induced through heterosynaptic plasticity at neighboring synapses, has been shown to occur even in the absence of inhibition (Gandolfi et al., 2020). Therefore, traditional techniques, like introducing NMDAR or mAChR antagonists, may not block this form of plasticity. However, GABA-B receptor antagonism, through application of saclofen, has been shown to inhibit potentiation of inhibitory synapses in rat V1 (Komatsu, 1996). GABA-B receptors are also implicated in BDNF signaling, so antagonism of TrkB kinase receptors may also block sequence learning, though this would not provide a conclusive result as BDNF signaling occurs in other contexts as well.

Another option would be to make a few adjustments to the sequence learning stimulus, but still attempt to test the notion of efficient coding. As mentioned above, passive exposure to a behaviorally irrelevant stimulus ought to induce suppression of firing rates with learning. In the case of a stimulus that has a fixed temporal frequency, the prediction of efficient coding is that the system ought to learn a notch filter for that frequency, in order to suppress it. So, a plausible experiment would test for a shift in the temporal frequency tuning of a large population of V1 neurons. Tuning would be measured in naïve mice and then again in mice after passive exposure to a variety of visual stimuli all with the same temporal frequency. Tuning ought to shift away from the trained frequency, quite the opposite effect expected from traditional Hebbian learning.

### **Finite Time, Finite Resources**

With modern technology and realistic time constraints, most of the experiments described in the previous section are impossible. The experiment that describes measuring the entropy of the total output of V1 might even be impossible in theory, due to the finite time we would have to record from any given animal and the difficulty in even defining what we mean by the outputs of V1. In this section, we will therefore focus on two potential experiments that are plausible.

Perhaps the most interesting experiment relevant to a more general neuroscience audience would be to test the hypothesis that feedforward inhibition at the first thalamocortical synapse in V1 serves as a predictive coder. There are effectively two testable predictions: 1) thalamorecipient PV and pyramidal neuron activity in layer 4 of V1 ought to be

decorrelated on fast timescales during natural viewing conditions. The ideal conditions would be as a mouse roams a natural environment in relative darkness. The precise timescale of decorrelation ought to be the same as the timescale of stimulus variation caused by the mouse's movement and movement within the environment. 2) The PV-to-pyramidal synapse ought to follow an Anti-Hebbian plasticity rule, such that correlated firing between the two neurons strengthens the inhibitory synapse. The first prediction is simpler to test. It would require the use of electrophysiology to capture the fast timescale of neural activity, and an opto-tagging setup in a PV-cre mouse model in order to accurately identify layer 4 PV neurons. Presumptive excitatory neurons would be those not tagged by optogenetic stimulation and, maybe, those with a wide spike waveform. Due to the relative sparsity of neurons in cortex, multiple implant surgeries would likely be necessary. A secondary experiment to complement this one would analyze the PV-pyramidal correlations at the same timescale under "non-natural" viewing conditions, such as a head-fixed mouse viewing drifting gratings. The mouse's internal predictive model would be optimized to its natural viewing conditions, and would almost certainly be sub-optimal for different environmental statistics.

The second potential experiment relates to the notion of heterosynaptic plasticity of inhibition, mediated by acetylcholine. An experiment to test this would introduce GABA antagonists during the training phase of sequence learning, as was done previously with scopolamine. GABA-A antagonists are convulsants, so this would not be advised, but GABA-B antagonists have been shown to have anti-convulsant effects. If local GABA-B antagonism blocked sequence learning, this would provide a starting point for continued

research. Of course, heterosynaptic plasticity at GABA-B receptors is not the only possibility; a thorough literature review of inhibitory plasticity mechanisms and interactions with acetylcholine may reveal alternatives.

### **Fin**

This brings the dissertation to a close. My sincere hope is that somebody, maybe me, will grow in wisdom from having read it.

## APPENDIX

### Probability and Information Theory

Probability is the quantification of uncertainty (Wasserman, 2004). Probability theory concerns how data-generating processes produce unique sets of outcomes, and the properties of those outcomes. A data-generating process, or source, might be a coin flip, the motions of the planets, or the sequence of action potentials generated by a neuron during an experiment. Probability theory provides a mathematical language to understand random processes, specifically what types of outcomes to expect from those processes and how often to expect them.

Information theory is also a quantification of uncertainty, though explicitly tied to the problem of communication in its original form (Cover & Thomas, 2005; Shannon, 1948). Information is transmitted on telephone wires or through the air for radio. Information resolves uncertainty. During a game of 20 questions, the questioner and the questioned transfer information; each question and answer resolve uncertainty as the questioner hones in on the solution. Information theory provides a formalism to quantify information, and it is built in the language of probability theory.

#### *Laws of Probability*

For a given data-generating process, such as an experiment, the sample space is the set of all possible outcomes or realizations. For a single roll of the die, there are six possible outcomes corresponding to the six sides of the die. The sample space is  $\Omega = \{1,2,3,4,5,6\}$ .

An event is any subset of the sample space. An event allows us to assign importance to some portion of the sample space. We might have, for example, the event that the die lands on a number less than 4,  $A = \{1,2,3\}$ , or the event that the die lands on an even number,  $B = \{2,4,6\}$ . We can then discuss the intersection or union of events:

$$\begin{aligned} \text{Union (A or B): } A \cup B &= \{1,2,3,4,6\} \\ \text{Intersection (A and B): } A \cap B &= \{2\} \\ \text{Disjoint if: } A \cap B &= \emptyset \end{aligned}$$

Based on these ideas and the fundamental axioms of probability, one can define a formal mathematical system to quantify uncertainty (Wasserman, 2004). Probability theory, through the axioms and laws of probability, allows us to make predictions regarding the outcomes of experiments simply by manipulating symbols.

Here are the three Kolmogorov axioms of probability (Wasserman, 2004):

$$\begin{aligned} P(A) &\geq 0 \\ P(\Omega) &= 1 \\ \text{If } A, B, C \text{ are disjoint events, } P(A \cup B \cup C) &= P(A) + P(B) + P(C) \end{aligned}$$

And some basic definitions:

$$\begin{aligned} P(AB) &\triangleq P(A \cap B) \\ P(A|B) &\triangleq \frac{P(AB)}{P(B)} \end{aligned}$$

Everything else is a direct consequence of set theory, the three axioms, and a few more basic definitions.

$$\begin{aligned} P(A) + P(A^c) &= 1 \text{ where } A^c \text{ is effectively "not A"} \\ P(A \cup B) &= P(A) + P(B) - P(AB) \end{aligned}$$

Two events are *independent* if:

$$P(AB) = P(A)P(B)$$

Chain Rule of Probability:

$$P(A_n \cap \dots \cap A_1) = \prod_{k=1}^n P(A_k | \bigcap_{j=1}^{k-1} A_j)$$

Law of Total Probability:

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

where the summation must be over a sequence of disjoint events  $(A_1, A_2, \dots)$  whose union spans the entire sample space.

Bayes' Theorem:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

In this formulation,  $P(A_i)$  is known as the *prior*, representing our knowledge or belief about  $A_i$  before seeing  $B$ .  $P(A_i|B)$  is the *posterior*, which represents our updated knowledge about  $A_i$  given that we have observed  $B$ .

Using the concepts of a sample space and the laws of probability, we can answer a wide variety of questions about random events that are difficult to answer intuitively (Wasserman, 2004). For example, suppose you have 3 playing cards. The first card is green on both sides ( $GG$ ), the second is red on both sides ( $RR$ ), and the third is red on one side and green on the other ( $RG$ ). One card of the three is chosen at random and one side of that

card is displayed. If the displayed side is green, what is the probability that the other side is also green?

To answer this question, we must define the sample space. There are actually two aspects to the space of all possible outcomes, the first pertaining to seeing the card initially, and the second pertaining to seeing the back side of the chosen card. If  $RG$  denotes the outcome of seeing a red card initially and then seeing that its back is green, the sample space is:

$$\Omega = \{GG, GG, GR, RG, RR, RR\}$$

If we are first shown green, then there are three possibilities: 1) we are looking at side 1 of the green-green card; 2) we are looking at side 2 of the green-green card; 3) we are looking at the green side of the green-red card. Thus, the probability of first seeing green and then flipping it over and seeing green again is  $2/3$ .

We can also answer the same question with the laws of probability. Define:

$$\begin{aligned} G_1 &\triangleq \text{event that displayed side is green} \\ G_2 &\triangleq \text{event that other side is green} \\ R_2 &\triangleq \text{event that other side is red} \end{aligned}$$

We wish to know:

$$P(G_2|G_1)$$

Considering the description in the problem and the definition of conditional probability, we recognize that:

$$\begin{aligned} P(G_2|G_1) &= P(G_1G_2)/P(G_1) \\ P(G_1G_2) &= 1/3 \\ P(G_1) &= 1/2 \end{aligned}$$

And therefore:

$$P(G_2|G_1) = 2/3$$

In this situation, it was simple enough to write down the entire sample space, but in more complicated scenarios, using the laws of probability becomes necessary.

### *Random Variables*

A random variable is a mapping from each outcome in the sample space to a real number (Wasserman, 2004). In a series of ten coin flips, the set of all possible numbers of heads can be mapped to a random variable, i.e.,  $\{0, \dots, 10\}$ . Depending on the probability of landing on heads, we can assign a unique probability to each possible outcome. For example, if  $P(\text{heads}) = 0.5$ , then  $P(0 \text{ heads in } 10 \text{ flips}) = 0.5^{10} = 9.77e - 4$ . For a discrete random variable such as this, these unique probabilities define a *probability mass function* over every possible realization,

$$f(x) \triangleq P(X = x)$$

For a continuous random variable, we can define a *probability density function*,  $f(x)$ , that satisfies:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$p(a < X < b) = \int_a^b f(x) dx$$

The probability mass and density functions allow us to answer questions about the space of all possible outcomes.

Some common probability mass functions are:

$$Z \sim \text{Binomial}(Z | n, p)$$

$$f(z; n, p) = \begin{cases} \binom{n}{z} p^z (1-p)^{n-z} & \text{for } z = 0, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

$$R \sim \text{Poisson}(R | \lambda)$$

$$f(r; \lambda) = e^{-\lambda} \frac{\lambda^r}{r!}$$

Some common probability density functions are:

$$X \sim \text{Uniform}(X | a, b)$$

$$f(x; a, b) = \frac{1}{b-a} \text{ for } x \in [a, b]$$

$$Y \sim \text{Normal}(Y | \mu, \sigma^2)$$

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Note that these are known as parametric distributions, because each distribution has a set of parameters, such as  $\lambda$  and  $\mu$ , that fully specify everything that could possibly be known about the distribution and the random variable.

### *Expectation*

For parametric distributions, the parameters provide a full characterization of the distribution. In general, though, we require a way to summarize distributions. The expectation, or expected value, is the most common summary of a random variable and its distribution. It is the average value of the random variable. For a discrete random variable, the expectation is:

$$\mathbb{E}[X] \triangleq \sum_x xP(X = x)$$

For a continuous random variable:

$$\mathbb{E}[X] \triangleq \int xp(x)dx$$

Because different outcomes occur with different frequencies, the expectation is a probability-weighted average. It is the center of mass of the distribution.

For functions of a random variable, the expectation is simply:

$$\mathbb{E}[r(X)] = \int r(x)p(x)dx$$

The expectation is a linear operator, so it has the following property:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

We can also use the expectation to define the variance as another distribution summary.

The variance summarizes the spread of a distribution around its center of mass:

$$\text{Var}(X) \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2]$$

The variance has the following properties:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ \text{Var}(aX + b) &= a^2\text{Var}(X)\end{aligned}$$

For a normal distribution, the parameters uniquely specify the expectation and the variance.

### *Statistics*

In general, statistics is the inverse of probability (Wasserman, 2004). While probability seeks to determine the properties of the outcomes of experiments, statistics attempts to use the outcomes, i.e., the data, to make inferences about the underlying data-generating

process. Note that *statistic*, singular, is any function of data. The average of a sample of data is a statistic.

### *Statistical Inference*

The most common problem in statistical inference is to use a dataset,  $(X_1, \dots, X_N)$ , to infer the underlying distribution of the data-generating process. For example, we might assume the data is generated from a normal distribution, though we do not know the values of the specific parameters,  $\mu$  and  $\sigma^2$ . We wish to discover a process by which we can reliably infer the parameters from the data, thereby inferring the underlying distribution.

For the normal distribution, the most common method to infer the parameters is to use the sample mean and the sample variance:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i$$
$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \hat{\mu})^2$$

where the hat symbol signifies that the parameters have been estimated from data.

Note that in practice, it is common to assume the data comes from some named parametric distribution, learn those parameters from the data, and then verify that assumption *post hoc* using a variety of goodness-of-fit techniques. There are, however, many techniques to learn the full probability distribution making minimal assumptions about its form.

*Maximum Likelihood Estimation*

Maximum likelihood estimation is an efficient and very general way to use data to estimate the parameters of a model, and therefore infer the underlying distribution of the data. The process begins by specifying a model, or joint probability distribution. This will depend on the specific features of the dataset. Once a model has been specified, you write down the likelihood function. The log likelihood function is the log of the joint probability of the data, assuming these are independent and identically distributed given the model:

$$\mathcal{L}(\theta; X) = \log\{p(X|\theta)\} = \log\left\{\prod_{i=1}^n f(x_i; \theta)\right\}$$

The likelihood function is a function of the parameters, with the data presumed fixed. In essence, the likelihood function summarizes the fit between different parameter values and the observed data. The set of parameter values that maximize the likelihood are those that are most likely to have generated the data. The maximum likelihood estimates for the parameters,  $\theta$ , is given by:

$$\hat{\theta}_{MLE} = \arg \max_{\theta} \mathcal{L}(\theta; X)$$

The maximum likelihood estimator has a number of important properties that make it a good estimator, which will not be explored here. However, we note these properties contribute to its widespread use in Statistics and Machine Learning. Almost all techniques in Bayesian inference, for example, use a likelihood function as part of the inference procedure and the Bayesian estimator is the maximum likelihood estimator for certain choices of prior distribution. In addition, many common models in Machine Learning and a variety of loss functions are actually likelihood functions in disguise.

### *Information Entropy*

In Physics, entropy is a measure of disorder or randomness. It is related to the number of unique ways a system can be configured and the likelihood of finding the system in different configurations. Claude Shannon introduced the related concept of information entropy to measure the uncertainty of a stochastic process, such as the sequence of letters in a handwritten note (Shannon, 1948). Information entropy characterizes the amount of uncertainty regarding which symbol will be chosen, or which message sent, among all possible alternatives. While the information generated by a single event is measured by  $I(x) = -\log_2(P(X = x))$ , the information or uncertainty associated with the entire sample space of possible outcomes is the expectation of the information, or information entropy,  $H$  (Cover & Thomas, 2005):

$$H(X) = \mathbb{E}[I(x)] = - \sum_{x \in \mathcal{X}} P(X = x) \log_2(P(X = x))$$

For a discrete random variable with a sample space of two outcomes, one with probability  $p$  and the other with probability  $1-p$ , the entropy is:

$$H(X) = -p \log_2(p) - (1 - p) \log_2(1 - p)$$

This is the average entropy for a single coin flip. Maximum entropy comes from a fair coin:  $p = 1 - p = 0.5$  and  $H(X|fair) = 1$ . For a biased coin, perhaps  $p = 0.05$ , then  $H(X|biased) = 0.286$ . The fair coin is the condition of maximum uncertainty among all possible values of  $p$ . Thus, in playing a game of 20 questions, the best strategy to maximize information is to ask a series of questions whose answers are expected to be split evenly

between “Yes” and “No”. Each answer will yield 1 bit for a total of 20 bits of information produced by the game on average:

$$I(No) = I(Yes) = -\log_2(0.5) = 1 \text{ bit}$$

$$I(20 \text{ Questions}) = 20 * H(X|fair) = 20 * [P(No)I(No) + P(Yes)I(Yes)] = 20 \text{ bits}$$

Alternatively, one could ask a series of questions that are expected to result in “No” 95% of the time. In this latter scenario, each “No” will provide

$$I(No) = -\log_2(0.95) = 0.07 \text{ bits}$$

and each “Yes”

$$I(Yes) = -\log_2(0.05) = 4.32 \text{ bits}$$

Every “Yes” provides a lot of information, but because “Yes” is an unlikely answer, over the course of the 20 questions there will only be about one “Yes”. Thus, the total information transmitted during the game will be approximately:

$$I(20 \text{ Questions}) = 20 * H(X|biased) = 19 * I(No) + 1 * I(Yes) = 5.73 \text{ bits}$$

significantly less than the game with equally likely outcomes. Thus, the statistics of the source, or data-generating process, govern the entropy, which is the amount of uncertainty resolved on average by each realization of the process or each message received.

We end this section by noting that there is a continuous analog to the entropy, known as the limiting density of discrete points. In Shannon’s original paper, however, he proposed the differential entropy:

$$h(X) = \mathbb{E}[-\log_2(p(x))] = - \int_{-\infty}^{\infty} p(x) \log_2(p(x)) dx$$

The differential entropy, unlike the entropy of a discrete distribution, is only meaningful in relative terms because it is not invariant to transformations of variables. Thus, it must be used to discuss a difference in entropy between distributions rather than as a raw measure of information. In practice, data is stored in digital form: it must first be “quantized”, or converted from a continuous space to a discrete one, and then its entropy can be computed using the discrete form.

A special case occurs for the normal distribution,  $\mathbf{x} \in \mathbb{R}^k$ :

$$\mathbf{x} \sim \text{Normal}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

such that the differential entropy is:

$$h(\mathbf{x}) = \frac{k}{2} + \frac{k}{2} \log_2(2\pi) + \frac{1}{2} \log_2(|\boldsymbol{\Sigma}|)$$

The entropy of a normal distribution is a function only of its variance (assuming constant dimensionality  $k$ ), given in this multivariate case by the determinant of the variance-covariance matrix. This provides additional insight into the nature of information and entropy. As variance increases, so does entropy and therefore the average amount of information conveyed by each realization of, or draw from, the distribution.

### *Mutual Information*

Mutual information is a measure of the average amount of information about one variable that is gained by observing another variable. It is a symmetric quantity. With the entropy defined as above, the mutual information is:

$$\begin{aligned} I(X; Y) &\triangleq H(X) + H(Y) - H(X, Y) \\ &= H(X) - H(X|Y) \end{aligned}$$

$$\begin{aligned}
 &= H(Y) - H(Y|X) \\
 &= H(X, Y) - H(X|Y) - H(Y|X)
 \end{aligned}$$

The joint entropy,  $H(X, Y)$ , measures the entropy of the joint distribution of  $X$  and  $Y$ , while the conditional entropy,  $H(X|Y)$ , is

$$H(X|Y) \triangleq - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(X = x, Y = y) \log_2(P(X = x|Y = y))$$

### *Efficient Coding*

According to Claude Shannon, who envisioned the key elements of information theory while working for Bell Telephone Labs in the 1940s:

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. (Shannon, 1948)

In transmitting messages, the sender makes certain choices about what to send and how. The message must be encoded in some form, whether that be through written text, Morse code, binary, or some other means. As a concrete example, consider sending a message from a four-letter alphabet,  $\theta = \{A, B, C, D\}$ . A typical message might look like BACAAABDAA. Assume that each letter appears independently of all other letters according to the probabilities:

$$P(A) = \frac{1}{2} \quad P(B) = \frac{1}{4} \quad P(C) = \frac{1}{6} \quad P(D) = \frac{1}{12}$$

The source entropy is,

$$H(\theta) = - \sum_{x \in \{A, B, C, D\}} P(X = x) \log_2(P(X = x)) = 1.73 \text{ bits}$$

such that the total information in a message of  $N$  symbols would be  $NH(\theta)$  bits on average. Encoding the message requires an encoding scheme, for example binary, and the source entropy of 1.73 bits provides a lower bound on the minimum amount of information that can be transmitted before information is lost (Shannon, 1948). This *source coding theorem* thereby provides a definition of an efficient code (Atick, 1992; H. B. Barlow, 1961; Shannon, 1948; Sterling & Laughlin, 2015): a code is efficient if messages are on average transmitted with the same number of bits as the entropy of the source, under the constraint that minimal information be lost. For a perfectly efficient code, no information is lost and the transmitted messages are fully compressed: all predictable information has been removed, leaving only unpredictable noise. This is precisely the definition of entropy, the irreducible uncertainty.

Imagine we were to encode our alphabet in the following way:

A	00
B	01
C	10
D	11

Every time a letter of the alphabet is transmitted, this scheme uses 2 bits. However, we know from Shannon's source coding theorem that we can do better, approaching the theoretical minimum of 1.73 bits. This particular code is inefficient due to redundancy: predictable information is still present in the transmitted messages. For example,  $P(A) = \frac{1}{2}$

and  $P(B) = \frac{1}{4}$ , so receiving a 0 for the first bit implies a 2/3 chance that the next bit will also be a 0.

Formally, the Shannon redundancy is given by (Atick, 1992; Cover & Thomas, 2005):

$$R = 1 - H(\theta)/C$$

where  $C = 2 \text{ bits}$  is the average length of a message in our encoding scheme. The redundancy is always between 0 and 1, with a perfectly efficient code having  $H(\theta) = C$  and  $R = 0$ . In this case,  $C = 2$ , so  $R = 0.135$ .

Consider, alternatively, the following 3-bit encoding scheme:

A	0
B	10
C	110
D	111

At first, it seems adding an extra bit would constitute a superfluous increase in our bit allotment. However, we must consider the statistics of the underlying message-generating process. In this case, the symbol *A* occurs 50% of the time, so using only 1 bit for the most prevalent symbol may be wise. Indeed, for the 3-bit scheme, the average information per message decreases:

$$C = P(A) * (1 \text{ bit}) + P(B) * (2 \text{ bits}) + (P(C) + P(D)) * (3 \text{ bits}) = 1.75 \text{ bits}$$

The redundancy is now  $R = 0.011$ . This is a very efficient code. In general, efficient codes use relatively fewer symbols to encode prevalent messages and relatively more symbols to encode rare messages. Efficient codes thereby signal surprise relative to expectation. This creates a code that is irreducibly random, with absolutely no predictability.

As explored in the section on Vision, efficient coding seems to be prominent in the nervous system. There, sensory information is encoded in spikes. For the spikes of many retinal ganglion cells to efficiently encode incoming light, they must make smart allotments of their total spike budget, avoiding significant redundancy. Unlike the alphabet, where each message was independent of all others (e.g.,  $P(ABC) = P(A)P(B)P(C)$ ), natural signals have statistical dependences across both space and time. These add complexity to the analysis, which we briefly explore here under certain simplifying assumptions (Atick, 1992; Atick & Redlich, 1990; Ocko et al., 2018).

Assume that a simple one-dimensional visual input,  $I \in \mathbb{R}^d$ , is passed through an array of neurons that output linearly-transformed versions of the input,  $O = KI$  with  $O \in \mathbb{R}^n$  and  $K \in \mathbb{R}^{n \times d}$ .  $K$  is often referred to as a neuron's receptive field. The problem of efficient coding asks, what is the optimal  $K$  such that the outputs encode all of the input information with minimal redundancy? Formally,

$$\hat{K} = \arg \min_K \sum_n H(O_n) \\ s. t. H(O) = H(I)$$

We seek the internal representation,  $O$ , that minimizes the sum of the entropies of the output neurons (i.e., we wish to minimize the entropy of the message,  $C$ ), subject to the constraint that the output information equal the input information. This is equivalent to minimizing redundancy. The transformation preserves mutual information between input and output. This constrained optimization problem can be re-written using the technique of Lagrange multipliers (Atick & Redlich, 1992; C. M. Bishop, 2006),

$$\hat{K} = \arg \min_K \sum_n H(O_n) - \lambda(H(O) - H(I))$$

To solve the problem, we must specify a probability distribution over inputs, which for natural scenes has been shown to be well-approximated by a Gaussian (D. W. Dong & Atick, 1995):

$$p(I) = \text{Normal}(I | 0, R)$$

where the covariance matrix  $R \in \mathbb{R}^{d \times d}$  has a special circulant structure due to translation invariance (the statistics of natural scenes are the same in every spatial location). Under this model, the outputs are distributed as

$$p(O) = \text{Normal}(O | 0, \tilde{R})$$

$$\tilde{R} \triangleq KRK^T$$

If we assume  $d = n$  (input dimensionality equal to the number of neurons), then the following equality holds:

$$H(O) = H(I) + \ln|K|$$

The marginal entropies of the individual neurons are directly related to the diagonal elements of the covariance of  $O$ , such that

$$\sum_n H(O_n) = \sum_n \frac{1}{2} \ln \tilde{R}_{nn} + \text{constant}$$

Therefore,

$$\hat{K} = \arg \min_K \sum_n \ln K_n R K_n^T - \lambda \ln |K K^T|$$

where  $K_n$  is the  $n$ -th row of  $K$ . Because the natural log is a convex function and all  $K_n R K_n^T$  are non-negative, minimizing  $\sum_n \ln K_n R K_n^T$  is equivalent to minimizing  $\sum_n K_n R K_n^T = \text{Trace}(K R K^T)$ , and so,

$$\hat{K} = \arg \min_K \text{Trace}(KKK^T) - \lambda \ln|KK^T|$$

Taking the appropriate derivatives and setting equal to zero,

$$\frac{\partial \mathcal{L}}{\partial K} = 0 = 2RK^T - 2\lambda K^T (KK^T)^{-1}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 = -\ln|KK^T|$$

From the second equation, we require the determinant of  $KK^T$  to equal 1, such that its natural log will be zero. This is achieved when  $KK^T = I_n$  (the identity matrix). Then, the first equation simplifies to:

$$RK^T = \lambda K^T$$

This is an eigenvalue problem and the optimal solution for the columns of  $K^T$  are the eigenvectors of the covariance matrix  $R$ . The optimal information-theoretic solution is therefore PCA, up to an arbitrary orthogonal rotation of  $K$ . This implies that the information-theoretic goal of establishing an efficient code is equivalent to creating an autoencoder capable of accurately reconstructing its inputs (Atick & Redlich, 1990, 1992; Ocko et al., 2018; Olshausen & Field, 1996).

The neurons, computing  $O = KI$ , project the input data into a transformed space with statistically independent and decorrelated outputs ( $\tilde{R} = KKK^T$ , the covariance of the outputs, is a diagonal matrix containing the eigenvalues of  $R$ ). This is a whitening operation on the inputs. Furthermore, because  $R$  is circulant, its eigenvectors are Fourier modes (sinusoids of increasing frequency). Thus, if the number of neurons is less than the dimensionality of the input,  $n < d$ , the optimal solution will include only the eigenvectors

with the largest eigenvalues, thereby excluding high-frequency signals and low-pass filtering the inputs. In the presence of internal noise, the optimal solution is instead a band-pass filter, establishing a trade-off between encoding fidelity and noise suppression (Atick & Redlich, 1992; Doi et al., 2012; Ocko et al., 2018).

Note that this solution does not strictly specify the spatial structure of the receptive field, as it is invariant to orthogonal rotations of  $K$  (in the Fourier domain, the equivalent statement is that we know the power spectrum of the optimal filter but not its phase) (Atick & Redlich, 1992; Doi et al., 2012). However, when realistic constraints are placed on the problem, such as physical spacing between photoreceptors, a firing-rate budget, and rectification, neurons do not represent specific Fourier modes. They instead show center-surround receptive fields, localized in space and time, qualitatively and quantitatively similar to what is actually observed in the retina (D. Dong & Atick, 1995; Ocko et al., 2018).

## BIBLIOGRAPHY

- Abbott, L. F., & Nelson, S. B. (2000). Synaptic Plasticity: taming the beast. *Nature Neuroscience*, 3. Retrieved from [https://www.nature.com/articles/nn1100\\_1178.pdf](https://www.nature.com/articles/nn1100_1178.pdf)
- Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2(2), 284–299. Retrieved from [http://persci.mit.edu/pub\\_pdfs/spatio85.pdf](http://persci.mit.edu/pub_pdfs/spatio85.pdf)
- Akhlaghpour, H. (2022). An RNA-based theory of natural universal computation. *Journal of Theoretical Biology*, 537, 110984. <https://doi.org/10.1016/J.JTBI.2021.110984>
- Ángel García-Cabezas, M., Zikopoulos, B., & Barbas, H. (2019). The Structural Model: a theory linking connections, plasticity, pathology, development and evolution of the cerebral cortex. *Brain Structure and Function*, 224, 985–1008. <https://doi.org/10.1007/s00429-019-01841-9>
- Aoi, M. C., Mante, V., & Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience*, 23, 1410–1420. <https://doi.org/10.1038/s41593-020-0696-5>
- Aoi, M. C., & Pillow, J. W. (2018). Model-based targeted dimensionality reduction for neuronal population data. *Neural Information Processing Systems*. Retrieved from <http://pillowlab.princeton.edu/jpillow/>
- Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 22(1–4), 213–251. <https://doi.org/10.3109/0954898X.2011.638888>
- Atick, J. J., Li, Z., & Redlich, A. N. (1992). Understanding Retinal Color Coding from First Principles. *Neural Computation*, 4, 559–572. Retrieved from <http://direct.mit.edu/neco/article-pdf/4/4/559/812346/neco.1992.4.4.559.pdf>
- Atick, J. J., & Redlich, A. N. (1990). Towards a Theory of Early Visual Processing. *Neural Computation*, 2(3), 308–320. <https://doi.org/10.1162/neco.1990.2.3.308>
- Atick, J. J., & Redlich, A. N. (1992). What Does the Retina Know about Natural Scenes? *Neural Computation*, 4, 196–210. Retrieved from <https://www.mitpressjournals.org/doi/pdf/10.1162/neco.1992.4.2.196>
- Atick, J. J., & Redlich, A. N. (1993). Convergent Algorithm for Sensory Receptive Field Development. *Neural Computation*, 5(1), 45–60. <https://doi.org/10.1162/neco.1993.5.1.45>

- Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*. US: American Psychological Association. <https://doi.org/10.1037/h0054663>
- Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C. C., & Richards, B. A. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Neural Information Processing Systems*, 35. Retrieved from <https://doi.org/10.1101/2021.06.18.448989>
- Balasubramanian, V., & Sterling, P. (2009). Receptive fields and functional architecture in the retina. *Journal of Physiology*. <https://doi.org/10.1113/jphysiol.2009.170704>
- Barbas, H. (2015). General Cortical and Special Prefrontal Connections: Principles from Structure to Function. *Annual Reviews Neuroscience*, 38, 269–289. <https://doi.org/10.1146/annurev-neuro-071714-033936>
- Barlow, H. (2001a). Redundancy reduction revisited. *Network: Computation in Neural Systems*. <https://doi.org/10.1088/0954-898X/12/3/301>
- Barlow, H. (2001b). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*. <https://doi.org/10.1017/S0140525X01000024>
- Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In *Sensory Communication*. <https://doi.org/10.7551/mitpress/9780262518420.003.0013>
- Barlow, H. B. (1989). Unsupervised Learning. *Neural Computation*, 1(3), 295–311. <https://doi.org/https://doi.org/10.1162/neco.1989.1.3.295>
- Barlow, H. B., & Földiák, P. (1989). Adaptation and decorrelation in the cortex. In *The Computing Neuron* (pp. 54–72).
- Bear, M. F. (2003). Bidirectional synaptic plasticity: from theory to reality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1432), 649–655. <https://doi.org/10.1098/rstb.2002.1255>
- Bear, M. F., Kleinschmidt, A., Gu, Q., & Singer, W. (1990). Disruption of experience-dependent synaptic modifications in striate cortex by infusion of an NMDA receptor antagonist. *Journal of Neuroscience*, 10(3), 909–925. <https://doi.org/10.1523/jneurosci.10-03-00909.1990>
- Bell, A. J., & Sejnowski, T. J. (1997). Edges are the “Independent Components” of natural scenes. *Advances in Neural Information Processing Systems*.
- Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., & Wu, Y. (2015). STDP as presynaptic activity times rate of change of postsynaptic activity. *ArXiv*, 1–10.

[https://doi.org/10.1162/NECO\\_a\\_00934](https://doi.org/10.1162/NECO_a_00934)

- Berry, M. J., Warland, D. K., & Meister, M. (1997). The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, *94*, 5411–5416. Retrieved from [www.pnas.org](http://www.pnas.org).
- Bhatt, D. H., Zhang, S., & Gan, W. B. (2009). Dendritic spine dynamics. *Annual Review of Physiology*, *71*, 261–282. <https://doi.org/10.1146/annurev.physiol.010908.163140>
- Bi, G. Q., & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, *18*(24), 10464–10472. <https://doi.org/10.1523/jneurosci.18-24-10464.1998>
- Bickford, M. E., Zhou, N., Krahe, T. E., Govindaiah, G., & Guido, W. (2015). Retinal and Tectal “Driver-Like” Inputs Converge in the Shell of the Mouse Dorsal Lateral Geniculate Nucleus. *Journal of Neuroscience*, *35*(29), 10523–10534. <https://doi.org/10.1523/JNEUROSCI.3375-14.2015>
- Bicknell, B. A., & Häusser, M. (2021). A synaptic learning rule for exploiting nonlinear dendritic computation. *Neuron*, *109*, 1–17. <https://doi.org/10.1016/j.neuron.2021.09.044>
- Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, *2*(1), 32–48. <https://doi.org/10.1371/journal.ppat.0020109>
- Bishop, C. M. (1999). Latent Variable Models. *Learning in Graphical Models*, 371–403. Retrieved from <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/bishop-latent-erice-99.pdf>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer.
- Bishop, C., & Tipping, M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, *61*(3), 611–622.
- Blais, B. S., Frenkel, M. Y., Kuindersma, S. R., Muhammad, R., Shouval, H. Z., Cooper, L. N., & Bear, M. F. (2008). Recovery From Monocular Deprivation Using Binocular Deprivation. *Journal of Neurophysiology*, *100*, 2217–2224. <https://doi.org/10.1152/jn.90411.2008>
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. <https://doi.org/10.1080/01621459.2017.1285773>

- Bliss, T. V. P., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology*, *232*(2), 331–356.  
<https://doi.org/10.1113/jphysiol.1973.sp010273>
- Briggs, F. (2010). Organizing principles of cortical layer 6. *Frontiers in Neural Circuits*, *4*(3). <https://doi.org/10.3389/neuro.04.003.2010>
- Bruno, J. P., Gash, C., Martin, B., Zmarowski, A., Pomerleau, F., Burmeister, J., ... Gerhardt, G. A. (2006). Second-by-second measurement of acetylcholine release in prefrontal cortex. *European Journal of Neuroscience*, *24*(10), 2749–2757.  
<https://doi.org/10.1111/j.1460-9568.2006.05176.x>
- Bucher, S., & Brandenburger, A. (2020). In what environments is divisive normalization an efficient computation? In *18th Annual Meeting for the Society of Neuroeconomics* (p. 18).
- Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London B*, *220*, 89–113. Retrieved from  
<https://royalsocietypublishing.org/>
- Burgid, M. F., Cadenaid, S. A., Denfieldid, G. H., Walkerid, E. Y., Toliasid, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLoS Computational Biology*, *17*(6), e1009028.  
<https://doi.org/10.1371/journal.pcbi.1009028>
- Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, *15*(4), e1006897.  
<https://doi.org/10.1371/journal.pcbi.1006897>
- Caporale, N., & Dan, Y. (2008). Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. *Annual Review of Neuroscience*, *31*, 25–46.  
<https://doi.org/10.1146/annurev.neuro.31.060407.125639>
- Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... Rust, N. C. (2005). Do We Know What the Early Visual System Does? *Journal of Neuroscience*, *25*(46). <https://doi.org/10.1523/JNEUROSCI.3726-05.2005>
- Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, *264*(5163), 1333–1336.  
<https://doi.org/10.1126/science.8191289>
- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation.

*Nature Reviews Neuroscience*, 13(1), 51–62. <https://doi.org/10.1038/nrn3136>

- Chalk, M., Marre, O., & Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences of the United States of America*, 115(1), 186–191. <https://doi.org/10.1073/pnas.1711114115>
- Chechik, G., Meilijson, I., & Ruppin, E. (1998). Synaptic Pruning in Development: A Computational Account. *Neural Computation*, 10(7), 1759–1777. <https://doi.org/10.1162/089976698300017124>
- Chen, Chen, Murphey, T. D., & MacIver, M. A. (2020). Tuning movement for sensing in an uncertain world. *ELife*, 9, e52371. <https://doi.org/10.7554/eLife.52371>
- Chen, Chinfai, & Regehr, W. G. (2000). Developmental remodeling of the retinogeniculate synapse. *Neuron*, 28(3), 955–966. [https://doi.org/10.1016/S0896-6273\(00\)00166-5](https://doi.org/10.1016/S0896-6273(00)00166-5)
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *37th International Conference on Machine Learning*, 1575–1585.
- Cho, K. K. A., Khibnik, L., Philpot, B. D., Bear, M. F., Huganir, R. L., Designed, M. F. B., & Performed, B. D. P. (2009). The ratio of NR2A/B NMDA receptor subunits determines the qualities of ocular dominance plasticity in visual cortex. *Proceedings of the National Academy of Sciences*, 31(13), 5377–5382.
- Collewijn, H., Kowler, E., R., B. P., H., C., H., C., N., C. T., ... L., B. (2008). The significance of microsaccades for vision and oculomotor control. *Journal of Vision*, 8(14), 20–20. <https://doi.org/10.1167/8.14.20>
- Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, 21(10), 576–586. <https://doi.org/10.1038/s41583-020-0355-6>
- Conwell, C., Buice, M. A., Alvarez, G. A., Katz, B., & Barbu, A. (2021). Neural Regression , Representational Similarity , Model Zoology & Neural Taskonomy at Scale in Rodent Visual Cortex. *Neural Information Processing Systems*, (NeurIPS 2021).
- Cooke, S. F., & Bear, M. F. (2010). Visual Experience Induces Long-Term Potentiation in the Primary Visual Cortex. *Journal of Neuroscience*, 30(48), 16304–16313. <https://doi.org/10.1523/JNEUROSCI.4333-10.2010>
- Cooke, S. F., & Bear, M. F. (2014). How the mechanisms of long-term synaptic

- potentiation and depression serve experience-dependent plasticity in primary visual cortex. *Philosophical Transactions of the Royal Society of London B*, 369. <https://doi.org/10.1098/rstb.2013.0284>
- Cooke, S. F., Komorowski, R. W., Kaplan, E. S., Gavornik, J. P., & Bear, M. F. (2015). Visual recognition memory, manifested as long-term habituation, requires synaptic plasticity in V1. *Nature Neuroscience*, 18(2), 262–271. <https://doi.org/10.1038/nn.3920>
- Cooper, L. N., & Bear, M. F. (2012). The BCM theory of synapse modification at 30: interaction of theory with experiment. *Nature Reviews Neuroscience*, 13(11), 798–810. <https://doi.org/10.1038/nrn3353>
- Cover, T. M., & Thomas, J. A. (2005). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing (2nd Ed.). <https://doi.org/10.1002/047174882X>
- Crandall, S. R., Cruikshank, S. J., & Connors, B. W. (2015). A Corticothalamic Switch: Controlling the Thalamus with Dynamic Synapses. *Neuron*, 86(3), 768–782. <https://doi.org/10.1016/j.neuron.2015.03.040>
- Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, 587–600. <https://doi.org/10.1038/nrn2457>
- Creutzig, F., & Sprekeler, H. (2008). Predictive coding and the slowness principle: An information-theoretic approach. *Neural Computation*, 20(4), 1026–1041. <https://doi.org/10.1162/neco.2008.01-07-455>
- Crozier, R. A., Wang, Y., Liu, C.-H., & Bear, M. F. (2007). Deprivation-induced synaptic depression by distinct mechanisms in different layers of mouse visual cortex. *Proceedings of the National Academy of Sciences*, 104(4), 1383–1388. Retrieved from [www.pnas.org/cgi/doi/10.1073/pnas.0609596104](http://www.pnas.org/cgi/doi/10.1073/pnas.0609596104)
- Dacey, D. M. (1994). Physiology, morphology and spatial densities of identified ganglion cell types in primate retina. *Ciba Foundation Symposium*, 184, 12–34. <https://doi.org/10.1002/9780470514610.ch2>
- Dan, Y., Atick, J. J., & Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *Journal of Neuroscience*, 16(10), 3351–3362. <https://doi.org/10.1523/jneurosci.16-10-03351.1996>
- Dayan, P., & Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press.

- De Valois, R L, Yund, E. W., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22(5), 531–544. [https://doi.org/10.1016/0042-6989\(82\)90112-2](https://doi.org/10.1016/0042-6989(82)90112-2)
- De Valois, Russell L, Morgan, H. C., & Snodderly, D. M. (1974). Psychophysical studies of monkey vision: III. Spatial luminance contrast sensitivity tests of macaque and human observers. *Vision Research*. Netherlands: Elsevier Science. [https://doi.org/10.1016/0042-6989\(74\)90118-7](https://doi.org/10.1016/0042-6989(74)90118-7)
- DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences*, 18(10), 451–458. [https://doi.org/10.1016/0166-2236\(95\)94496-R](https://doi.org/10.1016/0166-2236(95)94496-R)
- DeFelipe, J., López-Cruz, P. L., Benavides-Piccione, R., Bielza, C., Larrañaga, P., Anderson, S., ... Ascoli, G. A. (2013). New insights into the classification and nomenclature of cortical GABAergic interneurons. *Nature Reviews Neuroscience*, 14, 202–216. <https://doi.org/10.1038/nrn3444>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Computer Vision and Pattern Recognition*.
- Dennis, S. H., Pasqui, F., Colvin, E. M., Sanger, H., Mogg, A. J., Felder, C. C., ... Mellor, J. R. (2015). Activation of Muscarinic M1 Acetylcholine Receptors Induces Long-Term Potentiation in the Hippocampus. *Cerebral Cortex*, 26, 414–426. <https://doi.org/10.1093/cercor/bhv227>
- Dhande, O. S., Stafford, B. K., Lim, J.-H. A., & Huberman, A. D. (2015). Contributions of Retinal Ganglion Cells to Subcortical Visual Processing and Behaviors. *Annual Review of Vision Science*, 1, 291–328. <https://doi.org/10.1146/annurev-vision-082114-035502>
- Diamanti, E. M., Reddy, C. B., Schröder, S., Muzzu, T., Harris, K. D., Saleem, A. B., & Carandini, M. (2021). Spatial modulation of visual responses arises in cortex with active navigation. *ELife*, 10, 1–15. <https://doi.org/10.7554/elife.63705>
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434. <https://doi.org/10.1016/j.neuron.2012.01.010>
- Dipoppa, M., Ranson, A., Krumin, M., Pachitariu, M., Carandini, M., & Harris, K. D. (2018). Vision and Locomotion Shape the Interactions between Neuron Types in Mouse Visual Cortex. *Neuron*, 98(3), 1–14. <https://doi.org/10.1016/j.neuron.2018.03.037>
- Disney, A. A., Alasady, H. A., & Reynolds, J. H. (2014). Muscarinic acetylcholine

- receptors are expressed by most parvalbumin-immunoreactive neurons in area MT of the macaque. *Brain and Behavior*, 4(3), 431–445.  
<https://doi.org/10.1002/brb3.225>
- Disney, A. A., & Aoki, C. (2008). Muscarinic acetylcholine receptors in macaque V1 are most frequently expressed by parvalbumin-immunoreactive neurons. *Journal of Comparative Neurology*, 507(5), 1748–1762. <https://doi.org/10.1002/cne.21616>
- Doersch, C. (2016). Tutorial on Variational Autoencoders. *ArXiv*, 1–23. Retrieved from <http://arxiv.org/abs/1606.05908>
- Doi, E., Gauthier, J. L., Field, G. D., Shlens, J., Sher, A., Greschner, M., ... Simoncelli, E. P. (2012). Efficient coding of spatial information in the primate retina. *Journal of Neuroscience*, 32(46), 16256–16264. <https://doi.org/10.1523/JNEUROSCI.4036-12.2012>
- Dong, D., & Atick, J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2), 159–178. <https://doi.org/10.1088/0954-898x/6/2/003>
- Dong, D. W., & Atick, J. J. (1995). Statistics of Natural Time-Varying Images. *Network: Computation in Neural Systems*, 6(3), 345–358.
- Douglas, C. L., Baghdoyan, H. A., Lydic, R., & Baghdoyan, H. A. (2002). Prefrontal Cortex Acetylcholine Release, EEG Slow Waves, and Spindles Are Modulated by M2 Autoreceptors in C57BL/6J Mouse. *Journal of Neurophysiology*, 87, 2817–2822. <https://doi.org/10.1152/jn.01015.2001>
- Douglas, R. J., & Martin, K. A. C. (2004). Neuronal Circuits of the Neocortex. *Annual Review of Neuroscience*, 27(1), 419–451.  
<https://doi.org/10.1146/annurev.neuro.27.070203.144152>
- Douglas, R. J., Martin, K. A. C., & Whitteridge, D. (1989). A Canonical Microcircuit for Neocortex. *Neural Computation*, 1, 480–488.
- Dudek, S. M., & Bear, M. F. (1992). Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *Neurobiology*, 89, 4363–4367. Retrieved from <http://www.pnas.org/content/89/10/4363.long>
- Edelman, G. M., & Mountcastle, V. B. (1978). *The mindful brain: Cortical organization and the group-selective theory of higher brain function*. Oxford, England: MIT Press.
- El-Boustani, S., Ip, J. P. K., Breton-Provencher, V., Knott, G. W., Okuno, H., Bito, H., & Sur, M. (2018). Locally coordinated synaptic plasticity of visual cortex neurons in

- vivo. *Science*, 360. Retrieved from <http://science.sciencemag.org/content/sci/360/6395/1349.full.pdf>
- Elias, P. (1955). Predictive Coding—Part I & II. *IRE Transactions on Information Theory*, 1(1), 16–33. <https://doi.org/10.1109/TIT.1955.1055126>
- Espinosa, J. S., & Stryker, M. P. (2012a). Development and Plasticity of the Primary Visual Cortex. *Neuron*, 75(2). <https://doi.org/10.1016/j.neuron.2012.06.009>
- Espinosa, J. S., & Stryker, M. P. (2012b). Development and Plasticity of the Primary Visual Cortex. *Neuron*, 75(2), 230–249. <https://doi.org/10.1016/j.neuron.2012.06.009.Development>
- Failor, S. W., Carandini, M., & Harris, K. D. (2021). Learning orthogonalizes visual cortical population codes. *BioRxiv*, 1–24.
- Falconbridge, M. S., Stamps, R. L., & Badcock, D. R. (2006). A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images. *Neural Computation*, 18(2), 415–429. <https://doi.org/10.1162/089976606775093891>
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, 1, 1–47. Retrieved from <https://academic.oup.com/cercor/article/1/1/1/408896>
- Finnie, P. S. B., Komorowski, R. W., & Bear, M. F. (2021). The spatiotemporal organization of experience dictates hippocampal involvement in primary visual cortical plasticity. *Current Biology*, 31(18), 3996–4008. <https://doi.org/10.1101/2021.03.01.433430>
- Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V, Leinweber, M., & Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, (September). <https://doi.org/10.1038/nn.4385>
- Flindall, J. W., & Gonzalez, C. L. R. (2020). Revisiting Ungerleider and Mishkin: Two cortical visual systems. In *Brain and Behaviour: Revisiting the Classic Studies*. <https://doi.org/10.4135/9781529715064.n5>
- Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*. <https://doi.org/10.1007/BF02331346>
- Fong, M.-F., Finnie, P. S. B., Kim, T., Thomazeau, A., Kaplan, E. S., Cooke, S. F., & Bear, M. F. (2020). Distinct Laminar Requirements for NMDA Receptors in Experience-Dependent Visual Cortical Plasticity. *Cerebral Cortex*, 30, 2555–2572. <https://doi.org/10.1093/cercor/bhz260>

- Fournier, J., Saleem, A. B., Diamanti, E. M., Wells, M. J., Harris, K. D., & Carandini, M. (2020). Mouse Visual Cortex Is Modulated by Distance Traveled and by Theta Oscillations. *Current Biology*, *30*(19), 3811–3817.e6. <https://doi.org/10.1016/j.cub.2020.07.006>
- Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2006). Experience-Dependent Sharpening of Visual Shape Selectivity in Inferior Temporal Cortex. *Cerebral Cortex*, *16*, 1631–1644. <https://doi.org/10.1093/cercor/bhj100>
- Frenkel, M. Y., & Bear, M. F. (2004). How monocular deprivation shifts ocular dominance in visual cortex of young mice. *Neuron*, *44*(6), 917–923. <https://doi.org/10.1016/j.neuron.2004.12.003>
- Frenkel, M. Y., Sawtell, N. B., Cinira, A., Diogo, M., Yoon, B., Neve, R. L., & Bear, M. F. (2006). Instructive Effect of Visual Experience in Mouse Visual Cortex. *Neuron*, *51*, 339–349. <https://doi.org/10.1016/j.neuron.2006.06.026>
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 815–836. <https://doi.org/10.1098/rstb.2005.1622>
- Gandolfi, D., Bigiani, A., Porro, C. A., & Mapelli, J. (2020). Inhibitory Plasticity: From Molecules to Computation and Beyond. *International Journal of Molecular Sciences*, *21*(1805). <https://doi.org/10.3390/ijms21051805>
- Garner, A. R., & Keller, G. B. (2021). A cortical circuit for audio-visual predictions. *Nature Neuroscience*, *25*, 98–105. <https://doi.org/10.1038/s41593-021-00974-7>
- Garrett, M. E., Manavi, S., Roll, K., Ollerenshaw, D. R., Groblewski, P. A., Kiggins, J., ... Olsen, S. R. (2020). Experience shapes activity dynamics and stimulus coding of VIP inhibitory and excitatory cells in visual cortex. *ELife*, *9*, e50340. <https://doi.org/10.1101/686063>
- Garrett, M. E., Nauhaus, I., Marshel, J. H., & Callaway, E. M. (2014). Topography and areal organization of mouse visual cortex. *Journal of Neuroscience*, *34*(37), 12587–12600. <https://doi.org/10.1523/JNEUROSCI.1124-14.2014>
- Gavornik, J. P., & Bear, M. F. (2014). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature Neuroscience*, *17*(5), 732–737. <https://doi.org/10.1038/nn.3683>
- Gerstner, W., & Kistler, W. M. (2002). Mathematical formulations of Hebbian learning. *Biological Cybernetics*, *87*, 404–415. <https://doi.org/10.1007/s00422-002-0353-y>
- Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., ... Larkum,

- M. E. (2020). Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, *367*(6473). <https://doi.org/10.1126/science.aax6239>
- Gilbert, C. D. (1977). Laminar differences in receptive field properties of cells in cat primary visual cortex. *Journal of Physiology*, *268*, 391–421.
- Gillon, C. J., Pina, J. E., Lecoq, J. A., Ahmed, R., Billeh, Y., Caldejon, S., ... Zylberberg, J. (2021). Learning from unexpected events in the neocortical microcircuit. *BioRxiv*. <https://doi.org/10.1101/2021.01.15.426915>
- Gjorgjieva, J., Sompolinsky, H., & Meister, M. (2014). Benefits of Pathway Splitting in Sensory Coding. *Journal of Neuroscience*, *34*(36), 12127–12144. <https://doi.org/10.1523/JNEUROSCI.1032-14.2014>
- Goard, M., & Dan, Y. (2009). Basal forebrain activation enhances cortical coding of natural scenes. *Nature Neuroscience*, *12*(11), 1444–1449. <https://doi.org/10.1038/nn.2402>
- Goel, A., & Buonomano, D. V. (2014). Timing as an intrinsic property of neural networks: evidence from in vivo and in vitro experiments. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1637), 20120460. <https://doi.org/10.1098/rstb.2012.0460>
- Gómez-Palacio-Schjetnan, A., & Escobar, M. L. (2013). Neurotrophins and Synaptic Plasticity. *Current Topics in Behavioral Neuroscience*, *15*, 117–136. [https://doi.org/10.1007/7854\\_2012\\_231](https://doi.org/10.1007/7854_2012_231)
- Goodale, M. A., & Milner, A. D. (1992). Separate Visual Pathways for Perception and Action. *Trends in Neuroscience*, *20*(1), 20–25.
- Gordon, J. A., & Stryker, M. P. (1996). Experience-Dependent Plasticity of Binocular Responses in the Primary Visual Cortex of the Mouse. *Journal of Neuroscience*, *76*(10), 3274–3286.
- Gorrell, G. (2006). Generalized hebbian algorithm for incremental singular value decomposition in natural language processing. In *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*.
- Groleau, M., Kang, J. Il, Huppe-Gourgues, F., & Vaucher, E. (2015). Distribution and effects of the muscarinic receptor subtypes in the primary visual cortex. *Frontiers in Synaptic Neuroscience*, *7*(10), 1–9. <https://doi.org/10.3389/fnsyn.2015.00010>
- Guitchounts, G., Masís, J., Wolff, S. B. E., & Cox, D. (2020). Encoding of 3D Head Orienting Movements in the Primary Visual Cortex. *Neuron*, *108*(3), 512–525.e4.

<https://doi.org/10.1016/j.neuron.2020.07.014>

- Gulledge, A. T., Bucci, D. J., Zhang, S. S., Matsui, M., & Yeh, H. H. (2009). M1 receptors mediate cholinergic modulation of excitability in neocortical pyramidal neurons. *Journal of Neuroscience*, *29*(31), 9888–9902. <https://doi.org/10.1523/JNEUROSCI.1366-09.2009>
- Hammer, S., Monavarfeshani, A., Lemon, T., Su, J., & Fox, M. A. (2015). Multiple Retinal Axons Converge onto Relay Cells in the Adult Mouse Thalamus. *Cell Reports*, *12*(10), 1575–1583. <https://doi.org/10.1016/j.celrep.2015.08.003>
- Harris, J. A., Mihalas, S., Hirokawa, K. E., Whitesell, J. D., Choi, H., Bernard, A., ... Zeng, H. (2019). Hierarchical organization of cortical and thalamic connectivity. *Nature*, *575*(7781), 195–202. <https://doi.org/10.1038/s41586-019-1716-z>
- Harris, K. D., & Shepherd, G. (2015). The neocortical circuit: themes and variations. *Nature Neuroscience*, *18*(2), 170–181. <https://doi.org/10.1038/nn.3917>
- Hartveit, E. (1992). Simultaneous recording of lagged and nonlagged cells in the cat dorsal lateral geniculate nucleus. *Experimental Brain Research*, *88*, 229–232.
- Harvey, C. D., & Svoboda, K. (2007). Locally dynamic synaptic learning rules in pyramidal neuron dendrites. *Nature*, *450*(20). <https://doi.org/10.1038/nature06416>
- Hasselmo, M. E., & Stern, C. E. (2006). Mechanisms underlying working memory for novel information. *Trends in Cognitive Sciences*, *10*(11), 487–493. <https://doi.org/10.1016/j.tics.2006.09.005>
- Hawkins, J., & Ahmad, S. (2016). Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex. *Frontiers in Neural Circuits*, *10*(23). <https://doi.org/10.3389/fncir.2016.00023>
- Hayden, D. J., Montgomery, D. P., Cooke, S. F., & Bear, M. F. (2021). Visual recognition is heralded by shifts in local field potential oscillations and inhibitory networks in primary visual cortex. *Journal of Neuroscience*, *41*(29), 6257–6272. <https://doi.org/10.1523/JNEUROSCI.0391-21.2021>
- Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. Wiley. Oxford, England: Wiley.
- Henaff, O. J., Bai, Y., Charlton, J. A., Nauhaus, I., Simoncelli, E. P., & Goris, R. L. T. (2021). Primary visual cortex straightens natural video trajectories. *Nature Communications*, *12*(5982).
- Hénaff, O. J., Goris, R. L. T., & Simoncelli, E. P. (2019). Perceptual straightening of

- natural videos. *Nature Neuroscience*, 22, 984–991. <https://doi.org/10.1038/s41593-019-0377-4>
- Hendry, S. H. C., & Reid, R. C. (2000). The Koniocellular pathway in primate vision. *Annual Review of Neuroscience*, 23, 127–153. Retrieved from [www.annualreviews.org](http://www.annualreviews.org)
- Hennequin, G., Agnes, E. J., & Vogels, T. P. (2017). Inhibitory Plasticity: Balance, Control, and Codependence. *Annual Review of Neuroscience*, 40, 557–579. <https://doi.org/10.1146/annurev-neuro-072116>
- Hensch, T. K. (2005). Critical Period Plasticity in Local Cortical Circuits. *Nature Reviews Neuroscience*, 6, 877–888. <https://doi.org/10.1038/nrn1787>
- Hensch, T. K., & Quinlan, E. M. (2018). Critical periods in amblyopia. *Visual Neuroscience*, 35, E014. <https://doi.org/10.1017/S0952523817000219>
- Herrero, J. L., Roberts, M. J., Delicato, L. S., Gieselmann, M. A., Dayan, P., & Thiele, A. (2008). Acetylcholine contributes through muscarinic receptors to attentional modulation in V1. *Nature*, 454(7208), 1110–1114. <https://doi.org/10.1038/nature07141>
- Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2020). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. *Nature Communications*, 12(6456), 1–14. <https://doi.org/10.1038/s41467-021-26751-5>
- Hinton, G., & Sejnowski, T. J. (Eds.). (1999). *Unsupervised Learning: Foundations of Neural Computation*. The MIT Press. <https://doi.org/10.7551/mitpress/7011.001.0001>
- Hofer, S. B., Ko, H., Pichler, B., Vogelstein, J., Ros, H., Zeng, H., ... Mrsic-Flogel, T. D. (2011). Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nature Neuroscience*, 14(8). <https://doi.org/10.1038/nn.2876>
- Hofer, S. B., Mrsic-Flogel, T. D., Bonhoeffer, T., & Hübener, M. (2006). Lifelong learning: ocular dominance plasticity in mouse visual cortex. *Current Opinion in Neurobiology*, 16(4), 451–459. <https://doi.org/10.1016/j.conb.2006.06.007>
- Homann, J., Koay, S. A., Glidden, A. M., Tank, D. W., & Berry, M. J. (2017). Predictive Coding of Novel versus Familiar Stimuli in the Primary Visual Cortex. *BioRxiv*. <https://doi.org/10.1101/197608>
- Hooks, B. M., & Chen, C. (2020). Circuitry Underlying Experience-Dependent Plasticity

- in the Mouse Visual System. *Neuron*, *106*(1), 21–36.  
<https://doi.org/10.1016/j.neuron.2020.01.031>
- Hoon, M., Okawa, H., Della Santina, L., & Wong, R. O. (2014). Functional Architecture of the Retina: Development and Disease. *Progress in Retinal and Eye Research*, *42*, 44–84. <https://doi.org/10.1016/j.preteyeres.2014.06.003>
- Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*, 71–77. <https://doi.org/10.1038/nature03689>
- Hu, H., Gan, J., & Jonas, P. (2014). Fast-spiking, parvalbumin+ GABAergic interneurons: From cellular design to microcircuit function. *Science*, *345*(6196). <https://doi.org/10.1126/science.1255263>
- Hubel, D H, & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.
- Hubel, David H., & Wiesel, T. N. (1964). Effects of monocular deprivation in kittens. *Naunyn-Schmiedebergs Archiv Für Experimentelle Pathologie Und Pharmakologie*, *248*(6), 492–497. <https://doi.org/10.1007/BF00348878>
- Hubel, David H, & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*(1), 106–154.2. <https://doi.org/10.1523/JNEUROSCI.1991-09.2009>
- Hubel, David H, & Wiesel, T. N. (1965). Binocular interaction in striate cortex of kittens reared with artificial squint. *Journal of Physiology*, *28*, 1041–1059.
- Hubel, David H, & Wiesel, T. N. (1969). Anatomical Demonstration of Columns in the Monkey Striate Cortex. *Nature*, *221*, 747–750.
- Hubel, David H, Wiesel, T. N., & LeVay, S. (1977). Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *278*, 377–409. Retrieved from <https://royalsocietypublishing.org/>
- Huberman, A. D., & McAllister, A. K. (2002). Neurotrophins and visual cortical plasticity. *Progress in Brain Research*, *138*, 39–51. [https://doi.org/10.1016/S0079-6123\(02\)38069-5](https://doi.org/10.1016/S0079-6123(02)38069-5)
- Hughes, H., Schwartz, O., Pillow, J. W., Rust, N. C., & Simoncelli, E. P. (2006). Spike-triggered neural characterization. *Journal of Vision*, *6*, 484–507. <https://doi.org/10.1167/6.4.13>
- Huppé-Gourgues, F., Jegouic, K., & Vaucher, E. (2018). Topographic Organization of

- Cholinergic Innervation From the Basal Forebrain to the Visual Cortex in the Rat. *Frontiers in Neural Circuits*, 12(19). <https://doi.org/10.3389/fncir.2018.00019>
- Intrator, N., & Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5(1), 3–17. [https://doi.org/10.1016/S0893-6080\(05\)80003-6](https://doi.org/10.1016/S0893-6080(05)80003-6)
- Ishikawa, M., Otaka, M., Huang, Y. H., Neumann, P. A., Winters, B. D., Grace, A. A., ... Dong, Y. (2013). Dopamine triggers heterosynaptic plasticity. *Journal of Neuroscience*, 33(16), 6759–6765. <https://doi.org/10.1523/JNEUROSCI.4694-12.2013>
- Issa, N. P., Trachtenberg, J. T., Chapman, B., Zahs, K. R., & Stryker, M. P. (1999). The Critical Period for Ocular Dominance Plasticity in the Ferret's Visual Cortex. *Journal of Neuroscience*, 19(16), 6965–6978.
- Ji, W., Gămănuț, R., Bista, P., D'Souza, R. D., Wang, Q., & Burkhalter, A. (2015). Modularity in the Organization of Mouse Primary Visual Cortex. *Neuron*, 87(3), 632–643. <https://doi.org/10.1016/j.neuron.2015.07.004>
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., Hudspeth, A. J., & Mack, S. (2014). *Principles of Neural Science*. McGraw-Hill Companies (5th ed.).
- Kaneko, M., Fu, Y., & Stryker, M. P. (2017). Locomotion Induces Stimulus-Specific Response Enhancement in Adult Visual Cortex. *Journal of Neuroscience*, 37(13). <https://doi.org/10.1523/JNEUROSCI.3760-16.2017>
- Kaneko, M., Hanover, J. L., England, P. M., Stryker, M. P., & Keck, W. M. (2008). TrkB kinase is required for recovery, but not loss, of cortical responses following monocular deprivation. *Nature Neuroscience*, 11(4). <https://doi.org/10.1038/nn2068>
- Kaneko, M., Stellwagen, D., Malenka, R. C., & Stryker, M. P. (2008). Tumor Necrosis Factor- $\alpha$  Mediates One Component of Competitive, Experience-Dependent Plasticity in Developing Visual Cortex. *Neuron*, 58(5), 673–680. <https://doi.org/10.1016/j.neuron.2008.04.023>
- Kaplan, E. S., Cooke, S. F., Komorowski, R. W., Chubykin, A. A., Thomazeau, A., Khibnik, L. A., ... Dan, Y. (2016). Contrasting roles for parvalbumin-expressing inhibitory neurons in two forms of adult visual cortical plasticity. *ELife*, 5, 218–221. <https://doi.org/10.7554/eLife.11450>
- Kaplan, E., & Shapley, R. M. (1986). The primate retina contains two types of ganglion cells, with high and low contrast sensitivity (spatial vision/visual neurons/macaque monkey). *Proceedings of the National Academy of Sciences*, 83, 2755–2757.

- Karmarkar, U. R., & Dan, Y. (2006). Experience-Dependent Plasticity in Adult Visual Cortex. *Neuron*, *52*, 577–585. <https://doi.org/10.1016/j.neuron.2006.11.001>
- Karnani, M. M., Jackson, J., Ayzenshtat, I., Hamzehei Sichani, A., Manoocheri, K., Kim, S., & Yuste, R. (2016). Opening Holes in the Blanket of Inhibition: Localized Lateral Disinhibition by VIP Interneurons. *Journal of Neuroscience*, *36*(12), 3471–3480. <https://doi.org/10.1523/JNEUROSCI.3646-15.2016>
- Karnani, M. M., Jackson, J., Ayzenshtat, I., Tucciarone, J., Manoocheri, K., Snider, W. G. G., & Yuste, R. (2016). Cooperative Subnetworks of Molecularly Similar Interneurons in Mouse Neocortex. *Neuron*, *90*(1), 86–100. <https://doi.org/10.1016/j.neuron.2016.02.037>
- Katzner, S., Nauhaus, I., Benucci, A., Bonin, V., Ringach, D. L., & Carandini, M. (2009). Local Origin of Field Potentials in Visual Cortex. *Neuron*, *61*(1), 35–41. <https://doi.org/10.1016/j.neuron.2008.11.016>
- Keck, T., Toyozumi, T., Chen, L., Doiron, B., Feldman, D. E., Fox, K., ... Van Rossum, M. C. (2017). Integrating Hebbian and homeostatic plasticity: the current state of the field and future research directions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*. <https://doi.org/10.1098/rstb.2016.0158>
- Keller, G. B., Bonhoeffer, T., & Hübener, M. (2012). Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. *Neuron*, *74*(5), 809–815. <https://doi.org/10.1016/j.neuron.2012.03.040>
- Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435. <https://doi.org/10.1016/j.neuron.2018.10.003>
- Kelly, D. H. (1985). Visual processing of moving stimuli. *Journal of the Optical Society of America*, *2*(2), 216–225.
- Kepecs, A., & Fishell, G. (2014). Interneuron cell types are fit to function. *Nature*, *505*(7483), 318–326. <https://doi.org/10.1038/nature12983>
- Kerschensteiner, D., & Guido, W. (2017). Organization of the dorsal lateral geniculate nucleus in the mouse. *Visual Neuroscience*, *34*, e008. <https://doi.org/10.1017/S0952523817000062>
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.
- Kirkwood, A., Rioult, M. G., & Bear, M. F. (1996). Experience-dependent modification

of synaptic plasticity in visual cortex. *Nature*, 381, 526–528.

- Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2008). Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12), 4018. <https://doi.org/10.1371/journal.pone.0004018>
- Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., ... Machens, C. K. (2016). Demixed principal component analysis of neural population data. *ELife*, 5, e10989. <https://doi.org/10.7554/eLife.10989>
- Kolb, H. (2003). How the Retina Works. *American Scientist*, 91, 28–34. <https://doi.org/10.1511/2003.1.28>
- Komatsu, Y. (1996). GABA(B) receptors, monoamine receptors, and postsynaptic inositol trisphosphate-induced Ca<sup>2+</sup> release are involved in the induction of long-term potentiation at visual cortical inhibitory synapses. *Journal of Neuroscience*, 16(20), 6342–6352. <https://doi.org/10.1523/jneurosci.16-20-06342.1996>
- Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971), 244–247. <https://doi.org/10.1038/nature02169>
- Krahe, T. E., El-Danaf, R. N., Dilger, E. K., Henderson, S. C., & Guido, W. (2011). Morphologically Distinct Classes of Relay Cells Exhibit Regional Preferences in the Dorsal Lateral Geniculate Nucleus of the Mouse. *Journal of Neuroscience*, 31(48), 17437–17448. <https://doi.org/10.1523/JNEUROSCI.4370-11.2011>
- Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1, 417–446. <https://doi.org/10.1146/annurev-vision-082114-035447>
- Kuang, X., Poletti, M., Victor, J. D., & Rucci, M. (2012). Temporal encoding of spatial information during active visual fixation. *Current Biology*, 22(6), 510–514. <https://doi.org/10.1016/j.cub.2012.01.050>
- Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*. <https://doi.org/10.1214/aoms/1177729694>
- Kuo, M. C., Rasmusson, D. D., & Dringenberg, H. C. (2009). Input-selective potentiation and rebalancing of primary sensory cortex afferents by endogenous acetylcholine. *Neuroscience*, 163(1), 430–441. <https://doi.org/10.1016/j.neuroscience.2009.06.026>
- Kuśmierz, Ł., Isomura, T., & Toyozumi, T. (2017). Learning with three factors: modulating Hebbian plasticity with errors. *Current Opinion in Neurobiology*, 46, 170–177. <https://doi.org/10.1016/j.conb.2017.08.020>

- Lappe, M., Bremmer, F., & Van Den Berg, A. V. (1999). Perception of self-motion from visual flow. *Trends in Cognitive Sciences*, 3(9). [https://doi.org/10.1016/S1364-6613\(99\)01364-9](https://doi.org/10.1016/S1364-6613(99)01364-9)
- Larsen, R. S., & Sjöström, P. J. (2015). Synapse-type-specific plasticity in local circuits. *Current Opinion in Neurobiology*, 35, 127–135. <https://doi.org/10.1016/j.conb.2015.08.001>
- Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, 11(4), 475–480. [https://doi.org/10.1016/S0959-4388\(00\)00237-3](https://doi.org/10.1016/S0959-4388(00)00237-3)
- Laughlin, Simon B., De Ruyter Van Steveninck, R. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, 1(1), 36–41. <https://doi.org/10.1038/236>
- Law, M. I., Zaksas, K. R., & Stryker, M. P. (1988). Organization of Primary Visual Cortex (Area 17) in the Ferret. *The Journal of Comparative Neurology*, 278, 157–180.
- Leinweber, M., Ward, D. R., Sobczak, J. M., Attinger, A., & Keller, G. B. (2017). A Sensorimotor Circuit in Mouse Cortex for Visual Flow Predictions. *Neuron*, 95(6), 1420–1432.e5. <https://doi.org/10.1016/j.neuron.2017.08.036>
- Lien, A. D., & Scanziani, M. (2013). Tuned thalamic excitation is amplified by visual cortical circuits. *Nature Neuroscience*, 16(9), 1315–1323. <https://doi.org/10.1038/nn.3488>
- Lien, A. D., & Scanziani, M. (2018). Cortical direction selectivity emerges at convergence of thalamic synapses. *Nature*, 558(7708), 80–86. <https://doi.org/10.1038/s41586-018-0148-5>
- Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10), 2017–2031. [https://doi.org/10.1162/jocn\\_a\\_01544](https://doi.org/10.1162/jocn_a_01544)
- Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A Unified Theory of Early Visual Representations from Retina to Cortex Through Anatomically Constrained Deep CNNs. *International Conference on Learning*. Retrieved from <https://github.com/ganguli-lab/RetinalResources>.
- Lynch, G., Dunwiddie, T., & Gribkoff, V. (1977). Heterosynaptic depression: a postsynaptic correlate of long-term potentiation. *Nature*, 266, 737–739.
- Lyu, S. (2010). Divisive normalization: Justification and effectiveness as efficient coding transform. *Advances in Neural Information Processing Systems*, 1–9.

- Maheu, M., Dehaene, S., & Meyniel, F. (2019). Brain signatures of a multiscale process of sequence learning in humans. *ELife*, *8*, e41541.
- Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An Embarrassment of Riches. *Neuron*, *44*, 5–21. Retrieved from [https://learn.bu.edu/bbcswebdav/pid-5377688-dt-content-rid-18971475\\_1/courses/17fallengbe710\\_a1/LTP and LTD- An Embarrassment of Riches%281%29.pdf](https://learn.bu.edu/bbcswebdav/pid-5377688-dt-content-rid-18971475_1/courses/17fallengbe710_a1/LTP%20and%20LTD-An%20Embarrassment%20of%20Riches.pdf)
- Markram, H., Gerstner, W., & Sjöström, P. J. (2012). Spike-timing-dependent plasticity: A comprehensive overview. *Frontiers in Synaptic Neuroscience*, *4*(2), 1–3. <https://doi.org/10.3389/fnsyn.2012.00002>
- Markram, Henry, Luebke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. *Science*, *275*, 213–215.
- Marshel, J. H., Garrett, M. E., Nauhaus, I., & Callaway, E. M. (2011). Functional specialization of seven mouse visual cortical areas. *Neuron*, *72*(6), 1040–1054. <https://doi.org/10.1016/j.neuron.2011.12.004>
- Masland, R. H. (2001). The fundamental plan of the retina. *Nature Neuroscience*, *4*(9), 877–886. <https://doi.org/10.1038/nn0901-877>
- Mauk, M. D., & Buonomano, D. V. (2004). The Neural Basis of Temporal Processing. *Annual Review of Neuroscience*, *27*, 307–340. <https://doi.org/10.1146/annurev.neuro.27.070203.144247>
- Mauss, A. S., Vlasits, A., Borst, A., & Feller, M. (2017). Visual Circuits for Direction Selectivity. *Annual Review of Neuroscience*, *40*, 211–230. <https://doi.org/10.1146/annurev-neuro-072116>
- McCurry, C. L., Shepherd, J. D., Tropea, D., Wang, K. H., Bear, M. F., & Sur, M. (2010). Loss of Arc renders the visual cortex impervious to the effects of sensory experience or deprivation. *Nature Neuroscience*, *13*(4), 450–457. <https://doi.org/10.1038/nn.2508>
- McNamee, D., & Wolpert, D. M. (2019). Internal Models in Biological Control. *Annual Review of Control, Robotics, and Autonomous Systems*, *2*(1), 339–364. <https://doi.org/10.1146/annurev-control-060117-105206>
- Meister, M., & Berry, M. J. (1999). The Neural Code of the Retina. *Neuron*, *22*, 435–450.
- Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human Inferences about Sequences: A Minimal Transition Probability Model. *PLOS Computational Biology*, *12*(12), e1005260. <https://doi.org/10.1371/journal.pcbi.1005260>

- Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed. *Neuropsychologia*, *46*(3), 774–785.  
<https://doi.org/10.1016/j.neuropsychologia.2007.10.005>
- Minces, V. H., Alexander, A. S., Datlow, M., Alfonso, S. I., & Chiba, A. A. (2013). The role of visual cortex acetylcholine in learning to discriminate temporally modulated visual stimuli. *Frontiers in Behavioral Neuroscience*, *7*(16).  
<https://doi.org/10.3389/fnbeh.2013.00016>
- Modirshanechi, A., Kiani, M. M., & Aghajan, H. (2019). Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks. *NeuroImage*, *196*, 302–317.  
<https://doi.org/10.1016/J.NEUROIMAGE.2019.04.028>
- Montague, R. P., & Sejnowski, T. J. (1994). The Predictive Brain: Temporal coincidence and temporal order in synaptic learning mechanisms. *Learning & Memory*, *1*(1), 1–33.
- Montgomery, D. P., Hayden, D. J., Chaloner, F. A., Cooke, S. F., & Bear, M. F. (2022). Stimulus-Selective Response Plasticity in Primary Visual Cortex: Progress and Puzzles. *Frontiers in Neural Circuits*, *15*(815554), 1–18.  
<https://doi.org/10.3389/fncir.2021.815554>
- Mrsic-Flogel, T. D., Hofer, S. B., Ohki, K., Reid, R. C., Bonhoeffer, T., & Hübener, M. (2007). Homeostatic Regulation of Eye-Specific Responses in Visual Cortex during Ocular Dominance Plasticity. *Neuron*, *54*(6), 961–972.  
<https://doi.org/10.1016/j.neuron.2007.05.028>
- Nassar, C. (2001). Source Coding and Decoding. In *Telecommunications Demystified* (pp. 61–114). LLH Technology Publishing. <https://doi.org/10.1016/b978-0-08-051867-1.50010-x>
- Nicoll, R. A. (2017). A Brief History of Long-Term Potentiation. *Neuron*, *93*.  
<https://doi.org/10.1016/j.neuron.2016.12.015>
- Niell, C. M., & Stryker, M. P. (2008). Highly Selective Receptive Fields in Mouse Visual Cortex. *Journal of Neuroscience*, *28*(30), 7520–7536.  
<https://doi.org/10.1523/JNEUROSCI.0623-08.2008>
- Niell, Christopher M. (2015). Cell Types, Circuits, and Receptive Fields in the Mouse Visual Cortex. *Annual Review of Neuroscience*, *38*, 413–431.  
<https://doi.org/10.1146/annurev-neuro-071714-033807>
- Niven, J. E., & Laughlin, S. B. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*.  
<https://doi.org/10.1242/jeb.017574>

- Ocko, S. A., Lindsey, J., Ganguli, S., & Deny, S. (2018). The emergence of multiple retinal cell types through efficient coding of natural movies. *Neural Information Processing Systems*, 32. Retrieved from <https://github.com/ganguli-lab/RetinalCellTypes>.
- Ohki, K., Chung, S., Ch, Y. H., Kara, P., & Clay Reid, R. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433, 597–603. Retrieved from [www.nature.com/nature](http://www.nature.com/nature)
- Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*. <https://doi.org/10.1007/BF00275687>
- Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*. [https://doi.org/10.1016/S0893-6080\(05\)80089-9](https://doi.org/10.1016/S0893-6080(05)80089-9)
- Oja, E. (2002). Unsupervised learning in neural computation. *Theoretical Computer Science*, 287(1), 187–207. [https://doi.org/10.1016/S0304-3975\(02\)00160-3](https://doi.org/10.1016/S0304-3975(02)00160-3)
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 607–609. <https://doi.org/10.1038/381607a0>
- Olshausen, B. A., & Field, D. J. (1997). Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research*, 37(23), 3311–3325.
- Oord, A. van den, Li, Y., & Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *ArXiv*. Retrieved from <http://arxiv.org/abs/1807.03748>
- Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. *Neuron*, 92(2), 530–543. <https://doi.org/10.1016/J.NEURON.2016.09.038>
- Palmer, S. E., Marre, O., Berry, M. J., & Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22), 6908–6913. <https://doi.org/10.1073/pnas.1506855112>
- Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*, 15, 1191–1253. Retrieved from <http://www.stat.berkeley.edu/~binyu/summer08/L2P2.pdf>
- Park, I. M., & Pillow, J. W. (2011). Bayesian Spike-Triggered Covariance Analysis. *Advances in Neural Information Processing Systems*, 24, 1692–1700. Retrieved from [http://pillowlab.princeton.edu/pubs/ParkI\\_Pillow\\_BSTC\\_NIPS2011.pdf](http://pillowlab.princeton.edu/pubs/ParkI_Pillow_BSTC_NIPS2011.pdf)

- Park, M., & Pillow, J. W. (2011). Receptive Field Inference with Localized Priors. *PLoS Computational Biology*, 7(10), e1002219. <https://doi.org/10.1371/journal.pcbi.1002219>
- Pastuzyn, E. D., Day, C. E., Kearns, R. B., Kyrke-Smith, M., Taibi, A. V., McCormick, J., ... Shepherd, J. D. (2018). The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell*, 172(1–2), 275–288.e18. <https://doi.org/10.1016/j.cell.2017.12.024>
- Peddie, W. (1925). Helmholtz's Treatise on Physiological Optics. *Nature*, 116(2907), 88–89. <https://doi.org/10.1038/116088a0>
- Pehlevan, C., & Chklovskii, D. B. (2019). Neuroscience-Inspired Online Unsupervised Learning Algorithms: Artificial neural networks. *IEEE Signal Processing Magazine*, 36(6), 88–96. <https://doi.org/10.1109/MSP.2019.2933846>
- Perge, J. A., Koch, K., Miller, R., Sterling, P., & Balasubramanian, V. (2009). How the optic nerve allocates space, energy capacity, and information. *Journal of Neuroscience*, 29(24), 7917–7928. <https://doi.org/10.1523/JNEUROSCI.5200-08.2009>
- Pfeffer, C. K., Xue, M., He, M., Josh Huang, Z., & Scanziani, M. (2013). Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, 16(8). <https://doi.org/10.1038/nn.3446>
- Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*. <https://doi.org/10.1038/nature07140>
- Pinto, L., Goard, M. J., Estandian, D., Xu, M., Kwan, A. C., Lee, S.-H., ... Dan, Y. (2013). Fast modulation of visual perception by basal forebrain cholinergic neurons. *Nature Neuroscience*, 16(12), 1857–1863. <https://doi.org/10.1038/nn.3552>
- Pitkow, X., & Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nature Neuroscience*, 15(4). <https://doi.org/10.1038/nn.3064>
- Poort, J., Khan, A. G., Pachitariu, M., Nemri, A., Orsolich, I., Krupic, J., ... Hofer, S. B. (2015). Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex. *Neuron*, 86(6), 1478–1490. <https://doi.org/10.1016/j.neuron.2015.05.037>
- Price, B. H., Jensen, C. M., Khoudary, A. A., & Gavornik, J. P. (2022). Expectation violations produce error signals in mouse V1. *BioRxiv*, 1–23.
- Priebe, N. J., & Ferster, D. (2012). Mechanisms of Neuronal Computation in Mammalian

- Visual Cortex. *Neuron*, 75(2), 194–208.  
<https://doi.org/10.1016/j.neuron.2012.06.011>
- Rajan Dasgupta, X., Seibt, F., & Beierlein, X. M. (2018). Synaptic Release of Acetylcholine Rapidly Suppresses Cortical Activity by Recruiting Muscarinic Receptors in Layer 4. *Journal of Neuroscience*, 38(23).  
<https://doi.org/10.1523/JNEUROSCI.0566-18.2018>
- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. <https://doi.org/10.1038/4580>
- Rao, R. P. N., & Sejnowski, T. J. (2001). Predictive learning of temporal sequences in recurrent neocortical circuits. *Complexity in Biological Information Processing*.
- Rasmusson, D. D. (2000). The role of acetylcholine in cortical synaptic plasticity. *Behavioural Brain Research*, 115(2), 205–218. [https://doi.org/10.1016/S0166-4328\(00\)00259-X](https://doi.org/10.1016/S0166-4328(00)00259-X)
- Rauschecker, J. P. (2018). Where, When, and How: Are they all sensorimotor? Towards a unified view of the dorsal pathway in vision and audition. *Cortex*, 98, 262–268.  
<https://doi.org/10.1016/j.cortex.2017.10.020>
- Reid, R. C., & Alonso, J.-M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature*, 378(16), 281–284.
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, 2(11), 1019–1025. Retrieved from <http://neurosci.nature.com>
- Ringach, D L. (2004). Mapping receptive fields in primary visual cortex. *Journal of Physiology*, 558(3), 717–728. <https://doi.org/10.1113/jphysiol.2004.065771>
- Ringach, Dario L., Mineault, P. J., Tring, E., Olivas, N. D., Garcia-Junco-Clemente, P., & Trachtenberg, J. T. (2016). Spatial clustering of tuning in mouse primary visual cortex. *Nature Communications*, 7, 1–9. <https://doi.org/10.1038/ncomms12270>
- Rompani, S. B., Müllner, F. E., Wanner, A., Zhang, C., Roth, C. N., Yonehara, K., & Roska, B. (2017). Different Modes of Visual Integration in the Lateral Geniculate Nucleus Revealed by Single-Cell-Initiated Transsynaptic Tracing. *Neuron*, 93(4), 767–776.e6. <https://doi.org/10.1016/j.neuron.2017.01.028>

- Rossi, L. F., Harris, K. D., & Carandini, M. (2020). Spatial connectivity matches direction selectivity in visual cortex. *Nature*, *588*, 648–652. <https://doi.org/10.1038/s41586-020-2894-4>
- Rucci, M. (2008). Fixational eye movements, natural image statistics, and fine spatial vision. *Network: Computation in Neural Systems*, *19*(4), 253–285. <https://doi.org/10.1080/09548980802520992>
- Rucci, M., & Victor, J. D. (2015). The unsteady eye: an information- processing stage, not a bug. *Trends in Neurosciences*, *38*(4), 195–206. <https://doi.org/10.1016/j.tins.2015.01.005>
- Rudy, B., Fishell, G., Lee, S. H., & Hjerling-Leffler, J. (2011). Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Developmental Neurobiology*, *71*(1), 45–61. <https://doi.org/10.1002/dneu.20853>
- Russo, A. A., Bittner, S. R., Perkins, S. M., Seely, J. S., London, B. M., Lara, A. H., ... Churchland, M. M. (2018). Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. *Neuron*, *97*(4), 953-966.e8. <https://doi.org/10.1016/j.neuron.2018.01.004>
- Rust, N. C., & Dicarlo, J. J. (2010). Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. <https://doi.org/10.1523/JNEUROSCI.0179-10.2010>
- Rust, N. C., & Schwartz, O. (2005). Spatiotemporal Elements of Macaque V1 Receptive Fields. *Neuron*, *46*, 945–956. <https://doi.org/10.1016/j.neuron.2005.05.021>
- Saar, D., Grossman, Y., & Barkai, E. (2001). Long-lasting cholinergic modulation underlies rule learning in rats. *Journal of Neuroscience*, *21*(4), 1385–1392. <https://doi.org/21/4/1385> [pii]
- Sadahiro, M., Sajo, M., & Morishita, H. (2016). Nicotinic regulation of experience-dependent plasticity in visual cortex. *Journal of Physiology - Paris*. <https://doi.org/10.1016/j.jphysparis.2016.11.003>
- Sale, A., Berardi, N., Spolidoro, M., Baroncelli, L., & Maffei, L. (2010). GABAergic inhibition in visual cortical plasticity. *Frontiers in Cellular Neuroscience*, *4*(10). <https://doi.org/10.3389/fncel.2010.00010>
- Salgado, H., Bellay, T., Nichols, J. A., Bose, M., Martinolich, L., Perrotti, L., & Atzori, M. (2007). Muscarinic M2 and M1 Receptors Reduce GABA Release by Ca<sup>2+</sup> Channel Modulation Through Activation of PI 3 K/Ca<sup>2+</sup>-Independent and PLC/Ca<sup>2+</sup>-Dependent PKC. *Journal of Neurophysiology*, *98*, 952–965.

<https://doi.org/10.1152/jn.00060.2007>

- Sanger, T. D. (1989). Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network. *Neural Networks*, 2, 459–473.
- Sarkar, S., Reyes, C. M., Jensen, C. M., & Gavornik, J. P. (2022). M2 receptors are required for spatiotemporal sequence learning in mouse primary visual cortex. *BioRxiv*.
- Sarter, M., Hasselmo, M. E., Bruno, J. P., & Givens, B. (2005). Unraveling the attentional functions of cortical cholinergic inputs: Interactions between signal-driven and cognitive modulation of signal detection. *Brain Research Reviews*, 48(1), 98–111. <https://doi.org/10.1016/j.brainresrev.2004.08.006>
- Sato, M., & Stryker, M. P. (2008). Distinctive Features of Adult Ocular Dominance Plasticity. *Journal of Neuroscience*, 28(41), 10278–10286. <https://doi.org/10.1523/JNEUROSCI.2451-08.2008>
- Saul, A. B., & Humphrey, A. L. (1990). Spatial and temporal response properties of lagged and nonlagged cells in cat lateral geniculate nucleus. *Journal of Neurophysiology*, 64(1), 206–224. <https://doi.org/10.1152/jn.1990.64.1.206>
- Scholl, B., Connon, T. I., Ryan, M. A., Kamasawa, N., & Fitzpatrick, D. (2021). Cortical response selectivity derives from strength in numbers of synapses. *Nature*, 590, 111–114. <https://doi.org/10.1038/s41586-020-03044-3>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- Schultz, Wolfram. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, 18(1), 23–32. <https://doi.org/10.1038/nrn.2015.26>
- Schwartz, G. (Ed.). (2021). *Retinal Computation*. Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-819896-4.09993-5>
- Seabrook, T. A., Burbridge, T. J., Crair, M. C., & Huberman, A. D. (2017). Architecture, Function, and Assembly of the Mouse Visual System. *Annual Review of Neuroscience*, 40(1), 499–538. <https://doi.org/10.1146/annurev-neuro-071714-033842>
- Segal, M., Okabe, S., Kulik, A., De Chevigny, A., Runge, K., & Cardoso, C. (2020). Dendritic Spine Plasticity: Function and Mechanisms. *Frontiers in Synaptic Neuroscience*, 12(36). <https://doi.org/10.3389/fnsyn.2020.00036>

- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(4), 623–656. <https://doi.org/10.1002/j.1538-7305.1948.tb00917.x>
- Shatz, C. J., & Stryker, M. P. (1978). Ocular dominance in layer IV of the cat's visual cortex and the effects of monocular deprivation. *Journal of Physiology*, 281, 267–283.
- Shuler, M. G., & Bear, M. F. (2006). Reward Timing in the Primary Visual Cortex. *Science*, 311(21), 393–396. <https://doi.org/10.1126/science.1121879>
- Sidorov, M. S., Kim, H., Rougie, M., Williams, B., Siegel, J. J., Gavornik, J. P., & Philpot, B. D. (2020). Visual Sequences Drive Experience-Dependent Plasticity in Mouse Anterior Cingulate Cortex. *Cell Reports*, 32(11), 108152. <https://doi.org/10.1016/j.celrep.2020.108152>
- Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., ... Koch, C. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592, 86–92. <https://doi.org/10.1038/s41586-020-03171-x>
- Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2), 144–149. [https://doi.org/10.1016/S0959-4388\(03\)00047-3](https://doi.org/10.1016/S0959-4388(03)00047-3)
- Singer, Y., Teramoto, Y., Willmore, B. D. B., King, A. J., Schnupp, J. W. H., & Harper, N. S. (2018). Sensory cortex is optimised for prediction of future input. *ELife*, 7, 1–31. <https://doi.org/10.7554/eLife.31557>
- Sjöström, P. J., Turrigiano, G. G., & Nelson, S. B. (2001). Rate, Timing, and Cooperativity Jointly Determine Cortical Synaptic Plasticity. *Neuron*, 32, 1149–1164. [https://doi.org/10.1016/S0896-6273\(01\)00542-6](https://doi.org/10.1016/S0896-6273(01)00542-6)
- Sorg, B. A., Berretta, X. S., Blacktop, J. M., Fawcett, J. W., Kitagawa, X. H., Jessica, X., ... Miquel, X. M. (2016). Casting a Wide Net: Role of Perineuronal Nets in Neural Plasticity. *Journal of Neuroscience*, 36(45), 11459–11468. <https://doi.org/10.1523/JNEUROSCI.2351-16.2016>
- Spratling, M.W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, 112, 92–97. <https://doi.org/10.1016/j.bandc.2015.11.003>
- Spratling, Michael W. (2010). Predictive Coding as a Model of Response Properties in Cortical Area V1. *Journal of Neuroscience*. <https://doi.org/10.1523/JNEUROSCI.4911-09.2010>
- Spruston, N. (2008). Pyramidal neurons: dendritic structure and synaptic integration.

- Nature Reviews Neuroscience*, 9, 206–221. <https://doi.org/10.1038/nrn2286>
- Srinivasan, M. V, Laughlin, S. B., & Dubs, A. (1982). Predictive Coding: A Fresh View of Inhibition in the Retina. *Proceedings of the Royal Society of London*, 216(1205), 427–459. Retrieved from <https://www.jstor.org/stable/pdf/35861.pdf?refreqid=excelsior%3A7dd3e48784f2d65ab0b45c6ec9f0f51b>
- Stellwagen, D., & Malenka, R. C. (2006). Synaptic scaling mediated by glial TNF- $\alpha$ . *Nature*, 440(20), 1054–1059. <https://doi.org/10.1038/nature04671>
- Sterling, P., & Laughlin, S. (2015). *Principles of Neural Design*. Cambridge, MA: The MIT Press. <https://doi.org/10.2307/j.ctt17kk982>
- Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765), 361–365. <https://doi.org/10.1038/s41586-019-1346-5>
- Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437). <https://doi.org/10.1126/science.aav7893>
- Strong, S. P., De Ruyter Van Steveninck, R. R., Bialek, W., & Koberle, R. (1998). On the Application of Information Theory to Neural Spike Trains. *Pacific Symposium on Biocomputing*, 621–632.
- Sun, Y. J., Sebastian Espinosa, J., Hoseini, M. S., & Stryker, M. P. (2019). Experience-dependent structural plasticity at pre- and postsynaptic sites of layer 2/3 cells in developing visual cortex. *Proceedings of the National Academy of Sciences*, 116(43). <https://doi.org/10.1073/pnas.1914661116>
- T Goris, R. L., Anthony Movshon, J., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, 17(6), 858–865. <https://doi.org/10.1038/nn.3711>
- Takahashi, N., Ebner, C., Sigl-Glöckner, J., Moberg, S., Nierwetberg, S., & Larkum, M. E. (2020). Active dendritic currents gate descending cortical outputs in perception. *Nature Neuroscience*, 23, 1277–1285. <https://doi.org/10.1038/s41593-020-0677-8>
- Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S. A., & Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Neural Information Processing Systems*, 33, 1–11.
- Theunissen, F. E., David, S. V, Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*,

- 12(01), 289–316. <https://doi.org/10.1080/net.12.3.289.316>
- Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *ArXiv*, 1–16. Retrieved from <http://arxiv.org/abs/physics/0004057>
- Tomasev, N., Bica, I., McWilliams, B., Buesing, L., Pascanu, R., Blundell, C., & Mitrovic, J. (2022). Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? *ArXiv*. Retrieved from <http://arxiv.org/abs/2201.05119>
- Truccolo, W. (2004). A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects. *Journal of Neurophysiology*, 93(2), 1074–1089. <https://doi.org/10.1152/jn.00697.2004>
- Turrigiano, G. G. (2008). The Self-Tuning Neuron: Synaptic Scaling of Excitatory Synapses. *Cell*, 135(3), 422–435. <https://doi.org/10.1016/j.cell.2008.10.008>
- Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., & Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391, 892–896.
- Ungerleider, & Mishkin. (1982). Two Cortical Visual Systems. *Analysis of Visual Behavior*.
- van Hateren, J. H. (1992). A theory of maximizing sensory information. *Biological Cybernetics*, 68, 23–29. <https://doi.org/10.1007/s00422-003-0455-1>
- Van Hateren, J. H. (1993). Spatiotemporal contrast sensitivity of early vision. *Vision Research*, 33(2), 257–267. [https://doi.org/https://doi.org/10.1016/0042-6989\(93\)90163-Q](https://doi.org/https://doi.org/10.1016/0042-6989(93)90163-Q)
- Van Vreeswijk, C. (2001). Whence sparseness? *Advances in Neural Information Processing Systems*.
- Wang, L., & Maffei, A. (2014). Inhibitory Plasticity Dictates the Sign of Plasticity at Excitatory Synapses. *Journal of Neuroscience*, 34(4). <https://doi.org/10.1523/JNEUROSCI.4711-13.2014>
- Wang, Q., & Burkhalter, A. (2007). Area Map of Mouse Visual Cortex. *Journal of Comparative Neurology*, 502, 339–357. <https://doi.org/10.1002/cne.21286>
- Wang, Q., Gao, E., & Burkhalter, A. (2011). Gateways of ventral and dorsal streams in mouse visual cortex. *Journal of Neuroscience*, 31(5), 1905–1918. <https://doi.org/10.1523/JNEUROSCI.3488-10.2011>
- Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New

- York, NY: Springer New York. <https://doi.org/10.1007/978-0-387-21736-9>
- Weber, A. I., Krishnamurthy, K., & Fairhall, A. L. (2019). Coding Principles in Adaptation. *Annual Review of Vision Science*, 5. <https://doi.org/10.1146/annurev-vision-091718>
- Webster, M. A. (2011). Adaptation and visual coding. *Journal of Vision*, 11(5), 3–3. <https://doi.org/10.1167/11.5.3>
- Weliky, M., Fiser, J., Hunt, R. H., & Wagner, D. N. (2003). Coding of Natural Scenes in Primary Visual Cortex. *Neuron*, 37(4), 703–718. [https://doi.org/10.1016/S0896-6273\(03\)00022-9](https://doi.org/10.1016/S0896-6273(03)00022-9)
- Wiesel, T N. (1968). Receptive Fields and Functional Architecture of Monkey Striate Cortex. *Journal of Physiology*, 195, 215–243.
- Wiesel, Torsten N, & Hubel, D. H. (1963a). Effects of Visual Deprivation on Morphology and Physiology of Cells in the Cat's Lateral Geniculate Body. *Journal of Physiology*, 26, 978–993.
- Wiesel, Torsten N, & Hubel, D. H. (1963b). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of Physiology*, 26, 1003–1017.
- Woloszyn, L., & Sheinberg, D. L. (2012). Effects of Long-Term Visual Experience on Responses of Distinct Classes of Single Units in Inferior Temporal Cortex. *Neuron*, 74(1), 193–205. <https://doi.org/10.1016/j.neuron.2012.01.032>
- Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269(5232), 1880–1882. <https://doi.org/10.1126/science.7569931>
- Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., ... Zuo, Y. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature*, 462, 915–919. <https://doi.org/10.1038/nature08389>
- Yamins, D. L. K., & Dicarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3). <https://doi.org/10.1038/nn.4244>
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & Dicarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624. <https://doi.org/10.1073/pnas.1403112111>
- Yger, P., & Gilson, M. (2015). Models of Metaplasticity: A Review of Concepts.

- Frontiers in Computational Neuroscience*, 9(138), 1–14.  
<https://doi.org/10.3389/fncom.2015.00138>
- Yilmaz, M., & Meister, M. (2013). Rapid innate defensive responses of mice to looming visual stimuli. *Current Biology*, 23(20), 2011–2015.  
<https://doi.org/10.1016/j.cub.2013.08.015>
- Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *ArXiv*. Retrieved from <http://arxiv.org/abs/2103.03230>
- Zhang, C., Kolodkin, A. L., Wong, R. O., & James, R. E. (2017). Establishing Wiring Specificity in Visual System Circuits: From the Retina to the Brain. *Annual Review of Neuroscience*, 40, 395–424. <https://doi.org/10.1146/annurev-neuro-072116>
- Zhang, W., Basile, A. S., Gomeza, J., Volpicelli, L. A., Levey, A. I., & Wess, J. (2002). Characterization of central inhibitory muscarinic autoreceptors by the use of muscarinic acetylcholine receptor knock-out mice. *Journal of Neuroscience*, 22(5), 1709–1717. <https://doi.org/10.1523/jneurosci.22-05-01709.2002>
- Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), e2014196118. <https://doi.org/https://doi.org/10.1073/pnas.2014196118>
- Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., & Brendel, W. (2021). Contrastive Learning Inverts the Data Generating Process. *Proceedings of the 38th International Conference on Machine Learning*. Retrieved from <http://arxiv.org/abs/2102.08850>
- Zmarz, P., & Keller, G. B. (2016). Mismatch Receptive Fields in Mouse Visual Cortex. *Neuron*, 92(4), 766–772. <https://doi.org/10.1016/j.neuron.2016.09.057>
- Zylberberg, J., Murphy, J. T., & Deweese, M. R. (2011). A Sparse Coding Model with Synaptically Local Plasticity and Spiking Neurons Can Account for the Diverse Shapes of V1 Simple Cell Receptive Fields. *PLoS Computational Biology*, 7(10), e1002250. <https://doi.org/10.1371/journal.pcbi.1002250>

**CURRICULUM VITAE**

