2016

# Understanding The Implications Of Neural Population Activity On Behavior

John Briguglio
*University of Pennsylvania*, johnbri@sas.upenn.edu

# Understanding The Implications Of Neural Population Activity On Behavior

**Abstract**
Learning how neural activity in the brain leads to the behavior we exhibit is one of the fundamental questions in Neuroscience. In this dissertation, several lines of work are presented to that use principles of neural coding to understand behavior. In one line of work, we formulate the efficient coding hypothesis in a non-traditional manner in order to test human perceptual sensitivity to complex visual textures. We find a striking agreement between how variable a particular texture signal is and how sensitive humans are to its presence. This reveals that the efficient coding hypothesis is still a guiding principle for neural organization beyond the sensory periphery, and that the nature of cortical constraints differs from the peripheral counterpart. In another line of work, we relate frequency discrimination acuity to neural responses from auditory cortex in mice. It has been previously observed that optogenetic manipulation of auditory cortex, in addition to changing neural responses, evokes changes in behavioral frequency discrimination. We are able to account for changes in frequency discrimination acuity on an individual basis by examining the Fisher information from the neural population with and without optogenetic manipulation. In the third line of work, we address the question of what a neural population should encode given that its inputs are responses from another group of neurons. Drawing inspiration from techniques in machine learning, we train Deep Belief Networks on fake retinal data and show the emergence of Garbor-like filters, reminiscent of responses in primary visual cortex. In the last line of work, we model the state of a cortical excitatory-inhibitory network during complex adaptive stimuli. Using a rate model with Wilson-Cowan dynamics, we demonstrate that simple non-linearities in the signal transferred from inhibitory to excitatory neurons can account for real neural recordings taken from auditory cortex. This work establishes and tests a variety of hypotheses that will be useful in helping to understand the relationship between neural activity and behavior as recorded neural populations continue to grow.

**Degree Type**
Dissertation

**Degree Name**
Doctor of Philosophy (PhD)

**Graduate Group**
Physics & Astronomy

**First Advisor**
Vijay Balasubramanian

**Second Advisor**
Maria N. Geffen

**Keywords**
Decoding, Neural coding, Neural populations

UNDERSTANDING THE IMPLICATIONS OF NEURAL POPULATION ACTIVITY ON BEHAVIOR

John Briguglio

A DISSERTATION

in

Physics and Astronomy

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____

Vijay Balasubramanian

Professor of Physics and Astronomy

Co-Supervisor of Dissertation

_____

Maria Geffen

Assistant Professor of Otorhinolaryngology

Graduate Group Chairperson

_____

Ravi Sheth, Professor of Physics and Astronomy

Dissertation Committee

Vijay Balasubramanian, Professor of Physics and Astronomy

Maria Geffen, Assistant Professor of Otorhinolaryngology: Head and Neck Surgery

Mark Goulian, Professor of Biology

Eleni Katifori, Assistant Professor of Physics and Astronomy

Phil Nelson, Professor of Physics and Astronomy

*To Grammy, and what we can accomplish when we stop crying for a nice safe cage*

*To Grandad and Grandma Evelyn, because no one can ever take your education from you*

*To Mom and Dad, for all of the opportunities you've worked so hard to provide for me*

*To Ann, my role model, caretaker, biggest fan, and whatever I happened to need at the time*

## ACKNOWLEDGMENTS

I would like to thank Vijay Balasubramanian for taking me on a student, starting my path in neuroscience. Our time has taught me a great deal scientifically and interpersonally, but most importantly, about keeping my mind open to my own ideas.

I would like to thank Maria Geffen for affording me the opportunity to work with her and further for all of the advising she has provided. Our time has taught me my weaknesses, from which I hope to grow, while reminding me of my strengths.

I would like to thankn Phil Nelson for his unending patience and will to help with everything from details of data analysis to helping me navigate the academic system.

I would also like to thank my committee members, Mark Goulian and Eleni Katifori, for taking the time to read and listen to my work and provide valuable feedback.

Every project I have worked on has only been possible because I was able to surround myself with experts. The fascinating research in visual textures would not have been possible without Mary Conte, Ann Hermundstad, Gasper Tkacik, and Jonathan Victor. My involvement in studying the auditory cortex would have been impossible if not for the extensive amount of time, effort, and patience Mark Aizenberg and Ryan Natan showed me. Everything I have learned about applying machine learning techniques to neural activity has been through extensive discussions with David Schwabb.

I would additionally like to thank every member of the Balasubramanian Lab and the Geffen Lab with whom I have had many stimulating discussions.

My work would not have been possible without funding from the NEI Vision Training grant and the NSF.

# ABSTRACT

## UNDERSTANDING THE IMPLICATIONS OF NEURAL POPULATION ACTIVITY ON BEHAVIOR

John Briguglio

Vijay Balasubramanian, Maria Geffen

Learning how neural activity in the brain leads to the behavior we exhibit is one of the fundamental questions in Neuroscience. In this dissertation, several lines of work are presented to that use principles of neural coding to understand behavior. In one line of work, we formulate the efficient coding hypothesis in a non-traditional manner in order to test human perceptual sensitivity to complex visual textures. We find a striking agreement between how variable a particular texture signal is and how sensitive humans are to its presence. This reveals that the efficient coding hypothesis is still a guiding principle for neural organization beyond the sensory periphery, and that the nature of cortical constraints differs from the peripheral counterpart. In another line of work, we relate frequency discrimination acuity to neural responses from auditory cortex in mice. It has been previously observed that optogenetic manipulation of auditory cortex, in addition to changing neural responses, evokes changes in behavioral frequency discrimination. We are able to account for changes in frequency discrimination acuity on an individual basis by examining the Fisher information from the neural population with and without optogenetic manipulation. In the third line of work, we address the question of what a neural population should encode given that its inputs are responses from another group of neurons. Drawing inspiration from techniques in machine learning, we train Deep Belief Networks on fake retinal data and show the emergence of Garbor-like filters, reminiscent of responses in primary visual cortex. In the last line of work, we

model the state of a cortical excitatory-inhibitory network during complex adaptive stimuli. Using a rate model with Wilson-Cowan dynamics, we demonstrate that simple non-linearities in the signal transferred from inhibitory to excitatory neurons can account for real neural recordings taken from auditory cortex. This work establishes and tests a variety of hypotheses that will be useful in helping to understand the relationship between neural activity and behavior as recorded neural populations continue to grow.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

# 1. Introduction

## *The central problem in neuroscience*

Neuroscience concerns itself with understanding the brain, the organ most responsible for making us both, human and individuals. We are able to solve incredibly complex computational problems with little to no effort, including fixing our gaze on a particular object while moving our entire bodies and identifying objects in a complicated environment. There is something fundamentally interesting about trying to understand *how we work*. What does it mean to understand how the brain works? If the brain is a puzzle, we want to know the picture. The pieces, the things we have access to experimentally, are the small windows we have to view the picture. Developing an understanding of what the brain is doing may require only understanding what the larger picture is, and convincing ourselves that the pieces fit together to form such a picture. The importance of theory to neuroscience lies in its ability to draw specific pictures describing generically what the pieces may come together to form, regardless of the details of their individual shapes. That is, to turn *knowing* how the brain works into *understanding* how the brain works.

One recurring challenge encountered when trying to understand the brain relates to the general importance of abstraction. In early sensory systems, progress in understanding the neural code has been aided by the fact that we have some good sense about the type of representation we would expect to observe. For example, the retina has photoreceptors tiling the back of the eye (conceptually similar to the CCD mosaic in a camera), which leads to a natural guess that the representation used by early visual neurons may relate to the spatial patterning of the light. A model of the early visual

system where the light inputs are parameterized by their spatial distribution provides some of the canonical results in understanding the contributions from individual neurons. In the retina, for example, this model reveals that many neurons have a center-surround structure, while in primary visual cortex (V1), Gabor filters emerge.

Unfortunately, natural parameterizations aren't always so obvious for many of the problems the brain has to solve. For example, the encoding of *value* is inherently more difficult to quantify [1], but is essential in order for any organism to make wise decisions. More generally, the neural architecture evolution has stumbled upon to solve a particular problem may have no readily observed mapping into the kinds of algorithms we are accustomed to thinking about, despite using one. To illustrate this point, consider the problem of tracking your own hand position. One simple solution would be to encode a vector containing the angles of your shoulder, elbow, and wrist (as opposed to keeping track of the absolute spatial location). Any rotation of this vector would contain the same information as the original, but would obscure interpretations about the underlying representation. This makes the two representations difficult to distinguish by observing the neural responses, not because of any fundamental difference in the algorithms (in fact, there may be computational *advantages* of this kind of manipulation as it can information more diffusely available), but because recognizing the algorithm relies on our own ability to internally visualize it in a simple way. In light of this, keeping an open mind about the kinds of computations that may be going on is very important, since computational strategies that seem superficially dissimilar to biological ones may simply be embedded in a non-trivial way.

This dissertation presents several lines of work that use theoretical ideas about neural organization to predict behavior while avoiding issues with precise characterization of neural activity. By doing so, we are able to shed light on a number of issues of broad importance in computational neuroscience.

In chapter 2, we extend ideas of the efficient coding hypothesis to explain human perceptual sensitivity to visual textures. By simply examining statistics of natural scenes, we are able to predict the relative sensitivity humans display to a variety of visual textures. We avoid complex issues of representation that arise from dense correlated visual features by predicting directly the effects on behavior. In doing so, we show that the efficient coding hypothesis is a guiding principle for cortical organization, and we shed light on the differences in constraints between central and peripheral sensory processing. The work presented in the first part of this chapter is published in [2].

In chapter 3, we quantify the role auditory cortex plays in frequency discrimination acuity in mice. Optogenetic manipulations of the auditory cortex directly change its neural activity, but also change the frequency discrimination acuity of the animal. By examining the information-theoretic limitations on discrimination performance, we make individual frequency-discrimination predictions for each mouse, regardless of the manipulation performed. By doing so, we find not only that behavioral changes correlate with neural limitations, but that individual variability to a fixed manipulation is explained by neural activity. This reinforces the importance of treating subjects as individuals, as differences between behavior of mice is accounted for by differences in their neural activity. At the time of writing, the paper containing this work is in preparation.

In chapter 4, we take steps towards addressing the question of how neural circuits in cortex should organize given the fact that the inputs are *not* the external world, but rather the world as filtered by the senses. We examine the response properties of elements of Deep Belief Networks trained on the output of a fake retina to find, among other things, Gabor-like receptive fields that are common in cortex. This reaffirms that these filters are one way of efficiently representing natural stimuli, and provides an alternative learning rule that can produce these types of filters. Additionally, this work establishes that retinal responses are conducive to producing this kind of representation.

In chapter 5, we model excitatory-inhibitory network dynamics in auditory cortex and demonstrate that a single non-linearity in the inhibitory-to-excitatory synapse can account for a number of observed adaptive phenomena and optogenetic manipulations. This model establishes the simplest model that can account for the observed pyramidal neuron activity, and makes predictions about properties of the inhibitory neural population. The work presented in this section is published [3] [4].

In this following portion of this chapter, we will discuss relevant background information that provides context for the several of the subsequent chapters, including a discussion of the efficient coding hypothesis and a basic overview of neuronal function and the leverage optogenetic techniques provide to manipulate their activity.

## *The efficient coding hypothesis*

One concrete theory that has proven to be a helpful way to think about neural coding is the *efficient coding hypothesis*, first postulated by Barlow in 1961 [5]. The hypothesis states that evolution favors organisms more capable of sensing their environment. Put more precisely, the cost of neural resources to an organism will invoke selective pressure that favors individuals who maximize the mutual information their sensory organ provides about the environment. In some cases, this means that neurons have to remove redundancies in their input. In other cases, it means that noisy signals need to be combined in a manner that improves odds of detection. In all cases, the idea requires a "natural signal", and efficiency cannot be defined without it. In fact, the existence of a stable "natural signal" has to exist on evolutionary timescales in order for the organism to adapt to it, and so a number of timescales are at play. For example, if one computes the Fourier power spectrum of urban "natural" images and ones in nature, one will notice an overabundance of horizontal and vertical edges [6], likely resulting from e.g. buildings. Should we be more sensitive to these features by virtue of existing in modern society? While this may not have been present on evolutionary timescales, it is possible that evolution favored some degree of flexibility, and we have mechanisms in place that adapt to a number of different features in the world. It may be possible that we have adaptive processes that are capable of making us more sensitive to these features in relation to their increased presence, but strictly speaking, the hypothesis has little to say about this.

The efficient coding hypothesis has a long history of providing useful insight to early vision. The retina is a part of the brain whose output cells (retinal ganglion cells)

primarily lie on a single surface, making their responses relatively easy to access using multi-electrode arrays. Additionally, the optic nerve imposes a bottleneck on how much information the retina can pass to cortex, and therefore devote to any particular feature of the visual environment. Among mammals, the primate retina is unusual in that it is trichromatic, suggesting that the additional visual information was beneficial for us, and our day-to-day experiences tend to be visually dominated. The combination of ease of experimental access, evidence for selective pressure, and ease of controlling and measuring the input stimulus have made the retina a prime target for testing the efficient-coding hypothesis. In a 1990 paper [7], Atick and Redlich analytically optimize a coding scheme to minimize channel capacity requirements while maintaining a fixed information rate for a variety of luminance ratios for encoding of natural images. In doing so, they found numerical solutions for filters that were remarkably similar to retinal ganglion cell response profiles—including center-on/surround-off type responses when the signals are reliable [i.e. high contrast], and pooling over a large area when signal are unreliable. In concluding remarks, they remark that calculating a global optimum with respect to efficient representation is challenging, or even impossible, and therefore from a neural coding perspective, it makes sense to compute such an optimum only for a restricted family of filters, allowing each successive stage to improve in representation compared to the previous. Since then, a variety of additional ideas regarding principles of neural coding have been tested in the retina [8] [9].

The ideas of efficient representation have also been extended to try to explain cortical responses. As another example [10], Olshausen and Field examined in 1997 the idea that sparse representations may prove useful to better represent the underlying

structure of images, which has intuitive appeal because the images are generally composed of relatively few objects with particular boundaries. They present an algorithm for learning such sparse features, and when trained on natural images, the filters these structures derive resemble Gabor filters, characteristic of neural responses in V1. The idea of efficiently representing the environment appears helpful for making sense of cortical responses as well, although as we will show in chapter 4, these types of filters can emerge from other kinds models as well. One of the important takeaways is that it very well may not be the case that V1 is trying to optimize the cost function as explicitly written in one of these efficient coding papers, but the representation observed may nonetheless be highly efficient for a variety of similar cost functions. In chapter 2, we will examine other implications efficient coding has for behavior when applied to cortical coding.

Ideas of efficient coding have also been applied to the auditory pathway. In 2002 [11], Lewicki showed that performing Independent Component Analysis (ICA) on short snippets of a variety of natural sounds results in filters that are characteristic of responses in the auditory fiber. One of the major criticisms of the efficient coding hypothesis is the argument that, to biological systems, not all information is equally important. For simple organisms, this is likely a large factor. For more complex organisms with structures in place for making high-level decisions, there is a great deal of flexibility afforded to the organism by virtue of having a sensory system providing as much information as possible, while the higher structure can decide what to throw away. It is likely, then, that these principles will remain useful for understanding peripheral processing. At some point in the pathway, decisions must be made, and behavioral relevance becomes

unequivocally important. In chapter 4, we discuss the importance of sensory limitations in this context, and the implications it has for behavior.

### *Manipulating neural activity with optogenetics*

Neurons are the fundamental units of computation within the brain. What makes neurons unlike most other cells is that their cell membranes are highly electro-chemically sensitive, containing many voltage-gated ion channels. When the voltage difference between the interior and exterior of the cell membrane crosses a certain threshold, it starts a chain reaction of ion channels opening. This causes an extremely pronounced, stereotyped voltage response from the cell itself, called a *spike*. What makes neurons useful for computation and action is that they also have an *axon*, a long, cylindrical extension of the cell membrane that shares the features of electro-chemical excitability with the body. The spiking activity in the cell body is propagated through the axon, which can travel long distances (~1 meter for the sciatic nerve, for example). The activity pattern is decidedly discrete, as generically the output of the axon is silence punctuated with a few short, obvious pulses when the neurons spikes. Although things like external voltage fluctuations near the cell body can have large effects on the observed spiking activity (therefore analog computations may be quite relevant in understanding neural responses), the output of the neuron to distant brain or motor areas is decidedly discrete. This is also the reason why, in neural coding studies, emphasis is generally placed on the spiking activity of neurons, rather than the raw voltage traces, and the output of neurons is frequently treated as a digital stream. The vast majority of neurons also have dendrites, membrane protrusions responsible for connecting with axons from other. These axon-

dendrite interfaces, called *synapses*, are responsible for allowing neurons to receive electrochemical inputs from other neurons. In specialized cells, such as photoreceptors in the retina or inner hair cells in the inner ear, the inputs come from electrical or mechanical interactions with light and sound, allowing transduction of external signals. There are a number of different kinds of influences neurons can have on one another, and most neurons stereotypically excite or inhibit the ones that they form synapses with. Within cortex, roughly 80% of neurons are excitatory, and the remainder inhibitory. Since fibers projecting from one brain region to another typically contain bundles of axons from excitatory neurons, a useful simplified view is that excitatory neurons encode the results of any computation from a brain region, while the inhibitory neurons are necessary for the computation to take place. The inhibitory neurons in cortex can be divided into three subgroups called, PV ("parvalbumin") , SOM ("somatostatin"), and VIP ("vasoactive intestinal polypeptide") based on marker proteins they express, and represent ~40%, ~30%, ~30% of all inhibitory neurons in cortex, respectively [12]. We will primarily be concerned with the first one in chapter 3, and the first two in chapter 5.

The innovation of optogenetics revolutionized the kind of control experimentalists have over neurons. In green algae, channelrhodopsin is a protein that functions as a photosensitive ion channel used by green algae to "see", allowing it to move in the response to the presence of light. During the 2000's, a series of innovative approaches demonstrated techniques allowing neurons in other animals to express channelrhodopsin, allowing experimenters to control the activity of the neuron by shining visible light on it. Since its inception, significant improvements to temporal response, channels that allow activation or suppression of neurons, and genetic mouse (among other animals) lines

have been developed, allowing for very precise control of highly specific neural populations. In that past two sentences, I have trivialized a large body of work that is almost certainly Nobel prize-worthy. This is an incredibly rich field in its own right, and more information can be found in reviews such as [13]. One common usage of these techniques is to probe and elucidate the role specific neuronal subtypes play in cortical processing, as is the perspective we take in chapter 5 to examine the implications of the adaptive responses in auditory cortex on excitatory-inhibitory network state. In chapter 3, we take a slightly different perspective of their utility. We leverage the fact that each manipulation provides a different perturbation of the network to test a broad hypothesis about the role auditory cortex plays in frequency discrimination.

# 2. Behavioral evidence for efficient coding using visual textures

## *Principles of higher-order vision*

It has been colloquially said that the visual world is made up of "things" and "stuff", where "things" generically refer to obviously identifiable objects, and "stuff" is everything else. The point of this phrasing is that differentiating between what comprises a specific "object" and what comprises a "texture" is difficult, and not particularly well-defined. Most of our visual world is comprised of a series of objects of varying in size from large to small occluding one another. For example, a person may identify leaves on a front lawn in a close-up photograph as individual objects, but in a zoomed out picture of an entire house, leaves on grass may be better described as a visual texture. Visual textures can be thought of as patterns of localized statistics within an image that are repeated to cover a larger patch. In this example, the relevant statistical properties are contained within a length-scale approximately the size of a leaf, but are repeated to cover the size of the yard. One interesting feature about such large-scale image features is that the early cortical representation *must* be quite diffuse, as such texture can generally span a region much larger than the receptive field of early cortical neurons. We consider this complication a feature, rather than a concern, as most natural stimuli likely require activity from many neurons to encode/decode. We will see that efficient coding nevertheless makes useful predictions about the behavior that reflects the distribution of resources cortex devotes to various higher-order image features. This provides a different type of prediction from many of the previous efficient coding studies mentioned in the

introduction, one that may prove to be helpful in understanding coding of complex stimuli in other sensory modalities as well.

Previous work within our own collaboration has shown that high-order statistics that are predictable from lower-order ones are not encoded by cortex [14], which is consistent with suggestions proposed by van Hateren [15]. The intuition for the principle we will establish here is that, among natural signals that are unpredictable from lower-order ones, those with higher variability can better serve to differentiate between objects, materials, environments, etc. In order to measure this in correspondence in detail, we will first discuss the various regimes of efficient coding, and the implications they have for resource allocation in any coding population. With this established, we will examine a specific class of visual textures in order to establish that we can create and measure images containing specific, well-defined "texture" signals. With this well-defined signal in hand, we will then discuss how to characterize a natural image database using these signals. Then we will discuss the psychophysical measurements made in order to test human sensitivity to these textures. We will then compare the results of the natural image analysis to the behavioral results, keeping in mind the predictions made by the efficient coding hypothesis. After discussing the implications of this published work, we will show unpublished work with preliminary results extending these analyses to a larger class of visual textures and discuss the new questions that arise.

## *Two regimes of efficient coding*

The efficient coding hypothesis states that the neural circuitry should operate in a manner that maximizes the mutual information of the neural response about the

environment. We will examine the analytical results of a simple encoding problem to show two of the interesting coding regimes which arise.

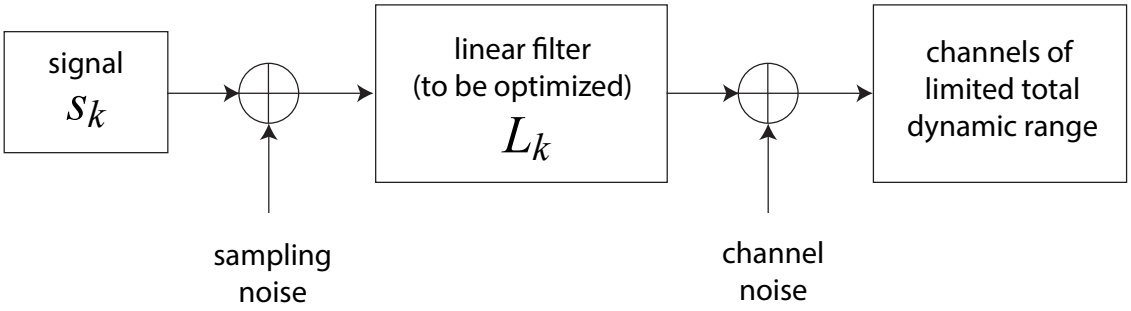


Figure 1: Schematic of optimization problem. In this problem we are constrained to encode signals $s_k$ in the presence of input and output noise with some linear filters, denoted $L_k$.

As in Figure 1, assume $s_k$ are the variance of Gaussian signals (indexed by $k$) we wish to encode using some type of linear filter, denoted $L_k$, in the presence of sampling (input) noise, channel (output) noise, with a limited bandwidth. Without loss of generality, we can take the sampling and channel noise to be unity, as we may rescale the signal size for the former and the total dynamic range size for the latter. We expect the sensitivity of the system to a particular signal to scale like the gain, $|L_k|$. We are still constrained by the total output power of the system, $P$, and so the problem can be formulated seeking to extremize the quantity $I = \sum_k I_k + \Lambda P$. Here $I$ is the quantity to be extremized with respect to $L_k$, $I_k$ is the mutual information between the channel input and output, and $\Lambda$ is the Lagrange multiplier used here to enforce the power constraint. Non-trivial solutions occur for $0 < \Lambda < 1$, and as $\Lambda$ moves from 0 to 1, the constraints switch

from being dominated by input-noise to being dominated by output-noise. This is worked out in detail in [15] by setting $\partial I/\partial L_k = 0$ and $\partial I/\partial \Lambda = 0$. The solutions are given by

$$|L_k|^2 = \frac{-(2+s_k^2) + \sqrt{s_k^4 + 4s_k^2/\Lambda}}{2(1+s_k^2)}$$

when the quantity is positive, and 0 otherwise. This quantity is positive as long as $s_k > \sqrt{\Lambda/(1-\Lambda)}$. This captures the intuition that sufficiently small signals are not worth encoding. For $0 < \Lambda < 1$, when $\Lambda$ is near 1, the critical value of $s_k$ becomes infinite, which corresponds to the transmission-limited, or output-noise limited regime. This implies nothing but the largest of signals are worth encoding at all. When $\Lambda$ is near 0, this critical value of $s_k$ approaches 0, which corresponds to the transmission limited, or input-noise limited regime. In this situation, virtually all signals are worth encoding. Numeric solutions depicting the resulting gain as a function of the signal strength are plotted in Figure 2.

In the transmission-limited regime ($\Lambda$ near 1), signals below the threshold value have zero gain, and for large signal values, the asymptotic limit of the gain equation for large signal strengths is given by

$$|L_k|^2 \sim \frac{1/\Lambda - 1}{1+s_k^2}$$

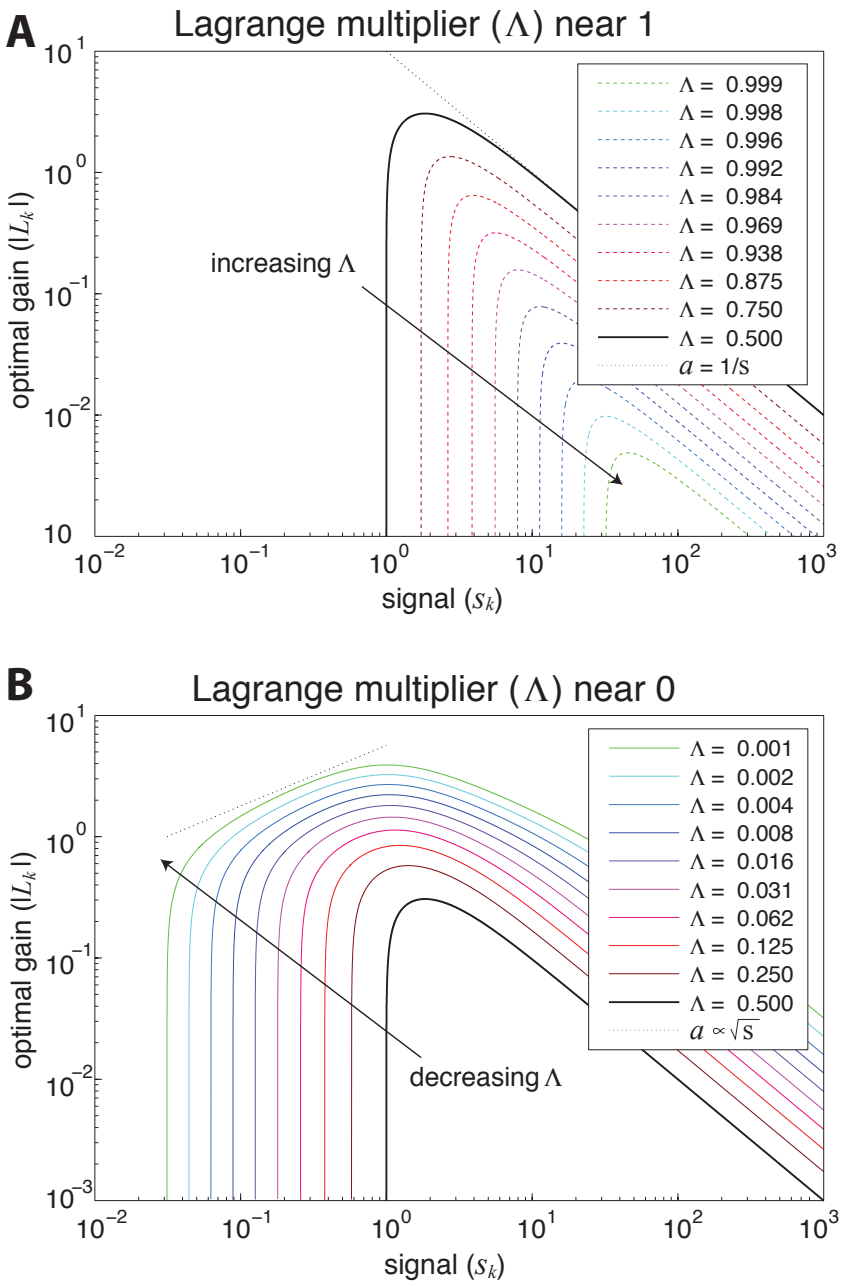


Figure 2: Numeric depiction of different efficient coding regimes. Plots show the optimal gain, $|L_k|$, as a function of signal strength for varying levels of $\Lambda$, the Lagrange multiplier which enforces the output power constraint. Whenever $s_k < \sqrt{\Lambda/(1-\Lambda)}$, the signal is not encoded. Panel A depicts $\Lambda$ near 1, the transmission limited regime, where the gain of a signal is inversely proportional to the signal strength ($|L_k| \sim 1/s_k$). Panel B depicts $\Lambda$ near 0, the sampling limited regime, where the gain of a signal is inversely proportional to the signal strength ($|L_k| \sim \sqrt{s_k}$).

This can be understood with the intuition that when signals are highly reliable, it is optimal to spend an equal amount of bandwidth encoding each one. The gain here is matched to compress the signal to fit into a fixed amount of bandwidth ($|L_k| \sim 1/s_k$). There is also a very sharp transition between the signals which are encoded according to this bandwidth-equalizing intuition, and those which are not worth encoding at all. This is depicted numerically in Figure 2**A**.

In the sampling-limited regime ($\Lambda$ near 0), signals below the threshold value still have zero gain, but there is a much larger transition region between the signals which are not encoded and the reliable signals. The asymptotic form of the gain equation under the conditions of $\Lambda$ near 0 is

$$|L_k|^2 \sim \frac{s_k}{1 + s_k^2} \sqrt{\frac{1}{\Lambda}}$$

If we examine the region $\sqrt{\Lambda} < s_k < 1$, where the signal is smaller than the sampling noise, but larger than the threshold for encoding, we see that the gain *increases* with the signal size ($L_k \sim s_k^{1/2} \Lambda^{-1/4}$). This is plotted in Figure 2**B**. This regime quantifies the intuition that, when signals are relatively unreliable, more resources should be spent on those which are more reliable. For the purposes of analyzing signals which inherently possess significant sampling limitations, this regime is likely to be more relevant. Consider, for example, visual textures. The relevant properties have significant statistical structure which needs to be averaged over some large homogeneous spatial region in order to have a measurement with small error, but spatial variations are significant in natural scenes, and the extent of homogeneity unpredictable (a priori). In order to retain

the important spatial variations, measurements of such statistics will be inherently noisy. This motivates our hypothesis that human perceptual sensitivity to a visual texture (quantified by a signal computed from images) should grow with the variability (measured from natural visual scenes) of its signal.

### *Parameterizing a tractable set of visual textures*

One of the powerful implications of the efficient coding hypothesis involves the sensitivity of the population which encodes the relevant features of the natural world. This sensitivity is something which any population, regardless of the particular encoding scheme, should achieve. It is therefore possible, as long as we have a well-controlled stimulus, to test the efficient coding hypothesis *without* knowing anything about the actual underlying representation. By simply examining behavioral sensitivity to a well-parameterized stimulus, and comparing the behavioral sensitivities to the presence of these signals in natural images we can test predictions of the efficient coding hypothesis at a macroscopic level. Our collaboration has previously tested this by looking at specific patterns and classifying them as either informative (belonging to the coding region) or uninformative (belonging to the zero gain region) based on whether or not they are informative about natural scenes [14]. Our goal here is to probe these predictions in greater depth by comparing the sensitivities to multiple patterns which are all predicted to be encoded by the sensory system.

Generically, visual textures are motifs with a particular small-scale structure that is repeated over a large region of the visual environment. The number of parameters one must keep track of for arbitrary visual textures grows exponentially with both the size and

the possible colorings of the regions. It is therefore most prudent to start with the simplest tractable subset of these textures which can capture important two-dimensional spatial structure. We therefore constrain ourselves to considering 2x2 pixel motifs containing only black and white pixels. There are $2^{2\times 2} = 16$ possible configurations such a grid can take, and a visual texture of this class can be described by the probabilities of each coloring. Probability summing to 1 and translation invariance reduce the number of free parameters to 10. A convenient basis to describe these is given by the general discrete Fourier transform and contains one first-order coordinate ($\gamma$), four second-order coordinates ($\beta_{-}, \beta_{|}, \beta_{/}, \beta_{\backslash}$), four third-order coordinates ( $\theta_{\llcorner}$ and rotations), and one fourth-order coordinate ($\alpha$). For more details about this showing this is a complete representation, see [16]. For any patch, computing these quantities is straightforward. Each of these coordinates has a specific configuration of pixels, and the value it takes for one example configuration is given by the parity of the pixels contained, taking black to be -1 and white to be +1. The coordinate value describing an image patch is the average across every matching configuration contained in the image patch. So $\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$ has $\beta_{-} = -1$, as there are two horizontal pixel configurations with the values $\beta_{-}(-1 \quad 1) = -1$ and $\beta_{-}(1 \quad -1) = -1$.

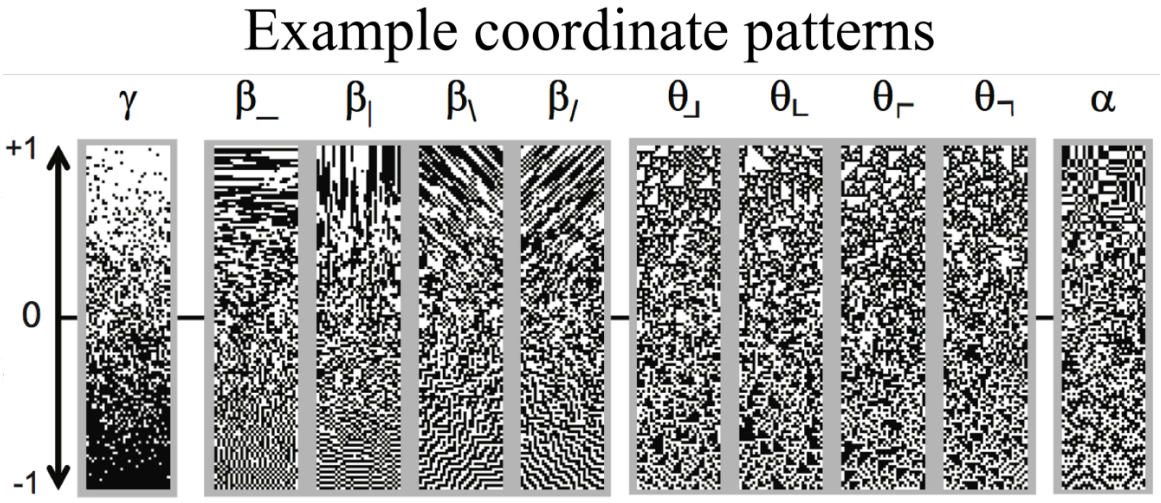


Figure 3: Visualizing 2x2 binary textures with single coordinates specified. Midline is white noise, and moving up or down in each column corresponds to increasing or decreasing the average value of the indicated coordinate. The emergent structures are easily visible at the extreme ends of the spectrum.

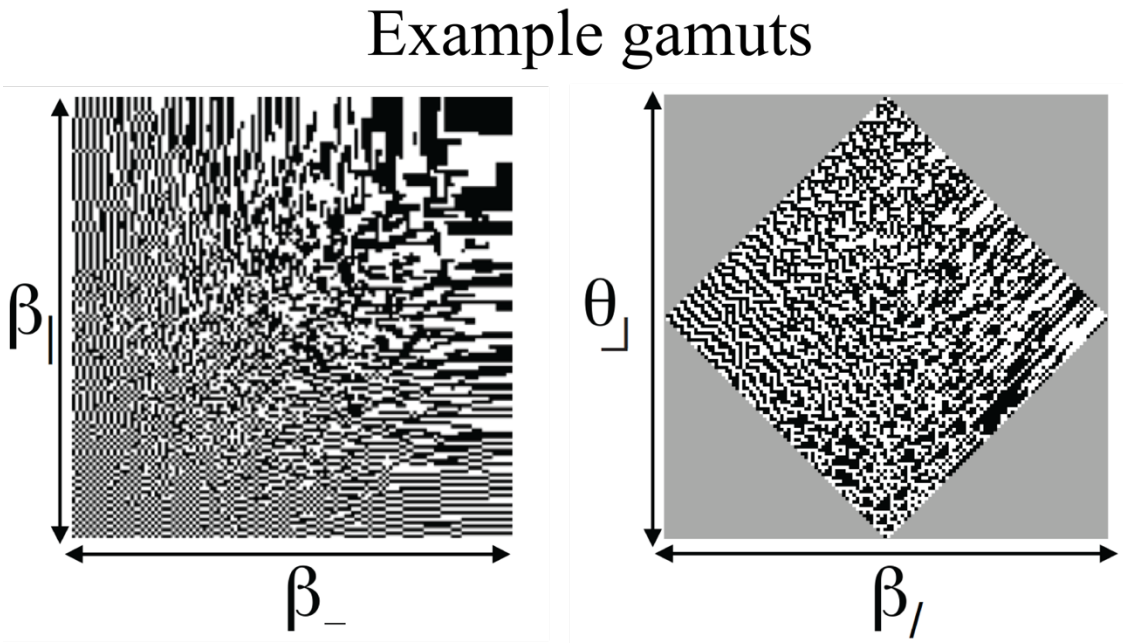


Figure 4: Visualizing 2x2 binary textures with multiple coordinates specified. Center point is white noise, and moving outward the patterns generated use increasingly strong coordinates. The emergent structures are easily visible at the extreme ends of the spectrum, and the combination of two coordinates provides significantly different patterns from only specifying one. There exist restricted regions (e.g. gray region of right panel) where no realizable pattern can give such statistic combinations.

It is possible to generate image samples which are maximum entropy subject to the constraint of having 1 or 2 coordinates specified [16] , and examples of the appearance of these patterns appear in Figures 3 and 4. To provide some intuition for this algorithm, consider the simple case of specifying single coordinates. It is easy to identify a boundary which contains only uncoupled pixels, which may be generated randomly. From here, the relevant template shape may be shifted in such a way that only one pixel is undefined. The pixel color is chosen from the Boltzmann distribution, enforcing the constraint on the average coordinate value for the image. For example, specifying the $\beta_{-}$ coordinate leaves independent rows, and so in each row, we may randomly generate the left-most pixel. We may sequentially generate pixel $i+1$ according to the distribution $p(c_{i+1}) = \frac{1}{Z} e^{-\text{atanh}\,(\beta_{-})c_i c_{i+1}}$. The functional form of this equation is identical to a formulation of the one-dimensional Ising model that specifies the spin-spin correlation rather than the coupling strength.

### *Characterizing the "natural" visual environment using visual textures*

The efficient coding hypothesis claims that sensory systems of organisms have evolved in order to be able to efficiently represent the types of stimuli they naturally encounter. In order to remain faithful to this claim, we use images from the UPenn Natural Image Database. The images are taken from natural baboon habitats in Botswana using a camera calibrated to faithfully capture the responses that L, M, and S cones of primates [17], although we cross-checked our work with another popular image database(the van Hateren Image Database). Since we will be eventually making a comparison to binary textures, we will consider the overall luminance at each point in the

image as the most relevant element, though certainly more generic visual textures of interest contain more generic color patterns. But what is the most sensible way to retain the structure of a grayscale image after converting to a binary image? Natural images have well-documented long-range correlations, which can be understood to a large extent by the properties of *translation-invariance* and *scale-invariance* [18]. The former can be understood by virtue of the fact that shifting a natural scene, for example to the left or to the right, yields another natural scene. The intuition explaining the notion of scale-invariance in natural images is as follows: if a particular environment or set of objects constitutes one natural image, then so does the same set of objects as viewed from either half the distance or twice the distance. These seemingly simple observations have powerful implications about the statistical properties of natural images, including the typical pair correlation between pixels. The fact that natural images have long-range pixel-pixel correlations implies that simply binarizing the grayscale image by itself (e.g. about the pixel intensity median) leaves large regions of the image either entirely black or entirely white and removes much of the small-scale structure of the image. This is a property which holds across the ensemble of images, and is itself unhelpful in distinguishing individual images from one another. By only removing the average pair correlation across the entire database (a procedure called whitening), we leave excess correlations that exist in specific images, and therefore don't lose any information that can be used to distinguish images. The difference in these two methods is illustrated in Figure 5. The whitening filter has a center-surround structure reminiscent of some retinal ganglion cells, and we have discussed arguments that the purpose of some early visual processing is to decorrelate the visual input in a similar way.

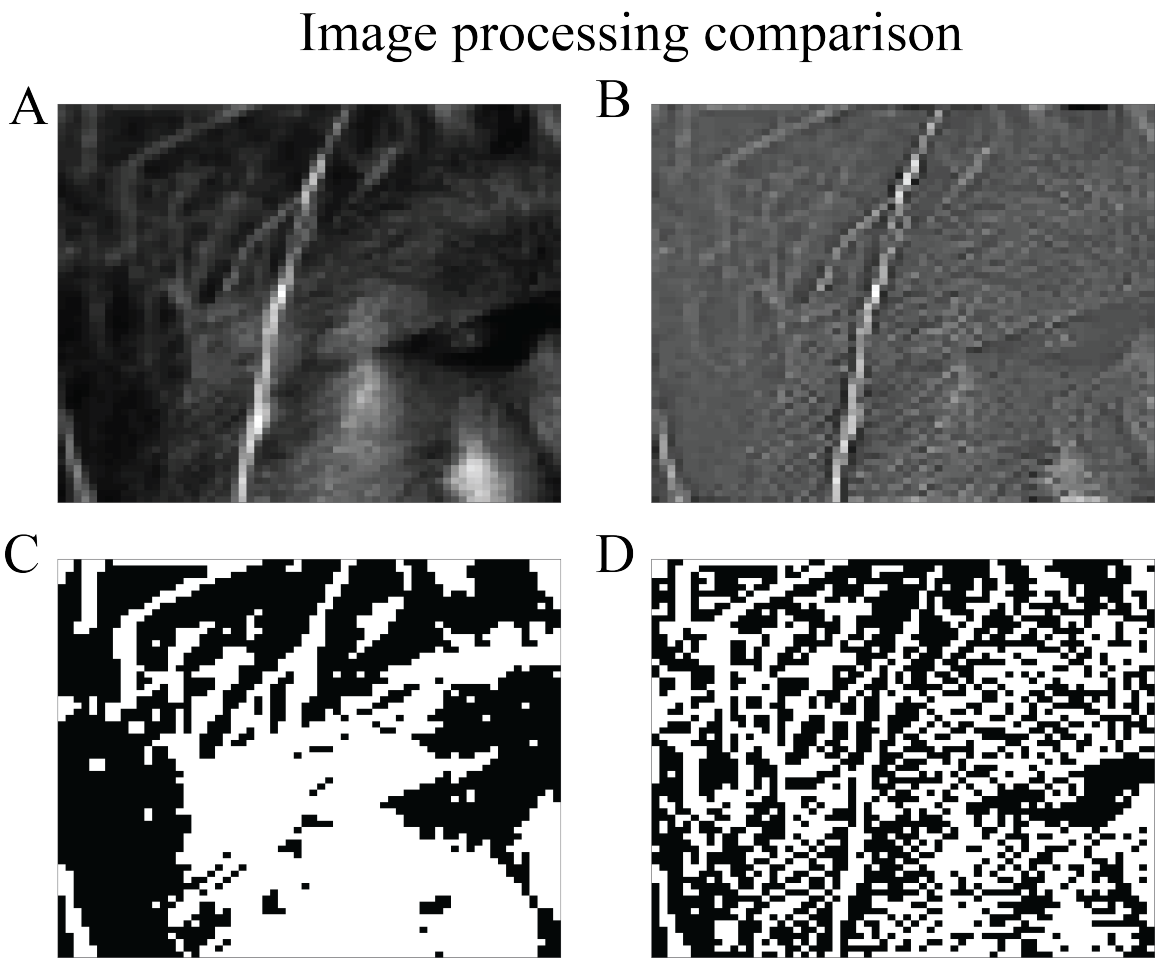


Figure 5: Comparing binarized images with and without removing average pair correlation. In panel **A**, the original image of a baboon's face slightly obscured by some brush. In panel **B**, the image has been filtered in order to remove the average pair correlation from the dataset. The significant features of the image are still visible. In panel **C**, the original image (panel **A**) has been binarized by setting all pixels with luminance higher than the median to 1, and all other to zero. Information about many local features, such as fur texture, are completely absent due to the strength of long-range correlations. In panel **D**, the whitened image (panel **B**) has been binarized about its median pixel intensity. Much more local information, such as the grass's countour and the hair texture, remains visible.

In order to check that scale does not affect the results, we introduce the block-average factor $N$ which sets the scale of the image by shrinking the image by a factor of $N$ in each direction, whose pixel values are the average of the corresponding $N \times N$ block in the original image. We do not assume scale invariance holds in the natural images, so we will remove the average pair correlation computed empirically from the natural image dataset used. This is done by flattening the average Fourier power spectrum, which relies on translation invariance.[1] To reliably estimate the pair-correlation for an image with $P$ pixels, we need approximately $P^2$ images (or $P$ images if we assume translation-invariance). Since our nice-sized databases have ~1000 images with ~1 Megapixels each, it is obvious that we will not have enough data to compute these quantities for full-sized images. Instead, we cut the original images into image patches of size $R \times R$ to form a larger database of smaller images. With these choices in image processing parameters, we can additionally test the results to see whether or not the scale of the image analyses has any bearing on the texture representations. The full processing procedure is pictured in Figure 6.

[1] Another way to achieve this result is by computing the principle components of the dataset and rescaling the image, as represented in the principle-component basis, by the inverse square-root of its variance. This also leaves pixels uncorrelated on average, and does *not* rely on the translation-invariance assumption, but is numerically unstable. Inevitably for large vector spaces like this, there will be principle components with variances near zero. These principle components that explain almost nothing about the data will be amplified by a numerically unstable amount using this method.

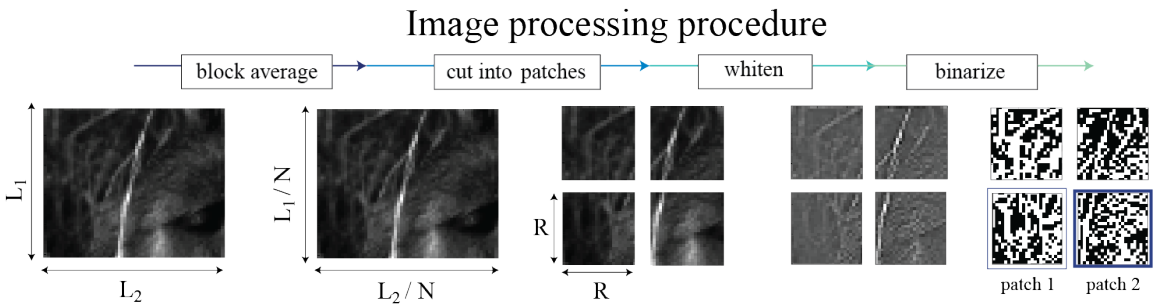


Figure 6: Depiction of image processing procedure. We first take an ensemble of images and make new pixels by averaging blocks of $N \times N$ pixels to make effective pixels in order to test the analysis across scales. We then divide the new image into patches of size $R \times R$ in order to be able to have enough samples to make meaningful ensemble statistics. This provides another check on scale invariance for the estimation of statistics. The image patches are then whitened in order to remove the mean pair-correlation. The whitened image is binarized at its pixel-intensity median, yielding a binary image which contains much of the structure at all length scales. The binarized image patches are used to compute the distribution of texture parameter values across natural images.

Once we have these image patches, we may compute the distributions of the various texture parameter values (in the manner described above) in order to see which are the most informative ones about natural scenes, and therefore, the ones to which we expect people to be most sensitive. We compute the mean of each of the texture parameters in each image patch, and our distribution contains one such vector for each image patch in the analysis. The standard deviation of this distribution, which we consider here to represent the strength of the signal from the above efficient coding calculation, is plotted for single coordinates in Figure 7. A single scale factor for the overall vector length was used for each set of image processing parameters. This can account for overall variance differences that can arise due to larger image patches having inherently smaller variances. Interestingly, non-trivial structure has already begun to emerge. We can see that the horizontal and vertical two-point correlations are the

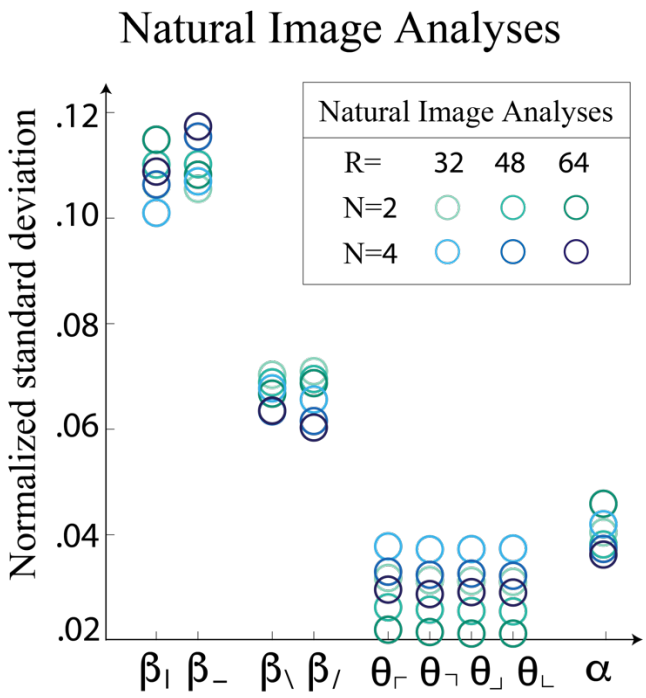


Figure 7: Normalized standard deviation of single coordinates. Here, we see the horizontal and vertical two-point correlations are most prominent, followed by diagonal two-point correlations. Four-point correlations are more prominent than three-point correlations of any orientation. Despite the apparent overlap in the cloud of points, rank ordering is preserved for each individual analysis.

strongest, followed by diagonal two-point correlations. Three-point correlations are the least prominent, with smaller variance than the four-point correlation. Performing this analysis on white noise yields equal standard deviation in each coordinate direction, suggesting that these are indeed novel features characterizing natural images.

### *Characterizing human sensitivity to visual textures*

To draw an analogy to the efficient coding hypothesis above, we interpret the ideal amount of gain to apply to a signal to be proportional to the sensitivity a subject displays to the signal. This means that we do not need to measure from the entire neural population to make a guess about the amount of neural resources devoted to the texture

signals, but rather we know the effective gain applied by measuring the psychophysical sensitivity. Additionally, it is worth noting that neural representations supporting discrimination of this kind of visual texture do not emerge until, at the earliest, secondary visual cortex (V2) [19]. In order to test human sensitivity to these visual textures, we use a four-alternative forced-choice task (see Figure 8**A**), in which a strip with a specific set of parameter values is placed in one of four locations (top, bottom, left, or right) and the rest of the image is filled with white noise. More specifically, the subject is asked to fixate at a point on a screen, after which the image changes to the structured target on white noise background for 120ms, before a white noise washout image is displayed to prevent the user from utilizing the afterimage. The task reflects the ability of the subject to distinguish the texture from white noise. This is done for a variety of coordinate values (specifying single and dual coordinate values), from which a *threshold* is defined as the strength of a parameter required for a subject to distinguish the location of the texture with an accuracy of 62.5% (halfway between chance and perfect) as schematized in Figure 8, panel **B**. An early observation about the psychophysical sensitivities shows that human subjects are symmetrically sensitive to positive correlations as negative correlations. There is no reason that this needs to be the case, although it is a property that an ideal observer would exhibit. The results from single-coordinate measurements feature the same rank-ordering as in the natural image analyses, $\beta_{-}, \beta_{|} > \beta_{/}, \beta_{\backslash} > \alpha > \theta$. Here, due to the indistinguishability of thresholds for some classes of texture parameters, single values were reported to represent sensitivity to that

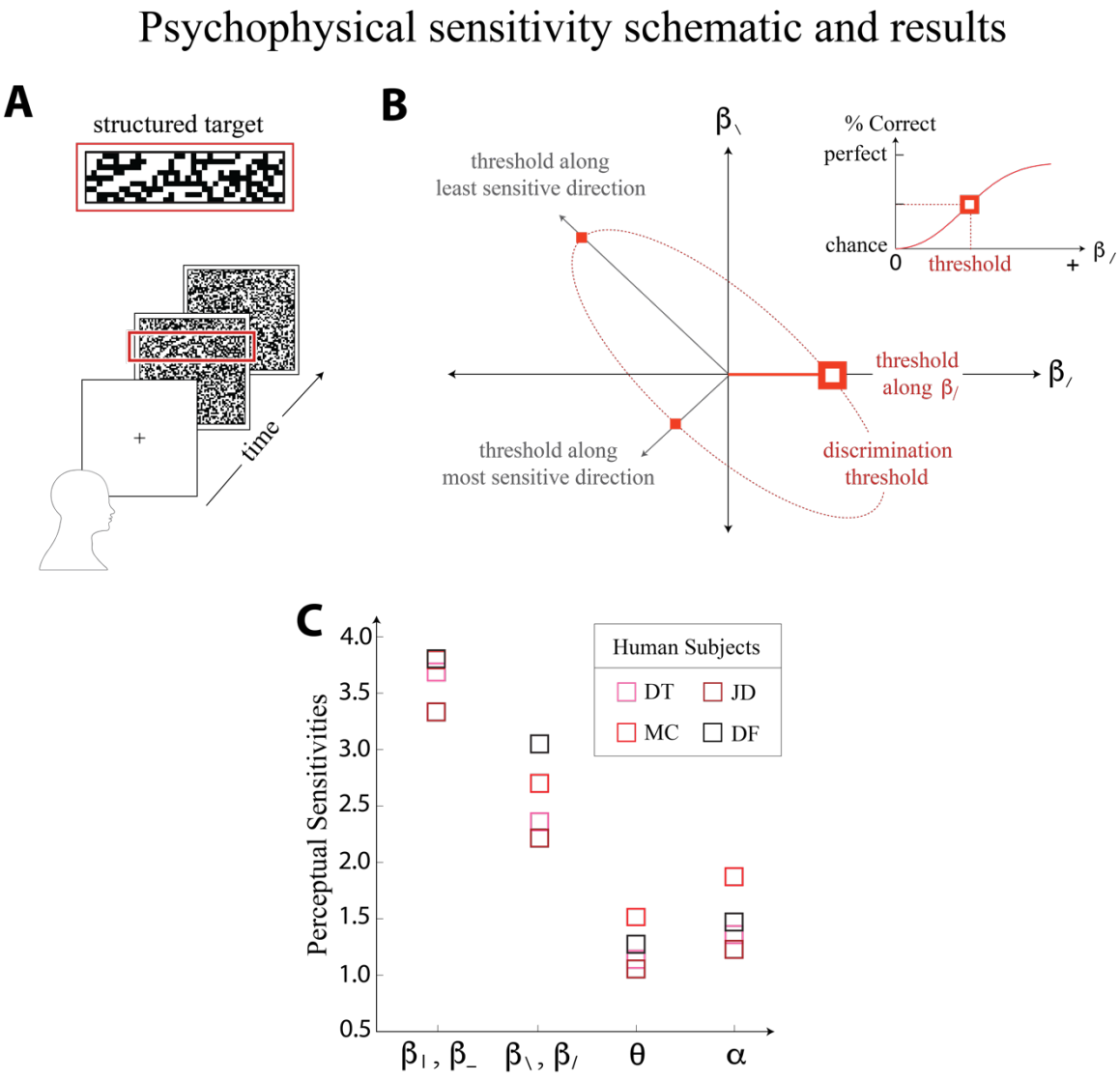


Figure 8: Depiction of psychophysical experimental procedure. The task (schematized in **A**) requires the subject to fixate on the center of the screen before the structured image is displayed. After 120ms, a white noise image is displayed to prevent burn in. The subject has to identify the location of the structured part of the image (top/bottom/left/right). This is done for a variety of texture parameter values, allowing the calculation of a threshold (where the subject reaches halfway between chance and perfect) for each coordinate, as well as oblique directions in each 2-dimensional subplane(panel **B**). The results for single-coordinates are displayed in panel **C**, featuring the same rank-ordering as found in the natural image analyses. Here, symmetries in psychophysical sensitivities suggested reporting single values for texture classes with indistinguishable thresholds.

class, as seen in Figure 8, panel **C**. This is very different from what an ideal observer would display, which would be equal sensitivity in each single coordinate direction [16]. Each subject performed 4320 trials per plane, totaling 47520. For more details about the psychophysical experimental procedures, see [20].

### *Comparing natural image statistics to human psychophysical sensitivities*

Since we expect the *sensitivity* to grow with the signal strength, and the psychophysical threshold to be small for parameters to which we are very sensitive, we should compare the standard deviations found in natural images to the inverse of the psychophysical threshold. After allowing for a single overall scale factor for each set of image processing parameters, plotting these quantities against one another (see Figure 9**A**) shows a striking degree of similarity. In addition to the robustly preserved rank-ordering, the relative magnitudes of the standard deviations match the relative magnitudes of the psychophysical sensitivities. It is also interesting to observe that the variability between image analysis parameters is similar to the variability between subjects.

Seeing this striking level of agreement for individual coordinates is very interesting, but our choice of single coordinates was simply using a convenient basis, rather than describing a fundamental set of independent parameters. We therefore need to examine the covariance structure of these signals, and compare the thresholds predicted by the inverse covariance matrix given from the natural image statistics to the threshold ellipses measured from human subjects. A comparison of these ellipses is shown in Figure 9**B**, for a single set of image processing parameters to reduce clutter, although

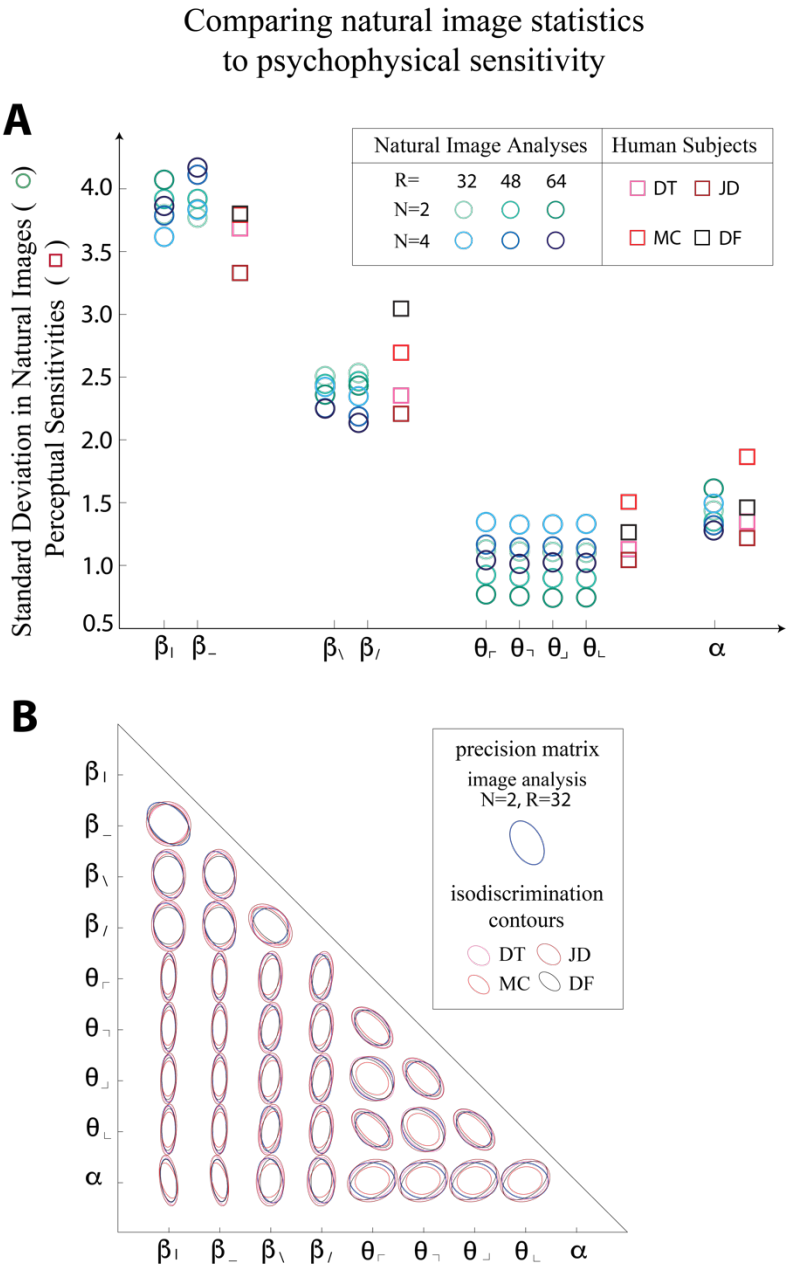


Figure 9: Comparing natural image statistics to psychophysical sensitivities. In panel **A**, the psychophysical sensitivity, given by 1/threshold, is plotted in red. Natural image standard deviations, plotted in green-blue, have each been allowed a single scale factor for each set of processing parameters, since the overall magnitudes need not directly reflect the psychophysical sensitivity. The degree of variability in image analyses is similar to the degree of variability between subjects. In panel **B**, the threshold ellipses for each subject are plotted in red along with the threshold ellipse predicted from the natural image statistics.

although results are similar across image analyses as well (for more detailed measurement, see [2]). We quantified the elliptical parameters, eccentricity and tilt, to measure the agreement (see Figure 10) between the ellipses, but note that when eccentricity is small, tilt becomes meaningless. Note that for the elliptical parameters, there *is no scale factor at all*, and the prediction made here has no free parameters, as the scale factor only affects the overall size of the ellipse.

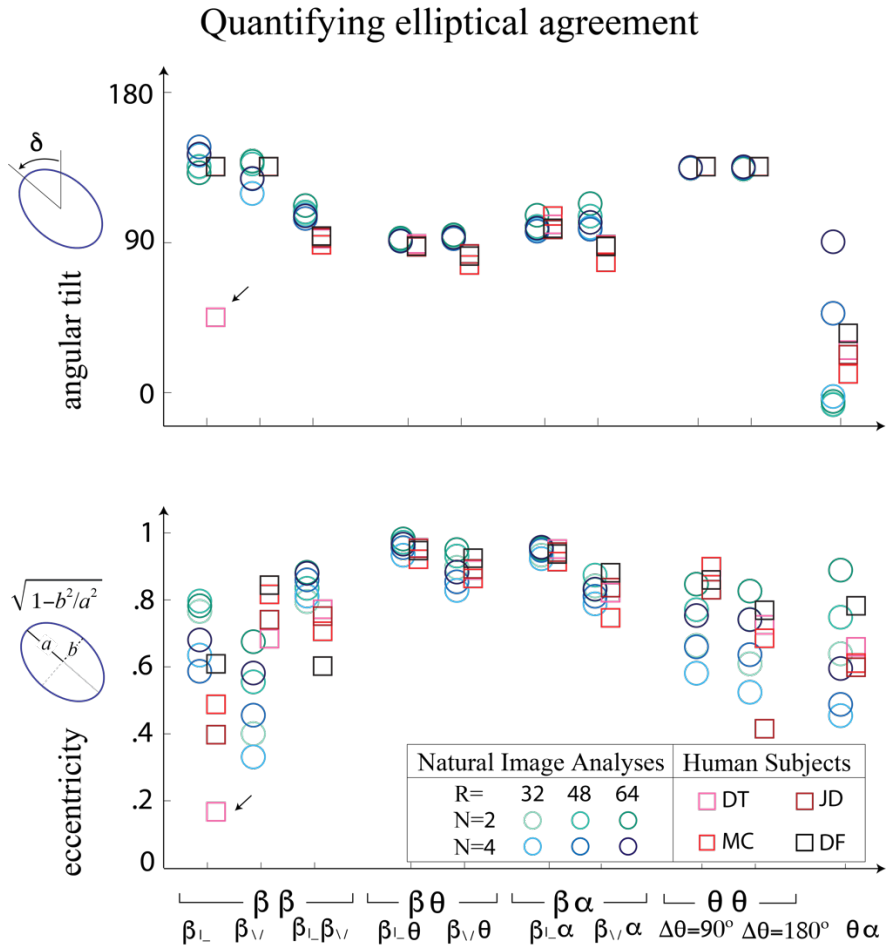


Figure 10: Quantifying elliptical agreement. Angular tilt (top) and eccentricity (bottom) plotted for a variety of image processing parameters and each human subject. The eccentricities and tilts agree to a large extent. This comparison is parameter free, as the scale factor only affects the overall size (area) of the ellipse.

*Discussion of binary results*

Here, we have proposed an idea governing the organization of neural circuits that makes predictions at the level of human behavior. This is a very powerful statement about the nature of neural circuit organization. For the purposes of this study, we did not even need to make direct measurements of cortical activity. The intuition behind the suggested coding scheme is that for signals which have relatively high uncertainty, it is worth devoting more resources to looking at signals that are more variable. In this case, sampling limitations for local texture features imply that signals with larger variability are more useful in distinguishing between natural images. It is interesting to note that a strength of the comparison made here is between a set of artificial textures and statistics computed from natural scenes. A strength of this study is that, despite the seemingly unnatural structure of the artificial stimuli, we were able to predict their salience to human subjects based on observations about how the texture parameters characterize natural images.

It is also interesting to note that the perceptual thresholds likely arise from cortical processing, as this implies that the efficient coding hypothesis is not only a useful tool to apply in the extreme sensory periphery, but can useful for understanding central processing as well. The stimuli contrast was very high, and the pixels were easily visible (14 arcmin), meaning that retinal limitations for contrast sensitivity and spatial resolution were not limiting factors for the discrimination. It has also been shown that cat retinal populations show no sensitivity to the four-point correlations, but simultaneous visual cortex field potential measurements do [21]. Similarly, neurons in macaque visual cortex elicit responses to three- and four-point correlations [19]. Furthermore, the efficient

coding regime that makes these predictions has input-noise as the dominating parameter limiting performance, which differs from the one traditionally applied to understand peripheral vision. In peripheral vision, the optic nerve applies a heavy constraint to output power, and output noise is the limiting factor. The 'whitening' regime, as it is called, calls for neural resources to be devoted with an inverse relation to the variability. For example, the retina has greater sensitivity for low spatial frequencies than high spatial frequencies, reflecting the $\sim 1/f^2$ power spectrum observed in natural images. This difference in coding constraints observed for peripheral and cortical vision could provide important insights into coding strategies used elsewhere in cortex.

It is also interesting to note that we observed more evidence for scale invariance in natural images. Image analysis parameters changing the scale of the scene (block average factor) did not significantly alter any of the significant findings of these texture statistics, which suggests scale invariance is a useful way of thinking about natural scenes in more ways than just predicting the frequently observed $1/f^2$ power spectrum.

This work is building on a larger class of studies examining the role of neural coding for visual texture perception. Previous studies within our own collaboration [14] have shown that certain high-order correlations that are present in natural scenes are not perceptually salient at all, finding that their presence in natural scenes is entirely explainable from shorter-range correlations. Other studies [22], have found manipulations to higher-order statistics in images that deform images in a manner that are undetectable to a fixating human(but readily observed when your gaze wanders). Both of these studies have quite a similar flavor, and are consistent with the coding model presented here as

elements which fit into the non-coding region. We have taken this a step further and shown that, for higher-order correlations that humans are sensitive to, we seem to be sensitive to them in proportion to their variability.

Another interesting implication of this line of work applies to situations that rely on human experts to examine highly unnatural images. In medical imaging, for example, it can be very difficult for an untrained eye to spot a defect or a fracture, particularly in a small bone or the appearance of a small tumor. It may be possible that these types of images, which certainly have highly different statistics from natural images, have a significant amount of information stored in local correlations that are difficult for humans to detect. If it were possible to effectively 'rotate' the coordinates so that the informative ones align with the ones humans are naturally sensitive to, it may make diagnoses based on medical image data much easier and more reliable. Some research in this direction has already begun.

### *Extension to grayscale images*

It is of course natural to want to extend these kinds analyses to grayscale images, as our experience of the world has nearly a continuum of luminance values, rather than just black and white. Analogous grayscale textures can be computed using the methods established in [16]for finite grayscale levels. We will start by examining textures with 3 grayscale levels in the same 2x2 pixel block. The basis we will use is related to the number theoretic Fourier transform, and spending some time describing this will be useful. In a 2x2 grid, we can label the pixels starting at the top left and going clockwise A, B, D, C. The manner of describing the relevant textural configuration is using these

letters, so $AB$ will denote a horizontal 2-point correlation, while $BCD$ denotes a 3-point correlation in a configuration that excludes the top-left corner. Previously, with only two grayscale levels, we used the parity of the block. Now, it is helpful to think of the patterns with respect to arithmetic mod 3, where a black pixel is labelled 0, a gray pixel labelled 1, and a white pixel labelled 2. Since each of these individual patterns is well-defined, we can deconstruct it into probabilities that the sum of individual grayscale values is equal to a specific value. For example, the $AB_{12}$ coordinate has probabilities associated with $P(A + 2B = 0\ mod\ 3)$, $P(A + 2B = 1\ mod\ 3)$, and $P(A + 2B = 2\ mod\ 3)$. These probabilities must sum to 1, and so there are only two free parameters describing this coordinate subspace. In the binary case, we had only a single value to describe these coordinates because there were two possible values the combined coloring could take. The patterns generated by this $AB_{12}$ are that $A = B$, so no change as pixels move in the horizontal direction [000…/111…/222… depending on initial value] when $P(A + 2B = 0\ mod\ 3) = 1$; the cyclic pattern [0210210…] when $P(A + 2B = 1\ mod\ 3) = 1$; and the cyclic pattern [012012…] when $P(A + 2B = 2\ mod\ 3) = 1$. The subspace of values these three probabilities can take lies within a triangle bounded by the three probabilities taking values between 0 and 1, and summing to 1. Overall, there are 33 different patterns, each with two degrees of freedom, totaling 66 dimensions (2 first-order, 16 second-order, 32 third-order, 16 fourth-order).

We can compute these quantities for natural images following a very similar processing pipeline as before, but instead of binarizing at the pixel intensity median, we "trinarize" with equal number of white, black, and gray pixels. This gives us the

distributions of these parameters in natural images. Psychophysical measurements have been carried out to analyze human sensitivity to several subspaces [23]. We will compare some our analyses of these natural scene statistics to the psychophysical measurements. Comparing the thresholds predicted using the same analyses to a subset of the planes containing 2-point correlations (plotted in Figure 11), we see agreement in the orientation and eccentricity of the ellipses for two subplanes ($AB_{12}$ and $AD_{12}$), but a lack of such strong agreement in two other less eccentric planes ($AB_{11}$ and $AD_{11}$). This is an interesting finding in its own right, and remains to be seen why agreement exists in some ways, but not in others. One possibility is that the neural mechanisms for encoding these highly complex features are heuristic, and therefore unable to capture every detail of the distribution, but prioritize coding the most important and salient features.

To further analyze this data, it is useful to use principle component analysis (PCA) to analyze where the bulk of the distribution is concentrated. Upon doing so, the first interesting feature that pops up is the eigenvalue spectrum (plotted in Figure 12**A**). There are 99 principle components because the covariance analysis here is performed using the full probability values, but 35 dimensions are null. This is expected because normalization reduces the number of free parameters to 66, and our "trinarization" process fixes the probability of having each, black, white and gray colored pixels. Then, we observe that the bulk of the eigenvalues are quite small compared to the variance of the first few components. In fact, nearly 75% of the variance in the dataset is contained within the first 10 principle components. These first 10 principle components are primarily composed of second-order statistics, with a few contributions from third- and

fourth-order statistics. This is consistent with the psychophysical observation that many second-order statistics are salient, but few third- and fourth-order ones are. Furthermore, the principle components provide insight into the natural structure of visual scenes, and may provide insight into the kinds of symmetries we may expect to observe psychophysically. As an example, plotted in Figure 12**B**, are coefficients of three of the first five principle components. The first column corresponds to the probability that the sum is equal to zero, the second column to the sum being one, the third to the sum being two. The principle component that contains the largest amount of variance in the data contains has the most significant contributions occurring equally from $AB_{12}$, $AC_{12}$, $BC_{12}$, and $AD_{12}$ (and the relevant sum equaling zero), all with positive coefficients. An interesting feature of this vector is that it is approximately symmetric under rotating the underlying image by $\pi/4$. Two more of the first few principle components contain similar contributions from two-point correlations that span a similar subspace as the most significant component, but differ in that positive and negative coefficients imply that this element is actually *antisymmetric* under the operation of rotating the image by $\pi/4$. This natural symmetry may manifest itself in an important way, and suggests that one non-trivial two-dimensional subspace of interest is, for example, the $P(AB_{12} = 0) - P(AC_{12} = 0)$ plane, because we have strong predictors along oblique directions within this plane. This analysis would additionally shed light on the role of this underlying approximate symmetry of natural images has on perception.

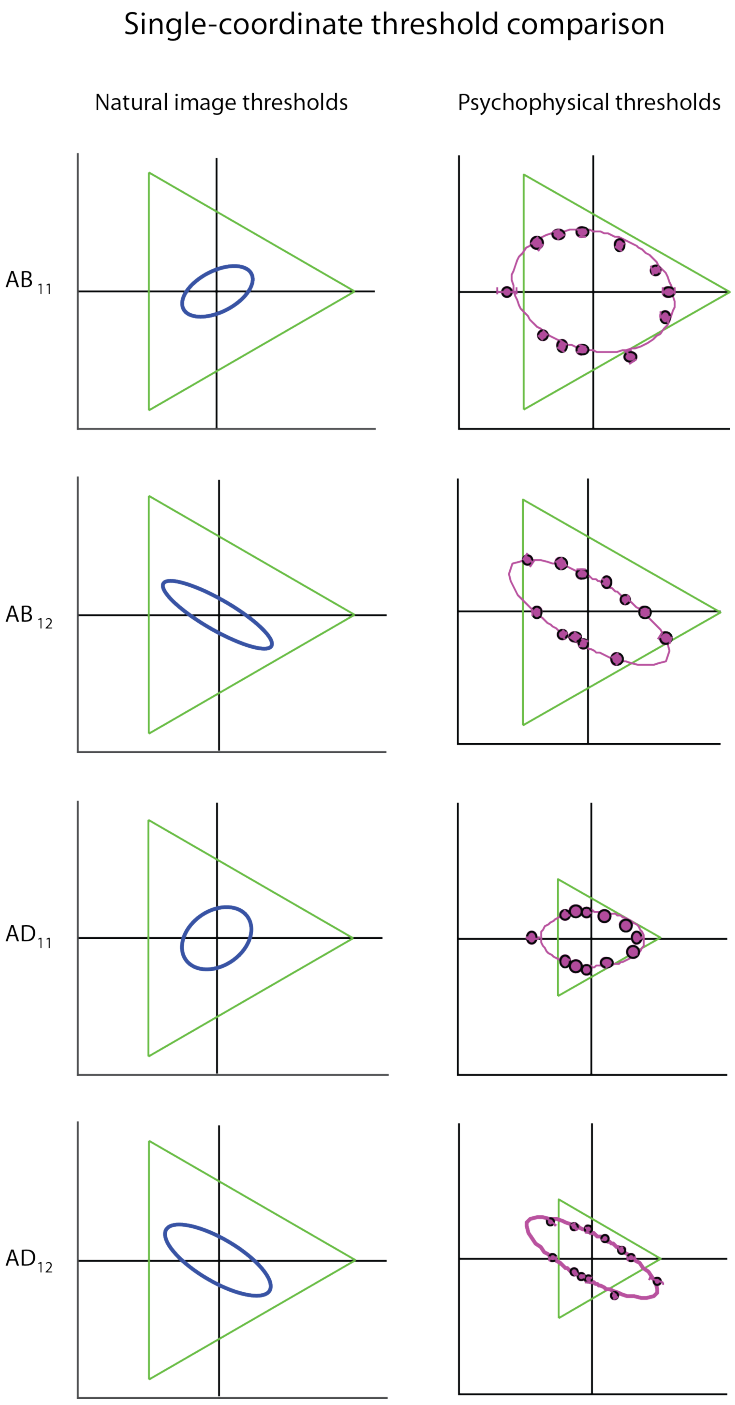


Figure 11: Comparing single-coordinate thresholds. Each plot is a projection of the coordinate space $P(X = i \ mod \ 3)$, where $X$ is the relevant coordinate equation (labelled to the left). The bottom left corner, right corner, and top left corner correspond to $P = 1$ when $i = 0, 1$, and 2, respectively. Psychophysical measurements are plotted for a single subject, though are representative of other subjects. We see agreement in the $AB_{12}$ and $AD_{12}$ subplanes, but a noticeable lack of agreement in $AB_{11}$ and $AD_{11}$ subplanes.

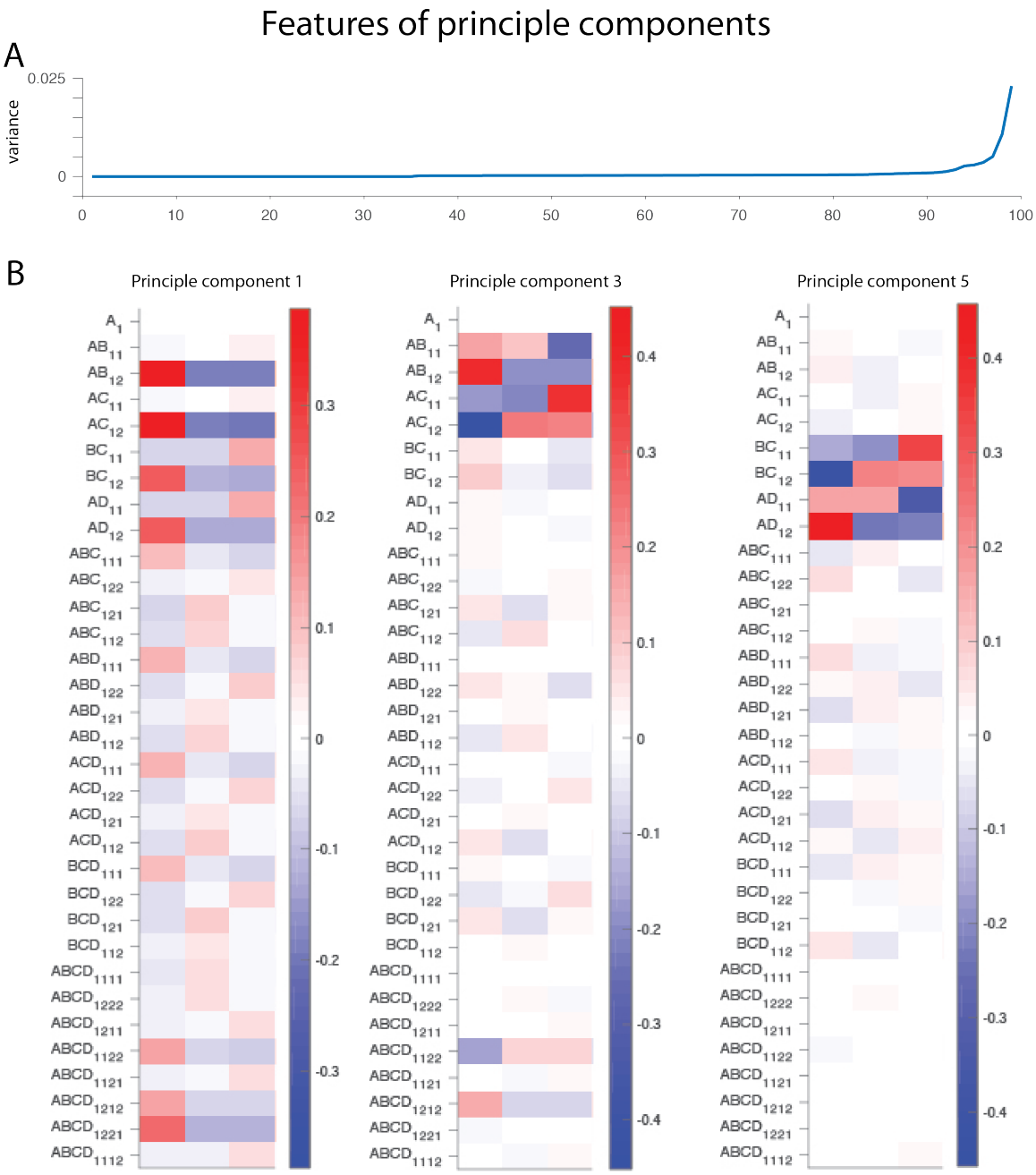


Figure 12: Features of principle components. In panel **A**, variance is plotted for each principle component (labelled in ascending order of variance), showing there are likely relatively few dimensions in the space where behavioral sensitivity is measurable. In panel **B**, we see the structure of some of the most significant principle components (labelled in descending order of variance) respects intuitive transformations of the environment. The largest principle component is approximately symmetric under rotations of $\pi/4$, while the third and fifth span a similar subspace, but are approximately antisymmetric under rotations of $\pi/4$.

Here we have taken steps in the direction of extending these analyses towards grayscale measurements, and we have seen that as the size of the space increases, the importance of natural image analysis is more important. Brute force cannot be used to measure thousands of 2-dimensional subspaces, so it is important to identify particularly important ones to look within, especially when we expect that few will be detectable at all. We have shown that the ideas formulated at the beginning of this analysis still provide useful predictions using more complex stimuli, but may have stumbled upon some instances where our theory begins to break down. This is where we may learn new things—whether it is about some kind of change in the coding scheme our visual system employs, or features of human perception limited by heuristic solutions used by our visual system, following our theory until it fails leads us to learn something new.

# 3. Neural populations predictive of frequency discrimination behavior in mice

## *Large neural populations and information encoding*

An important step toward understanding the neural code is establishing limitations it provides for behavior. In the previous chapter, this was exemplified using natural images as the source. In this chapter, we inspect neural responses in auditory cortex to different tone frequencies in order to see whether or not this activity can explain behavioral limitations of the animal. Previous studies have tried to address similar questions for identifying heading direction [24] and for sound localization [25], but no direct link has been drawn for frequency tuning. In fact, the role auditory cortex plays in frequency discrimination has a few subtleties to it. For example, some studies have found that pharmacological suppression [26] and lesions of human AC [27] impair frequency discrimination. However, other lesioning [28] and pharmacological [29] studies have shown little effect. Many neurons in the auditory cortex are frequency tuned, and respond more strongly to some frequencies than others. Moreover, this frequency tuning can be changed by learning [30] [31] [32] [33] [34]. Recent work within our lab has shown that optogenetic manipulations of auditory cortex change the behavioral frequency discrimination performance of mice. More specifically, activating PV interneurons in auditory cortex on average led to *improvements* in frequency discrimination performance, while suppressing them led to impaired performance [3]. It is quite interesting that a manipulation improved performance, because that rules out the possibility that the neural circuitry in auditory cortex is tuned to optimize performance in this kind of sensory task.

If this were the case, *any* manipulation of the circuitry would impair performance. This information suggests that, even though auditory cortex is not necessary for frequency discrimination, it still plays an important modulatory role. We will examine this role on an individual-by-individual basis, and our work suggests that individual differences in frequency discrimination performance may be tied to differences in the underlying activity of the AC.

In order to establish a link between neural activity and behavioral frequency discrimination, there are several challenges to overcome. Neural recordings significantly subsample the population (there are ~$10^5$ neurons in mouse auditory cortex), and there is no clear mapping from the subset of neurons in one mouse to those in another. The way we control for this effect is by using the same neural responses, and taking advantage of the fact that optogenetic manipulations of neurons in AC lead to changes in (i) the behavioral thresholds exhibited by individual mice and (ii) the neural responses exhibited to tones. We can therefore make a direct comparison between the thresholds predicted from the population and the behavioral thresholds in both, light-on and light-off conditions. Although the absence of recordings for many neurons from the population may make predicting the *absolute* behavioral threshold challenging, the *change* in threshold should be similar if we have the same subset of neurons embedded in the same population, so long as the changes in the subpopulation are representative. We will first present the methods for computing the limitations the neural activity gives for frequency decoding. We then discuss characterization of neural responses in AC, and how they may be used to calculate an empirical estimate of frequency discrimination performance. Methods for measuring behavioral frequency discrimination follow. We then compare the

thresholds found neurometrically to those found behaviorally, and discuss the implications of our findings. Finally, the chapter closes by proposing follow-up work that could shed light on the role AC plays in learning and the implications these sensory limitations have on behaviorally relevant stimuli.

## *Computing Fisher information from a neural population*

How can one quantify the discriminability between two inputs from something which encodes them? For example, if we know the neural response (including variability in the response) to two different tones, we should be able to be able to estimate how well the neural activity can distinguish them. For two randomly selected tones in the auditory spectrum, the neural responses will most likely be drastically different, allowing one to easily determine which tone was played using the neural responses. However, if the two tones happen to be quite close to one another, we need a way of estimating the distinguishability of the tones. *Fisher information* is a useful quantity to examine whenever the underlying signal is naturally described as a continuous variable (tone frequency, orientation of a bar, or velocity of moving object, to name a few), and is defined by:

$$I_F(f) = \sum_{\vec{n}} P(\vec{n}|f) \left( \frac{\partial}{\partial f} \log P(\vec{n}|f) \right)^2$$

where $f$ is the signal being encoded (*frequency* for the uses in this chapter), $\vec{n}$ is the vector denoting the response of neural population (where each dimension represents a neuron in the population, and its entry is an integer specifying the number of times it spiked), and $P(\vec{n}|f)$ is the likelihood function describing the probability that a particular spiking pattern is observed *given* that the signal input is *f*. The Fisher information is large

wherever the probability distribution changes quickly, which captures the intuition that distinguishing nearby frequencies requires the neural response to change rapidly as the frequency shifts. In fact, any unbiased estimator $\hat{f}$ based on the neural responses will have a lower bound on its variance calculable from the Fisher information, $Var(\hat{f}) \geq \frac{1}{I_F(f)}$. In other words, this quantity sets the length scale in signal space of how far apart another frequency must be for any criterion level of detection, and therefore bounds the optimal performance. The neurometric threshold, which describes the length-scale in signal space for a criterion performance, is therefore defined to be $I_F(f)^{-1/2}$.

But how do we apply this to the responses of neural populations? One useful approximation is to assume that neurons respond independently of one another. This is clearly untrue in general, as most neurons are excited directly by other neurons, but when considering a set of neurons with inputs dominated by inputs from a different brain region, it is not a bad one. If neural responses to the input are independent, we can write

$$P(\vec{n}|f) = \prod_i P(n_i|f)$$

where $P(n_i|f)$ is the probability that neuron $i$ will spike $n_i$ times in response to the frequency. This quantity is much easier to measure experimentally. This additionally makes computation of the Fisher information simpler, because the sum factors into a sum of the Fisher information of individual neurons. Without independence, $k^N$ probabilities must be computed, where $k$ is the maximum number of spikes possible during the time period of interest and $N$ is the number of neurons. With the assumption of independence,

only $N * k$ probabilities must be computed, allowing the analyses to scale for reasonably large populations.

Another useful approximation is that individual neurons respond with Poisson statistics, which is the case when a neuron receives inputs that bias it to fire at a certain rate, but the individual spiking events rely on a stochastic process to occur. Mathematically, this means

$$P(n_i|f) = \frac{e^{-\mu_i(f)}\mu_i(f)^n}{n!}$$

where $\mu_i(f)$ is the mean number of spikes expected in response to stimulus $f$. One property of the Poisson distribution is that the *Fano factor*, defined as the ratio of the variance to the mean of the response, is equal to one. It has been observed that though it is a good approximation, this does not always hold for real neurons [35], and many neurons have larger Fano factors than this. We will ultimately be interested in relaxing this constraint, and so we will also use the generalized Poisson distribution, defined by

$$P(n_i|f) = \frac{\alpha_i(f)(\alpha_i(f) + n_i\lambda_i)^{n-1}e^{-(\alpha_i(f)+n_i\lambda_i)}}{n_i!}$$

where the additional parameters $\alpha_i(f)$ and $\lambda_i$ are related to the moments of the distribution. More specifically, $E[N_i] = \frac{\alpha_i(f)}{1-\lambda_i}$ and $Var[N_i] = \frac{\alpha_i(f)}{(1-\lambda_i)^3}$. $\lambda_i$ can be expressed in terms of the Fano factor, $F_i$ as $\lambda_i = 1 - F_i^{-1/2}$, which leaves $\alpha_i(f) = \mu_i(f) * F_i^{-1/2}$ . When $\lambda_i$ is zero, this reverts to the standard Poisson distribution, but allows an extra degree of freedom to control the ratio of variance to mean, allowing us to capture more realistic properties of real neural populations. With these tools, we able to compute from a neural population what the limits on its performance will be. This will allow us to test

the hypothesis that the ability to decode frequency from responses in auditory cortex predicts the behavioral thresholds exhibited.

### *Measuring from a neural population in auditory cortex*

The key to utilizing neural responses to predict something about behavioral frequency discrimination is to characterize the response to tones. Neural activity was measured in awake, head-fixed mice using 50 tones spaced logarithmically between 1 and 80 kHz at 8 different sound intensity levels (from 10-80dB). Tones were presented pseudorandomly for a duration of 50 ms with 450 ms between them (Schematized in Figure 13**A**). The stimulus was counter-balanced to allow for an analogous measurement with optogenetic manipulation. For tones with optogenetic manipulation, the light was delivered for 200ms, starting 100ms prior to tone onset. From this, we computed the frequency response function of the neuron by averaging the firing rate to each frequency at the 3 highest sound pressure levels. This was then fit by a Gaussian frequency response function, as depicted in Figure 13**B**. After retaining only neurons where the Gaussian fit has $R^2 > .6$, we pooled the neural population across each individual mouse. From this set of tuning curves, the Fisher information may be calculated (see Figure 13**C**), from which a predicted threshold may be derived (see Figure 13**D**). More details of recordings can be found in [3].

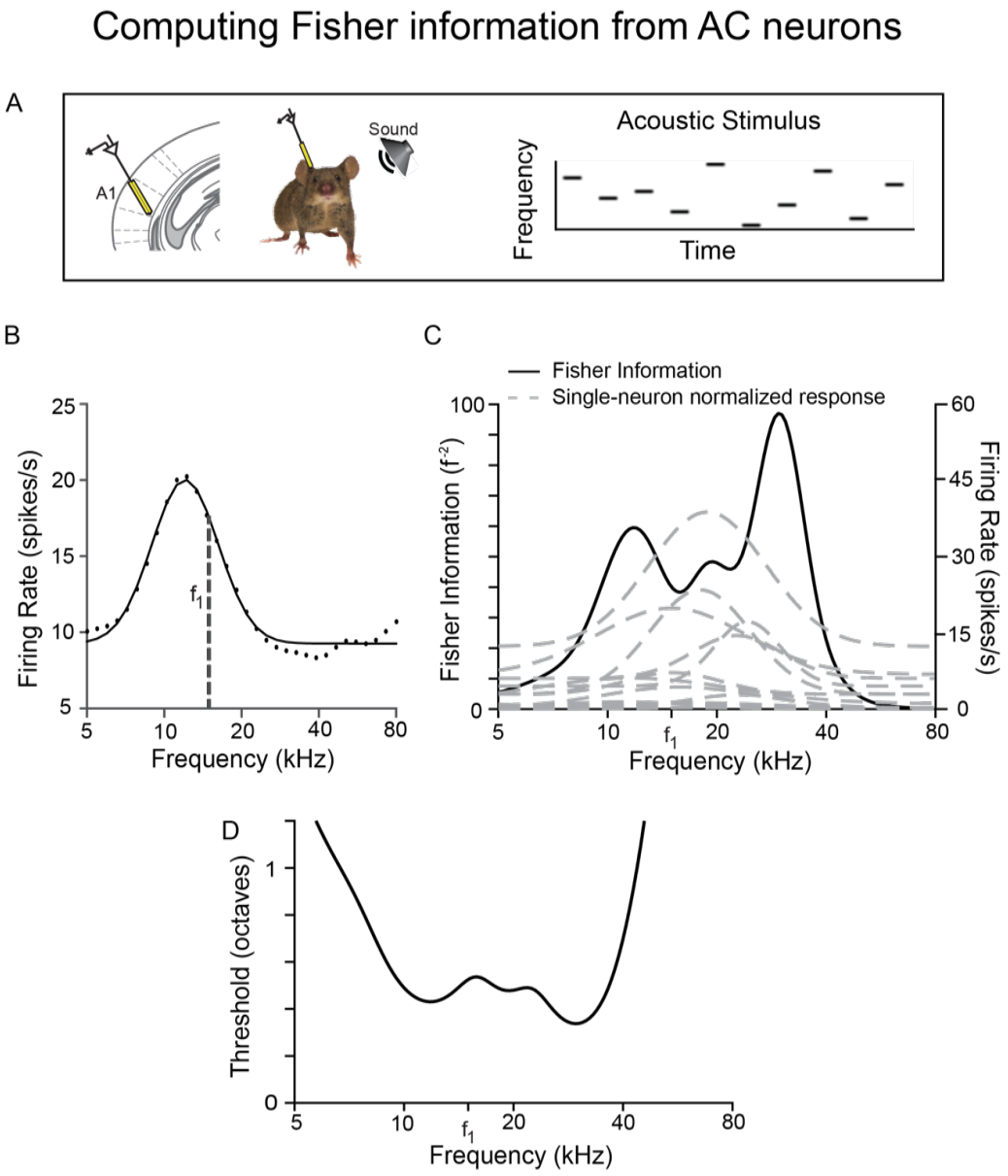


Figure 13: Computing Fisher information from neurons in AC. In panel **A**, neural recordings from AC are performed while a presenting frequencies chosen pseudo-randomly from 50 tones spaced logarithmically from 1-80kHz. Each neuron has a frequency response function (panel **B**) which is fit to a Gaussian (solid line). From the population of tuning curves for a particular mouse, we can compute the Fisher information (panel **C**), which sets a bound on the frequency discrimination profile possible from this population. The threshold is predicted by $I_F^{-1/2}$, plotted in panel **D**. We will be interested in looking at a particular frequency, $f_1$, which represents the frequency at which behavioral frequency discrimination acuity is measured.

*Assessing behavioral discrimination in mice*

In order to measure behavioral performance in mice, we used a pre-pulse inhibition procedure. Essentially, by playing a background tone, followed by a relatively loud burst of white noise, the animal exhibits a startle response. In order to utilize the startle response to measure frequency discrimination acuity (Figure 14**A**), while the animal is standing on a platform that measure paw pressure, we play "pre-pulse" tone for 60ms (10.2, 12.6, 13.8, 14.7, 15.0 kHz) between the background tone (15kHz for a randomly chosen 10-20s) and the noise burst (100dB SPL broadband noise for 20ms). When the pre-pulse tone is indistinguishable from the background tone, there is no reduction in the startle response (Figure 14**B**), but when the pre-pulse tone becomes distinguishable, the startle response is suppressed. This reduction in startle response is termed pre-pulse inhibition (PPI). A sigmoid is fit to the PPI, which is computed from the acoustic startle response by $PPI(f) = \frac{ASR(0)-ASR(f)}{ASR(0)}$. A sigmoid is fit to the PPI curve, and the behavioral threshold is defined as the frequency difference that leads to 50% of the maximum PPI (Figure 14**C**). A major advantage PPI has over other tasks that measure the same quantity, such as go/no-go or 2-alternative forced choice task, is that it is an innate response. The measured acuity is therefore not confounded by the ability of the animal to learn the task, as we will not mistake a decision-making error for a perceptual one. For more details about experimental measurements, see [3].

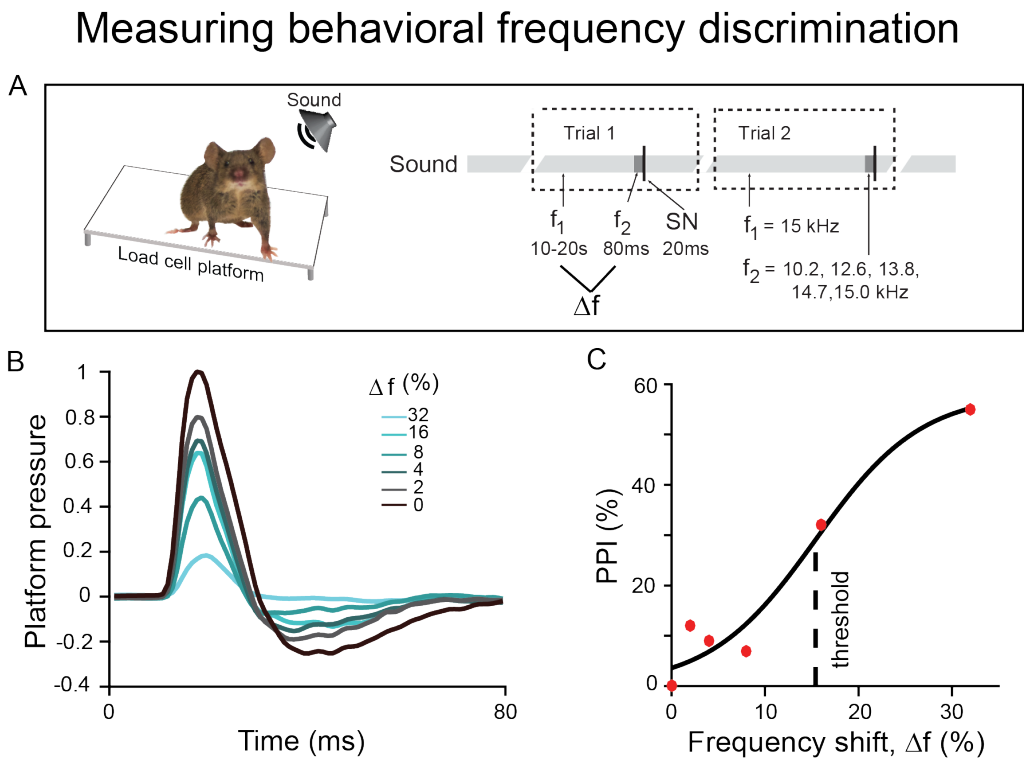


Figure 14: Measuring behavioral frequency discrimination. Animals stand on a platform that measure paw pressure while a series of three sounds are played (panel **A**). A background tone is played for 10-20s, followed by a tone of variable frequency, before finally a brief burst of broadband noise. The acoustic startle response (ASR) is reduced as the pre-pulse tone becomes increasingly different from the background tone (panel **B**). The pre-pulse inhibition (panel **C**) measures the ASR reduction as a function of frequency shift. The threshold is defined by the frequency difference yielding 50% of maximum PPI.

## *Effects of optogenetic manipulations on behavior and recordings*

The broad hypothesis here states that regardless of the specifics of the manipulation, the effects that are salient to behavior will be captured by changes in Fisher information of the individual neurons. We used 3 different optogenetic manipulations, including expressing ChR2 in in PV+ interneurons, Arch in PV+ interneurons, and ChR2 in pyramidal neurons. This allows us to excite PV+ interneurons (inhibiting typical pyramidal neurons), inhibit PV+ interneurons (disinhibiting typical pyramidal neurons), and excite pyramidal neurons, respectively. For more information on how the optogenetic

manipulations were performed in this case, see [3]. We observed salient changes in the baseline activity of neurons during optogenetic activation of each class of neurons. Activating PV interneurons tended to reduce the baseline activity of most recorded neurons (Figure 15**A**), while suppressing PV interneurons led to a slight increase in typical baseline activity (Figure 15**B**). Activation of pyramidal neurons led to primarily increases in neural activity (Figure 15**C**). We also observed changes in the behavioral thresholds between baseline and optogenetically modified conditions. Exemplar PPI curves for each type of optogenetic manipulation are plotted in Figure 15, panels **D-F**. Most animals with optogenetic activation of PV interneurons had improved discriminability (reduced threshold), but not all. In the small sample sizes reported here, animals with suppressed PV interneurons displayed increased and decreased thresholds, while the mice whose optogenetic manipulations activated pyramidal neurons displayed increased thresholds. The changes in frequency response function properties under the influence of optogenetic manipulations lead to changes in the Fisher information profile of the population, and therefore a change in the predicted threshold. The threshold curves predicted with and without optogenetic manipulations (assuming a Poisson noise model) are plotted along with the measured threshold in Figure 15 panels **G-I**. Note that the behavioral thresholds are much lower than the predicted neurometric thresholds. This is expected because the neural populations had between 10 and 100 neurons, a small fraction of the neurons in auditory cortex that contribute. Also note that the behavioral threshold is valid only where it was measured, at $f_1$. In each of these three examples, we see that the change in the threshold of the neural population is in the same direction as the

changes in the behavioral thresholds. This is suggestive, but we still need to control for differing population sizes and actually measure how much the change in neural threshold correlates with the change in behavioral threshold.

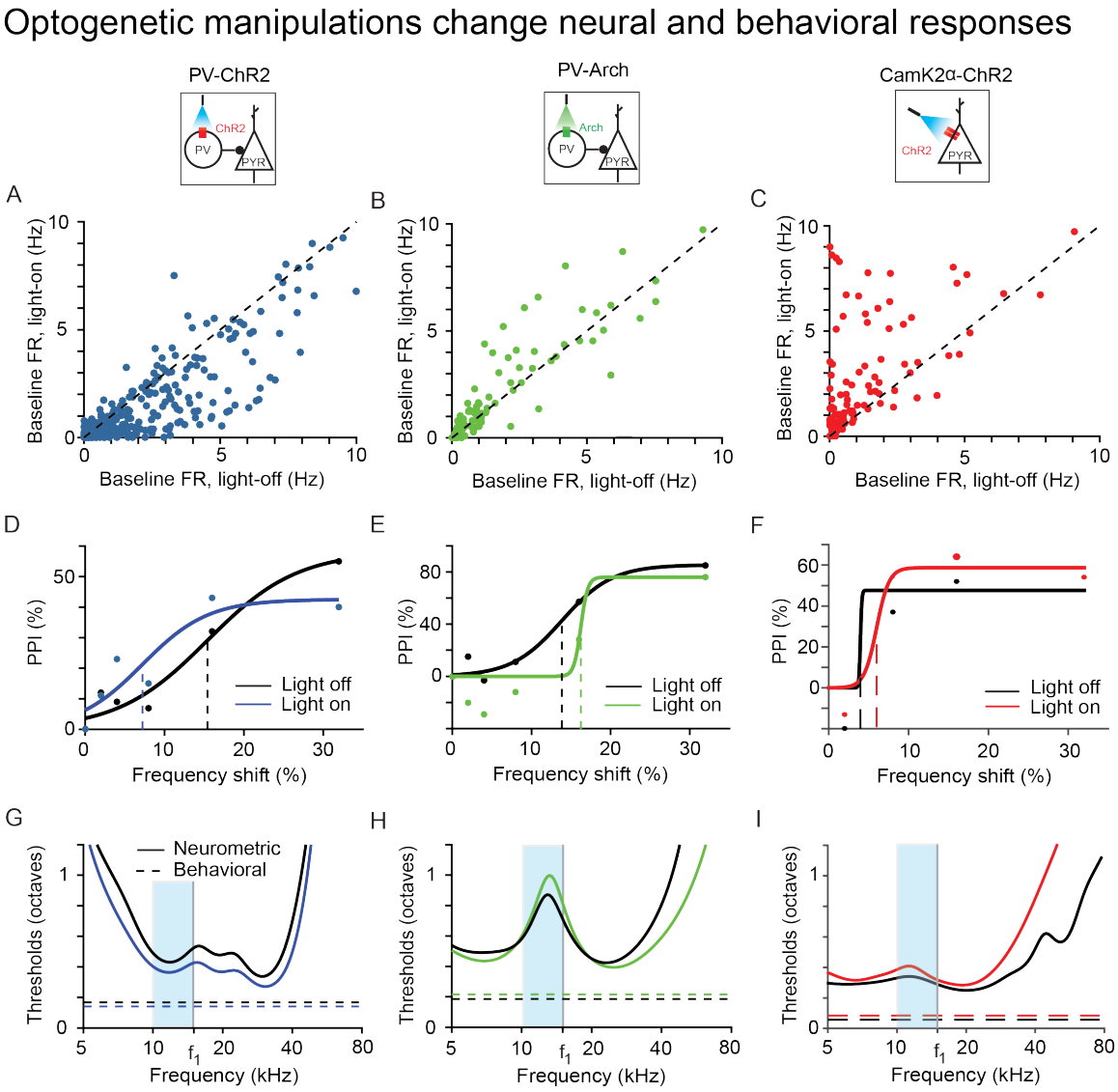


Figure 15: Optogenetic manipulations change neural and behavioral responses. The left column corresponds to optogenetically activating PV interneurons, the middle column corresponds to optogenetically inhibiting PV interneurons, and the right column corresponds to optogenetically activating Pyramidal neurons. Panels **A**, **B**, and **C** show that activation of PV interneurons leads to a reduced baseline firing rate for most neurons, suppression of PV interneurons leads to a slight increase in baseline activity for most neurons, and activation of Pyramidal neurons leads to increases in the activity of most neurons. Panels **D-F** show exemplar mice from each type of manipulation. Most mice in the PV activation category displayed improved threshold with the optogenetic manipulation. Panels **G-I** show the Fisher information plots under both, light on and light-off conditions in comparison to their behavioral thresholds (measured in **D-F**). The curves are continuous because they are computed from Gaussian fits to neural responses, rather than directly to data.

*Trends across mice*

As previously explained, we do not expect neural thresholds to capture the full behavioral acuity due to the significant subsampling of cortical neurons. To compare predicted sensitivities across mice, since each mouse had a different number of neurons, we normalized the predicted threshold. Assuming independent neurons and that Fisher information per neuron was representative of the other neurons in the animal, Fisher information can be written as $I_F^{tot} = \sum_i I_F^i = N * I_F^{avg}$. The threshold is then given by $t_{neu} = \left(I_F^{avg}\right)^{-1/2} * N^{-1/2}$. The number of neurons assumed about the population controls only the magnitude of the thresholds, and the average Fisher information controls the relative sizes. The resulting normalized neural thresholds are plotted assuming 400 neurons (chosen because it is approximately the number needed to reconcile the absolute magnitude of behavioral discrimination with the neural threshold predictions) in Figure 16**A**, which include a light-off and a light-on measurement for each mouse connected by a grey line. There is a statistically significant correlation between these quantities (C=.35, p=.03, N=38, including a light-off and light-on measurement for each mouse), which suggests that the neural thresholds predicted from individual neural measurements is informative about the behavioral acuity displayed by the animals. The correlation strength is not particularly strong, but it is surprising to see a significant effect at all because, in addition to the sampling limitations, there is no way to control for the subset measured corresponds in any meaningful way to the subset predicted elsewhere.

A more salient effect is found by examining the index of change under the light-off and light-on conditions, given by $I_{change} = \frac{t_{on}-t_{off}}{t_{on}+t_{off}}$. This quantity is equal to zero when there is no change, and equal to 1 when the threshold with the optogenetic manipulation increases significantly. Note also that it is unaffected by the scale factor used to compare the absolute magnitude of predicted neural thresholds. This measure (plotted in Figure 16**B**) quantifies, in some sense, the size and direction of the grey lines in Figure 16**A**. The behavioral index of change is significantly correlated with the neurometric index of change (C=.59, p=.008, N=19). The line of best fit (plotted in gray) has a slope of .25, which may have implications for the type of decoding being performed. The index of change also has the advantage that we are comparing the same subset of neurons embedded within a population in the same manner. It is important to note that in several instances, the same optogenetic manipulation evokes *different* behavioral responses in different individuals, and this difference is often predicted by the neural population.

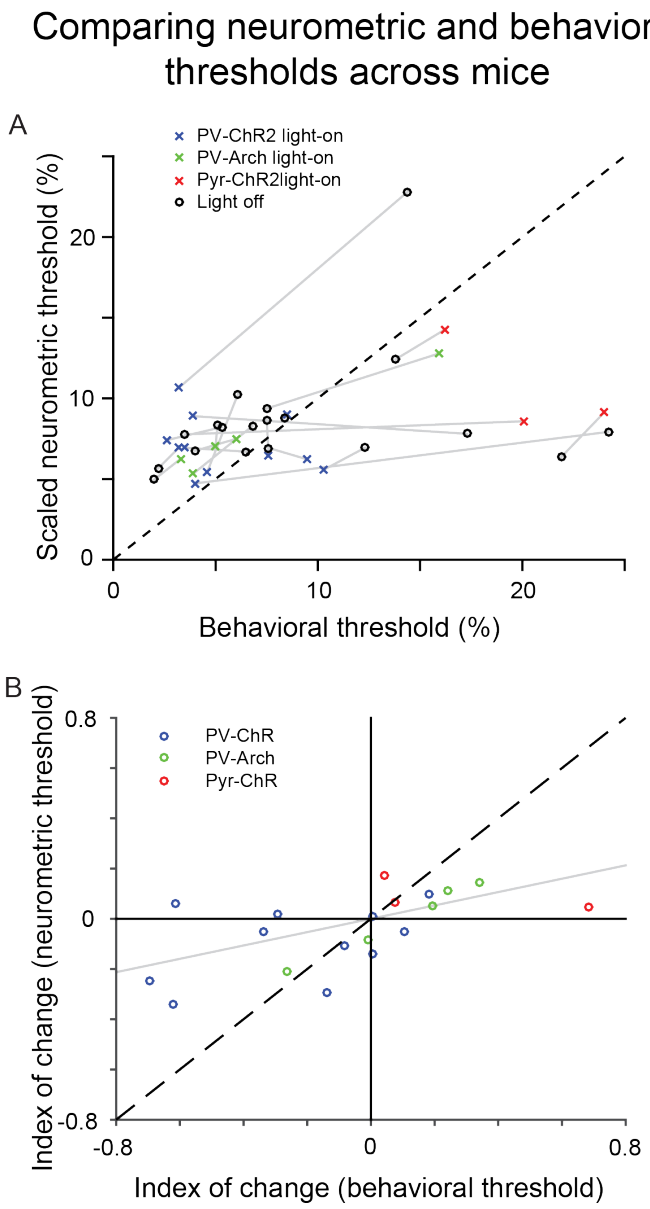


Figure 16: Comparing neurometric and behavioral thresholds across mice. In panel **A**, we have plotted the scaled neurometric threshold against the behavioral threshold. Neurometric threshold is scaled to reflect a population of 400 neurons with the same Fisher information density to account for differences between number of reliable units recorded for each mouse. Black 'x's correspond to the measurement of an animal without any optogenetic manipulation. Colored circles (attached to x of the same mouse by a light grey line) indicate the threshold measured during corresponding optogenetic manipulation. In panel **B**, we see the index of change, which is the difference in thresholds under light-on and light-off conditions divided by the sum. The top right quadrant exhibit higher thresholds with manipulation, while the bottom left quadrant had improved acuity with manipulation. The light grey line is the line of best fit.

*Accounting for neural variability and correlations*

The results presented so far have assumed independence in neural responses and Poisson-like variability in the number of spikes. This is not generally true in neural systems. It is important to measure these quantities to ensure that our approximation is a reasonable one for our dataset, and that these neglected feature are not important in explaining the observed phenomenon. Because our data had relatively few repeats of any specific stimulus (5 at each frequency and intensity), we will have to pool across trials which may have an underlying rate difference.

In order to calculate the Fano factor, which measures the ratio of variance in neural spike count to mean activity rate, we calculated the mean and variance of each recorded neuron at each frequency and intensity. We took the slope of these quantities to represent the effective Fano factor for the neuron in this population. These Fano factors were measured under both, light-off and light-on conditions, and the probability distribution is plotted for both conditions in Figure 17, panel **A-C**. None of the optogenetic manipulations made any significant differences to the Fano factor distributions. The mean Fano factor was about 1.2, which suggests that Poisson variability is a good approximation for this neural population. We also tested that none of the optogenetic manipulations had any systematic effects on the Fano factors measured (PV-Chr2: $t_{335} = .4, p = .69$; PV-Arch: $t_{89} = .92, p = .36$; Pyr-ChR2: $t_{133} = -.2, p = .84$).

Additionally, some studies have shown that neurons that are more active tend also to have higher variability [35]. This is relevant because a neuron with a larger Fano factor and all other response parameters the same has relatively less Fisher information (see

Figure 17**G**), and neurons with the highest firing rates tend to contribute most significantly to the Fisher information from the population. A bias towards over-representing the information contribution of these neurons could lead to a systematic error in measurement of thresholds. Using the generalized Poisson model to include the measured Fano factors into the Fisher information calculation leads to a different set of threshold predictions that are quite similar. A plot comparing the neurometric thresholds computed using the generalized Poisson model against the thresholds found using the standard Poisson model is found in Figure 17**H**. The threshold values change very little. It is worth noting that all of the thresholds using the generalized Poisson model increased. This is guaranteed because the generalized Poisson model is only a well-defined probability distribution for variance-to-mean ratios greater than 1. It has been observed that most cortical neurons have a Fano factor greater than 1, and other models that attempt to take into account this increased variability, such as the negative binomial distribution, only allow for Fano factors greater than 1, as well. We therefore set any Fano factors measured to be less than 1 equal to 1 for the purpose of this calculation. Analogous plots to Figure N using the generalized Poisson model are not reproduced here due to redundancy—they are difficult to distinguish visually and correlation coefficients and p values differ by less than 3%.

Correlation coefficients were also measured in both optogenetic conditions. Once again, due to limited samples for any specific stimulus input, we computed the correlation in two steps. and it was observed that they do not change significantly between light-on and light-off conditions (plotted in Figure 17**D-F**). First, we computed a reduced measure of deviation from the mean for each neuron:

$$s_i^k(f,d) = \frac{r_i^k(f,d) - \overline{r_i}(f,d)}{\sqrt{F_i \overline{r_i}(f,d)}}$$

where $k$ denotes the repetition number (1-5), $i$ denotes the neuron, $f$ denotes the frequency, $d$ denotes the intensity, $r$ denotes the evoked response, $\bar{r}$ denotes the average firing rate of a neuron to a particular frequency and intensity, and $F$ is the measured Fano factor for that neuron. This reduced measure is useful because, for a generalized Poisson process, it has zero mean and unit variance (because variance is proportional to the mean). The correlation between the neurons is computed by

$$C_{i,j} = \langle s_i^k(f,d) s_j^k(f,d) \rangle_{k,f,d}$$

The probability distribution of this correlation is plotted for light-off and light-on conditions with each optogenetic manipulation in Figure 17, panels **D-F**. We observed that correlations had a significant, non-zero mean ($\bar{C}_{PV-ChR2} = .09$, $t_{1937} = 28$, $p = 4.6 * 10^{-141}$; $\bar{C}_{PV-Arch} = .13$, $t_{524} = 22$, $p = 2.2 * 10^{-76}$; $\bar{C}_{Pyr-ChR2} = .13$, $t_{982} = 32$, $p = 1.1 * 10^{-155}$). The distributions, however, had no systematic changes under the influence of optogenetic manipulations (paired *t* test ns: PV-ChR2 $t_{1937} = .26$, $p = .80$; PV-Arch $t_{524} = -1.3$, $p = .18$; Pyr-ChR2 $t_{982} = -1.7$, $p = .09$). Similar models attempting to assess the effect of correlations on discrimination threshold have found that they lead to small increases in the discrimination threshold computed from the population [24]. Between the lack of a systematic effect from optogenetics and the small effect observed previously, it is unlikely that changes in the correlations account for the differences in threshold changes when manipulating cortex.

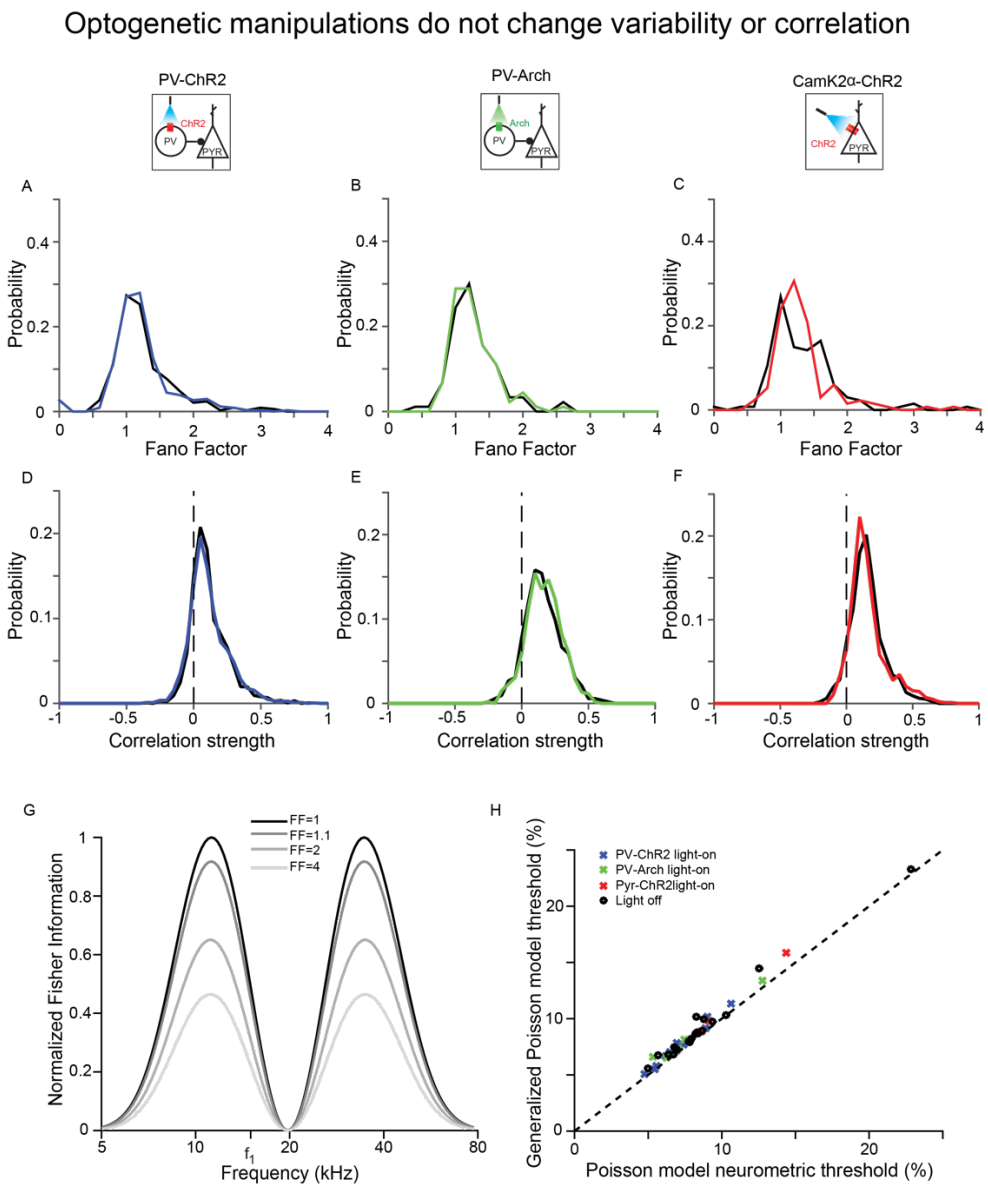


Figure 17: Optogenetic manipulations do not change neural variability or correlation. In panels **A**-**C**, we see the probability distributions of neural Fano factor measured in light-on (colored curves) and light-off (black curves) conditions. Fano factors have no systematic change from any optogenetic manipulation. Panels **D**-**F** show the correlation strength distribution measured in light-on and light-off conditions. Correlations tend to be slightly positive, and exhibit no systematic change under optogenetic manipulations. In panel **G**, we see the Fisher information for a single neuron decreases as Fano factor increases (amplitude = 8 spikes/s, center frequency 20kHZ, tuning width = 0.2 decades, baseline=spikes/s). Panel **H** compares thresholds computed with the generalized Poisson model to thresholds computed with a standard Poisson model, and demonstrates the additional variability parameter makes only a small difference to any threshold prediction.

*Discussion*

Here we have used Fisher information to quantify the discriminability in auditory cortex. Despite having too few neural units to accurately predict absolute behavioral discriminability, the change in behavior under the influence of optogenetic manipulations correlated well with the predicted change from neural computation. This is the first direct prediction about behavioral frequency discrimination acuity based on decoding a neural population in auditory cortex, though similar techniques have been used for studying sound localization [25]. The result suggests that there is a relevant measure for behavioral performance, namely frequency decoding threshold. This can be contrasted with the null hypothesis that AC is simply a part of the circuit responsible for processing spectral information, and therefore any manipulation can change the discrimination acuity of the animal in an unprincipled manner.

Our results also have important implications for the role of inhibitory neurons in the context of frequency encoding. It has been proposed that a number of important tuning properties of excitatory neurons are shaped by inhibition, including tuning width, response variability, magnitude of response, and strength of correlations between neurons [36] [37] [38]. We manipulated PV interneurons, the most common type of interneuron accounting for 40% of interneurons in cortex [12]. We observed no systematic changes in reliability (measured here as the Fano factor) or correlations between neurons as we manipulated the activity of interneurons. We observed changes in the response strength and tuning width for some units, which has been previously reported [3]. Despite the consistency in many of these observations, these same manipulations sometimes evoked opposite behavioral effects in different animals. It is possible that differences in

inhibitory properties within AC may contribute to differences in auditory behavior of the animals.

Another interesting observation was the slope of the index of change plot is only about .25, while true optimal Bayesian decoding from AC neurons would allow decoding at exactly the limits placed by a Fisher information, and therefore predict a slope of exactly 1. This does not strictly preclude decoding from AC consistent with an optimal Bayesian decoder because of the duration of the optogenetic manipulation. It is possible that achieving an optimal decoding scheme requires learning and utilizes plasticity on longer timescales than the optogenetic manipulation is applied. It is also possible that decoding is not optimal, and that another type of decoding is utilized that does not optimize information use. This is a very interesting question and will be accessible when the measured population sizes increase, as this will allow prediction of absolute frequency discrimination. Our recordings had between 10 and 100 frequency-tuned neurons per animal, and extrapolating from the measured population indicates that ~1000 neurons are typically required to explain in order to account for behavioral discrimination acuity. Since the mouse cortex has ~$10^5$ neurons/$mm^3$, the AC is ~1 $mm^3$ in size, about 30-50% of neurons are frequency tuned, and the tuning width is of order ~1/10 of the auditory spectrum, anatomically we would estimate order $10^3$ neurons responding to any given tone. Discrepancies in absolute predicted and measured thresholds will be revealing about whether or not the animals are able to discriminate at the limit established by neural responses.

While many optogenetic studies emphasize the power of manipulating a specific type of neuron in order to trace and understand its role in cortex, here we emphasize a

different perspective. We seek to test a general theory about stimulus encoding in AC that depends on the state of neural circuitry during stimulus exposure, and we utilize the optogenetic manipulation as a way of altering the state of the neural circuitry while keeping other elements the same (including the physical neural network and the same subset of neurons sampled). This would be impossible to compare across animals because there is no one-to-one mapping between neurons for animals as complex as mice or humans. By changing the state of the auditory cortex while controlling other elements, we are able to test our hypothesis about AC function within an animal, despite having too few neurons to make a prediction about absolute thresholds. This perspective shows the utility of optogenetic techniques in providing robust, controlled tests of any model relating cortical activity to behavior.

The circuitry within auditory cortex had unique responses to the optogenetic manipulations, which is demonstrated by the differences in behavioral effects between individuals and the differences in neurometric predictions between animals. Had we combined our results across individuals of a fixed manipulation, we would have seen small average effects and viewed variability across individuals as noise. This would have obscured the role of the auditory cortex in frequency discrimination because the correlations between individual circuit changes and individual behavioral changes would be missed. It because we treated the mice as individuals and tested a hypothesis that applies generically to neural responses under any optogenetic manipulation that we were able to observe the general role AC plays in behavioral frequency discrimination. Treating differences between individuals as a signal rather than as noise will become

more important as we are able to measure from larger neural populations, and therefore probe more explicitly the role cortex plays in shaping behavior.

### *Toward understanding plastic changes in an environment with costs*

Emotional and task-specific learning has been shown to cause changes in the spectral representations in auditory cortex [30] [31] [32] [33] [34]. It has also been shown to cause changes in behavioral frequency discrimination that can be altered with optogenetic manipulation [3] [39]. Asking whether or not changes in spectral representation within auditory cortex explain, by themselves, the difference in behavioral frequency discrimination acuity is natural follow-up question to these observations and a natural extension of the methods used here. This could be done using a similar strategy by recording responses from neurons before and after fear conditioning. Looking to see whether there an analogous correlation between change in neurometric threshold and change in behavioral threshold before and after fear conditioning would tell us whether the new thresholds are predicted from the change in neural responses alone. However, there are good reasons why this might not be the case.

Mice that are fear conditioned by applying footshock during presentation of a specific tone elicit freezing responses when tones are presented, even when those tones are well above their frequency discrimination threshold [3] [39]. This behavior is not entirely surprising, given that it is better to err on the side of caution, but it demonstrates that accounting for animal behavior requires more than simply establishing the limits on sensory system performance. In fact, limits on the sensory system can still serve to

constrain the performance, even when the behavior sits in the more complicated context of a cost landscape.

Let us consider the output of a sensory system trying to estimate some parameter of the environment associated with an appetitive or aversive stimulus. For example, the auditory system estimates the frequency of a tone to decide whether the tone it heard indicates that footshock is incoming ($f_+$) or not ($f_-$). The sensory system provides an estimate, $\hat{f}$, of the frequency, and the probability of that estimator differs for the two tones( $P(\hat{f}|f_+) \neq P(\hat{f}|f_-)$ ) or else there is no information provided. If $f_+ > f_-$, a simple decoding scheme is to set a threshold, $f$, and whenever $\hat{f} > f$, the animal freezes. It is useful to define the cumulative distribution function $\Phi_{+/-}(f) = \int_{-\infty}^{f} P(f'|f_{+/-})df'$. If the $+$ and $-$ event occurs with probability $A_+$ and $A_-$, respectively, then the probability of false-positive and false-negatives are given by $P_{-,1}(f) = A_-(1 - \Phi_-(f))$ and $P_{+,0}(f) = A_+\Phi_+(f)$, where we have used 1 and 0 to denote the binary decision of the presence of the aversive stimulus. Since the actual presence of the aversive stimulus is uncontrollable, the only costs associated with this sensory system are with mis-identifying the stimulus. In other words, the costs of misidentification are only relative to correct identification. During normal auditory exposure, it may be that misidentifying a tone as another is symmetric, and so the costs are the same. However, if the presence of one tone signifies an aversive stimulus, then this cost is asymmetric. The total cost is then given by:

$$\langle C \rangle(f) = C_{-,1}P_{-,1}(f) + C_{+,0}P_{+,0}(f) = C_{-,1}A_-\big(1 - \Phi_-(f)\big) + C_{+,0}A_+\Phi_+(f)$$

For simple distributions, we may extremize this quantity by setting the derivative equal to zero, which returns the simple solution $C_{-,1}A_{-}P(f|f_{-}) = C_{+,0}A_{+}P(f|f_{+})$. If these distributions are Gaussian, which is a reasonable approximation given the large number of neurons relevant for decoding in conjunction with the central limit theorem, this can be reduced in terms of the Gaussian parameters:

$$\frac{(f-f_{+})^2}{2\sigma_{+}^2} - \frac{(f-f_{-})^2}{2\sigma_{-}^2} + \log\frac{C_{-,1}A_{-}\sigma_{+}}{C_{+,0}A_{+}\sigma_{-}}$$

Taking the standard deviations to be the same (because the useful insights are easier to glean), this is solved for

$$f^{*} = \frac{f_{+}+f_{-}}{2} + \frac{\sigma^2}{f_{+}-f_{-}}\log\frac{C_{-,1}A_{-}}{C_{+,0}A_{+}}$$

Plots of these quantities numerically solved for $f_{+} = 2$, $f_{-} = 0$, $\sigma = 1$, $A_{-} = A_{+} = 1$, $C_{+,0} = 1$, and $C_{-,1} = 2$ are shown in Figure 18**A**, along with a mutual-information maximizing solution. First, we should observe that the cost is incorporated into the solution in the same way as the prior likelihood of the events occurring. Second, when the system is symmetric, the solution is information maximizing, but any asymmetry in the costs leads to a solution which no longer maximizes information. Third, the correction for asymmetric costs grows with the variance of the sensory signals. This last part is especially important, as it suggests that animals with more reliable sensory systems will be less likely to generalize their fear response. However, sufficiently aversive stimuli can cause *overgeneralization*. Whenever the cost is sufficiently large, the cost-optimizing threshold will be shifted so that *any* stimulus of this type will be treated like the aversive one, a condition with similarities to post-traumatic stress disorder. The

fact that this leads to a *specific* operating point on to optimize this kind of cost implies that there will be a specific false-negative rate associated with the false-positive rate (see Figure 18**B**). If we have a measure of the capacity of the sensory system, we can constrain the false-positive and false-negative rate. If we are able to control the relative costs of false-positives and false-negatives, we can make a prediction of exactly the false-positive and false-negative rate. This could be tested using, for example, a two-alternative forced choice task where correctly guessing one tone leads a larger reward than the other.

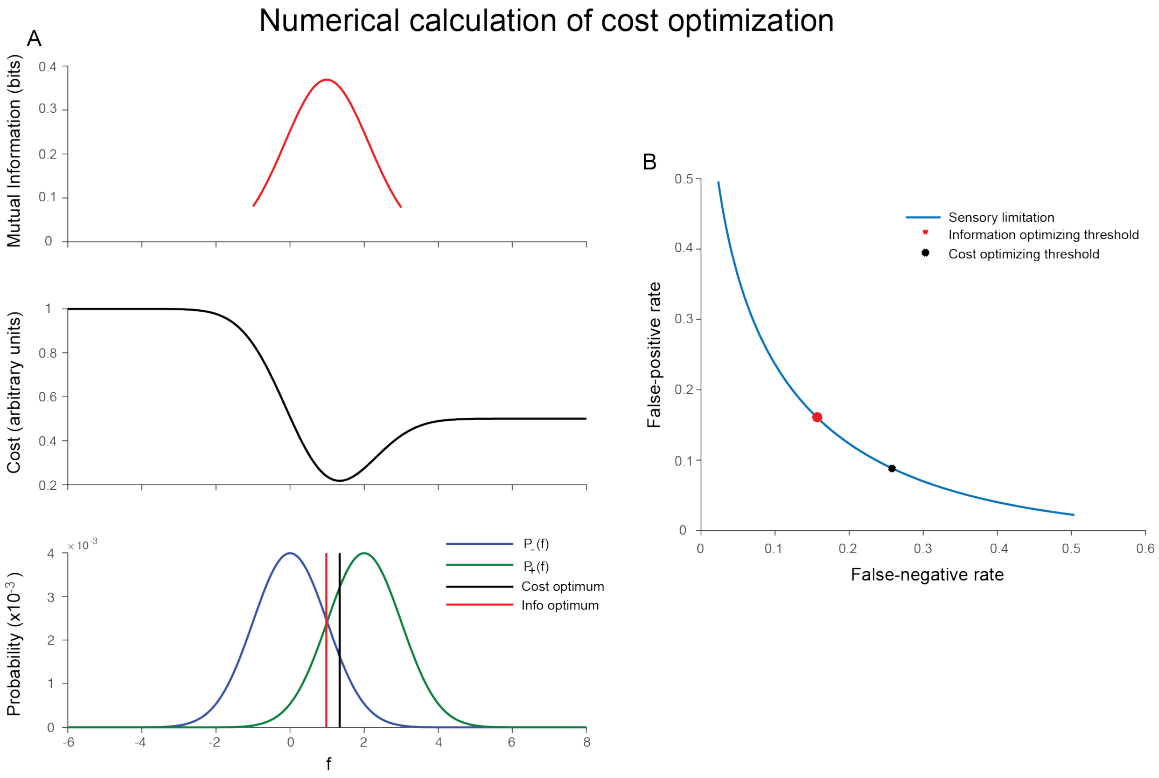


Figure 18: Numerical calculation of cost optimization. In panel **A**, we see the mutual information (top row) and cost (middle row) as a function of the threshold. The probability distributions are plotted along with the information-maximizing and cost-optimizing thresholds in the third row. In panel **B**, we see the curve that limits sensory discrimination performance, and the operating points predicted by information optimization and cost-optimization. Note that the cost-optimizing solution trades off a higher false-negative rate in exchange for a reduced false-positive rate.

We have proposed ways to test the role auditory cortex plays in the frequency discrimination changes associated with neural plasticity in a more complex environment. Whether or not the auditory cortex can fully explain these using the methods established here is an interesting question in its own right, but having a measure of the discrimination ability of the sensory system will allow a detailed prediction of false-positive vs false-negative rates when the animal is required to act in a complex environment with multiple

different costs. As we are able to access larger populations of these neurons, we will get closer and closer to unraveling the mysteries relating neural activity to behavior.

# 4. Learning features through neural input

## *Cortical coding uses only neural inputs*

So far, we have seen a number of theories that predict features of neural representation based on the input stimuli, and we have simultaneously improved our understanding of neural organization and behavior. However, *real* cortical inputs are not the stimuli, but rather neural responses from the preceding sensory neurons. An unanswered question is how neural circuits should organize in order to accommodate potentially diverse inputs. It may even be the case that many sensory cortices perform the same or similar procedures for representing their inputs, and the nature of the input layer. Some studies have suggested that some straightforward learning rules can lead to information maximizing and ICA-like representations of inputs [40]. However, it is not clear how representations in subsequent layers could progress if this type of learning would apply, as the inputs of the next layer would already be independent of one another, and no new information would be gained.

Tools within the machine learning community have shown promising and intriguing results over the course of the past 20 years in problems such as image recognition and speech recognition, and many of the techniques are inspired by biological neural networks. One advantage of some such approaches is that they intrinsically scale well to large input sizes. Computing full probability distributions in other traditional neural network models, such as the Ising model, tend to scale poorly, as there are $2^N$ states if there are N neurons. Deep Belief Networks implicitly create a generative model of the data that may be efficiently sampled without strictly calculating the probability of

each state separately. This is a feature, as most specific patterns in a real neural system will never be observed in the lifetime of the organism. Naturally, these neural-inspired models have also been used to model real biological neural networks. In one study [41], a two-layer sparse deep belief network is trained on image patches from the van Hateren image database and they examine the resulting filters. In the first layer, filters look similar to V1 responses, and they show features learned by a second layer and claim similarities to V2 responses. We draw inspiration from these analyses, but instead turn our attention to ask what happens when the input to the system is more realistic—inputs from the retina itself. Given a different input, there is no guarantee that the same types of filters would be learned. The work in this chapter is unpublished. We will first talk about how we model the retinal responses that enter the deep belief network. We will then go on to discuss the details of what a Restricted Boltzmann machine (RBM) and Deep Belief Network (DBN) are, and how they are trained. We will then examine the filters that occur as a result of performing this training procedure on our simulated retinal responses. We will then discuss the implications this analysis has for real neural networks, and explain how future work utilizing this procedure can test the importance of subtle features of the neural code (such as the role of correlations or real neural variability).

### *Modeling retinal ganglion cell outputs*

We use modeled retinal data, but we have designed the rest of the analysis so that the methods can be repeated using real retinal data. For preliminary analysis, using a retinal model gives us more control over what relevant features are included. To model the output of the retina, we use a common, simplified model of retinal outputs. We use a

linear-nonlinear model containing independent spatial and temporal kernels. The spatial structure is given by a difference of Gaussians, providing the prototypical center-surround structure of retinal receptive fields. We take the surround size to be 3 times larger than the center. The center Gaussian was set to have a standard deviation of 5 pixels for the purpose of convolving with natural movies in order to avoid sampling artifacts from smaller receptive field sizes. Receptive fields were arranged on a grid with separation equal to the standard deviation of the center. The grid was a 25 neuron square, totaling 625 neuron center-positions. We include a population with two neural populations: on-center and off-surround, and off-center and on-surround. This brings the total number of "neurons" to 1250. On/off and off/on cells were arranged with the same center locations and receptive field sizes. The temporal kernel used is biphasic, and given by the equation

$$K(t,\alpha) = \alpha * e^{-\alpha t}\left(\frac{(\alpha t)^5}{5!} - \frac{(\alpha t)^7}{7!}\right)$$

and we take $\alpha = 1/15\ ms^{-1}$. This kernel has positive contributions at small times, and negative contributions at large times. The kernel is approximately negligible at times more distant than 250ms. Non-linearities were chosen so that there were approximately 1-10 spikes/s, typical firing rates for retinal ganglion cells.

Receptive fields were convolved with natural movies provided by Stephanie Palmer of the University of Chicago, and included a movie of a butterfly flying, a tree blowing in the wind, and fly larvae wriggling, combining to be equivalent to hours of neural responses, sampled at ~60Hz.

*Restricted Boltzmann machines and Deep Belief Networks*

In this section, we will cover some basics about Restricted Boltzmann machines and Deep Belief Networks. More information and very helpful tutorials containing everything from practical uses to detailed mathematical explanations of the techniques can be found at Geoffrey Hinton's website, www.cs.toronto.edu/~hinton/ . Restricted Boltzmann machines can be used as a generative model for a vector input. Consider two layers of nodes, an input layer that corresponds to the data to be modeled with responses denoted $\vec{v}$ with size $n$, and a separate hidden layer that will be used to model the input layer with activity denoted $\vec{h}$ with size $m$. For the purposes of our uses with neural networks, we will consider the case where the activity of these nodes are binary. Each element will contain its own bias to fire, which we will call $\vec{a}$ for the visible units and $\vec{b}$ for hidden units. The connection weights $W_{ij}$ will be allowed only between individual elements of $\vec{v}$ and $\vec{h}$, but importantly not within a single layer. The probability of a particular observed state is given by:

$$P(\vec{v},\vec{h}) = \frac{1}{Z}\exp\left[\vec{a}\cdot\vec{v} + \vec{b}\cdot\vec{h} + \vec{v}^{t}W\vec{h}\right]$$

This is illustrated in Figure 19.

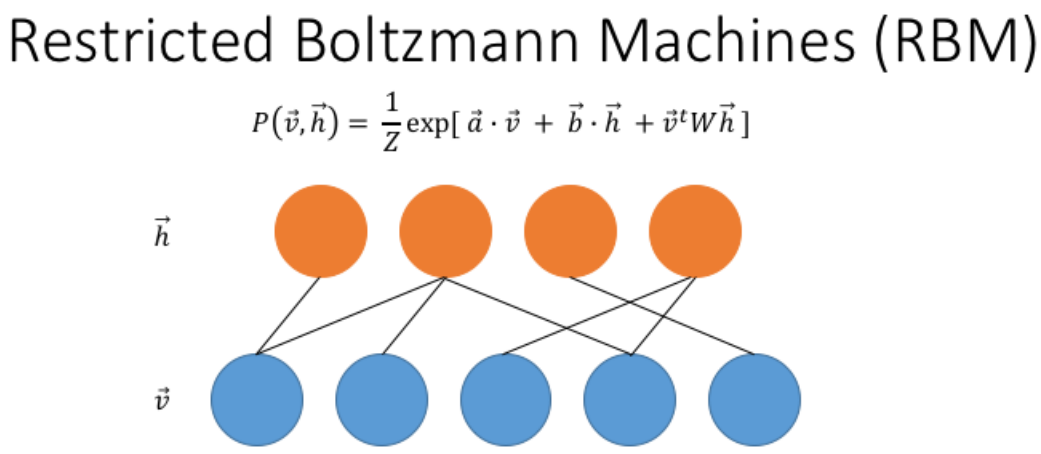


Figure 19: Restricted Boltzmann Machines schematic. A typical connection topology has connections between the hidden units and visible units, but no connections within layers.

The technique we use to train the hidden layer in order to be able to provide a useful generative model of the data is called *contrastive divergence*, and was developed by Geoffrey Hinton [42]. It relies on being able to calculate the marginal probability distributions $P(\vec{v}|\vec{h})$ and $P(\vec{h}|\vec{v})$, which can be expressed:

$$P(\vec{v}|\vec{h}) = \prod_i P(v_i|\vec{h}) = \prod_i \sigma\left(a_i v_i + v_i \sum_j W_{i,j} h_j\right)$$

$$P(\vec{h}|\vec{v}) = \prod_j P(h_j|\vec{v}) = \prod_j \sigma\left(b_j h_j + h_j \sum_i v_i W_{i,j}\right)$$

which relies on the fact that no intra-layer connections exist. This is important, as it allows the calculation to be performed by computing only $2n$ and $2m$ probabilities, instead of having to compute the full $2^n$ and $2^m$ probabilities required to describe a generic joint distribution. The basic method for training with contrastive divergence is

carried out by taking a raw data sample (corresponding, in our case, to the retinal response) , $\vec{v}^{(0)}$, and select a hidden layer response, $\vec{h}^{(0)}$, by sampling from the hidden layer response probabilities implied by the visible state. Then, select a visible layer response, $\vec{v}^{(1)}$, by sampling from the visible layer response probabilities implied by $\vec{h}^{(0)}$. Repeat this once more to compute $\vec{h}^{(1)}$. The weights will then be updated by:

$$\Delta W_{i,j} = \epsilon\left(\langle v_i^{(0)} h_j^{(0)}\rangle - \langle v_i^{(1)} h_j^{(1)}\rangle\right)$$

$$\Delta a_i = \epsilon\left(\langle v_i^{(0)}\rangle - \langle v_i^{(1)}\rangle\right)$$

$$\Delta b_j = \epsilon(\langle h_j^{(0)}\rangle - \langle h_j^{(1)}\rangle)$$

where the expectation values are usually taken over small batches of data samples to stabilize the gradient calculation. This Markov sampling procedure approximates a gradient descent algorithm that tries to maximize the average log-probability of generating a sample from the original training set, and in principle needs to be repeated for several steps (rather than just 1) until the final states are decorrelated from the initial states (though in most cases, taking 1 step is sufficient). Gibbs sampling can be used to generate fake visible layer responses by simply treating $\vec{v}^{(n)}$ as an "observation".

Deep Belief Networks (DBNs) have multiple hidden layers, and hidden layer $k$ serves as the hidden layer of an RBM to layer $k-1$, and the input layer to hidden layer $k+1$. These are typically trained sequentially by first extensively training layer 2 on responses by layer 1 responses (the visible layer/data), then using a set of layer 2 responses sampled from the visible layer responses as the "data" for training layer 3. This process is repeated for each layer, from which a trained DBN is formed. Having multiple

layers allows for more complex representations of the data to emerge. Not only does this allow for better representation of data, but it allows more complex features to emerge.

We trained our deep belief networks using fake retinal data and using a modified training algorithm that additionally encourages sparse activation of the hidden layer network, again to be consistent with realistic neural firing rates. We used several different hidden layer architectures to examine the resulting spatial receptive fields developed in higher layers. The spatial receptive fields in higher layers were calculated by convolving the connection weights with lower layers with the lower level's spatial receptive field. The input layer had 1250 units, as previously described. We tested a compressive architecture containing 600 units in layer 2 and 150 units in layer 3, an expansive architecture containing 250 units in layer 2 and 500 units in layer 3, and an equal architecture containing 300 units in layer 2 and 300 units in layer 3.

### *Emergent representations in DBNs*

Broadly speaking, each network architecture and each layer developed spatial receptive fields that fell into one of three categories, with a few exemplars depicted in Figure 20. The first category describes spatial responses that are similar to Gabor filters, which are traditionally associated with neural responses in V1. These constitute approximately 10-25% of observed hidden layer filters. The second category contains neural responses that are diffusely responsive in alternating sign to large regions of the image. This kind of receptive field is similar to the principal components computed from a natural image ensemble [43], and accounts for approximately 10-25% of observed filters. The third type of receptive field classification contains receptive fields that look

similar to the center-surround structure of the individual receptive fields of the fake retinal responses, but have a larger associated length scale. These are most common in highly compressive stages, indicating that there may have been redundancy in representing the responses of nearby fake neural output. By placing a large compressive constraint, it became efficient to represent the data in a manner that pools spatially localized responses with similar sign.

Surprisingly, the various architectures had little effect on the representations used. There may have been changes in the relative frequency of each type of component, but the differences were relatively small. This may be in part due to the fact that, because of data limitations, it was necessary to include an initially compressive step in the DBN architecture.

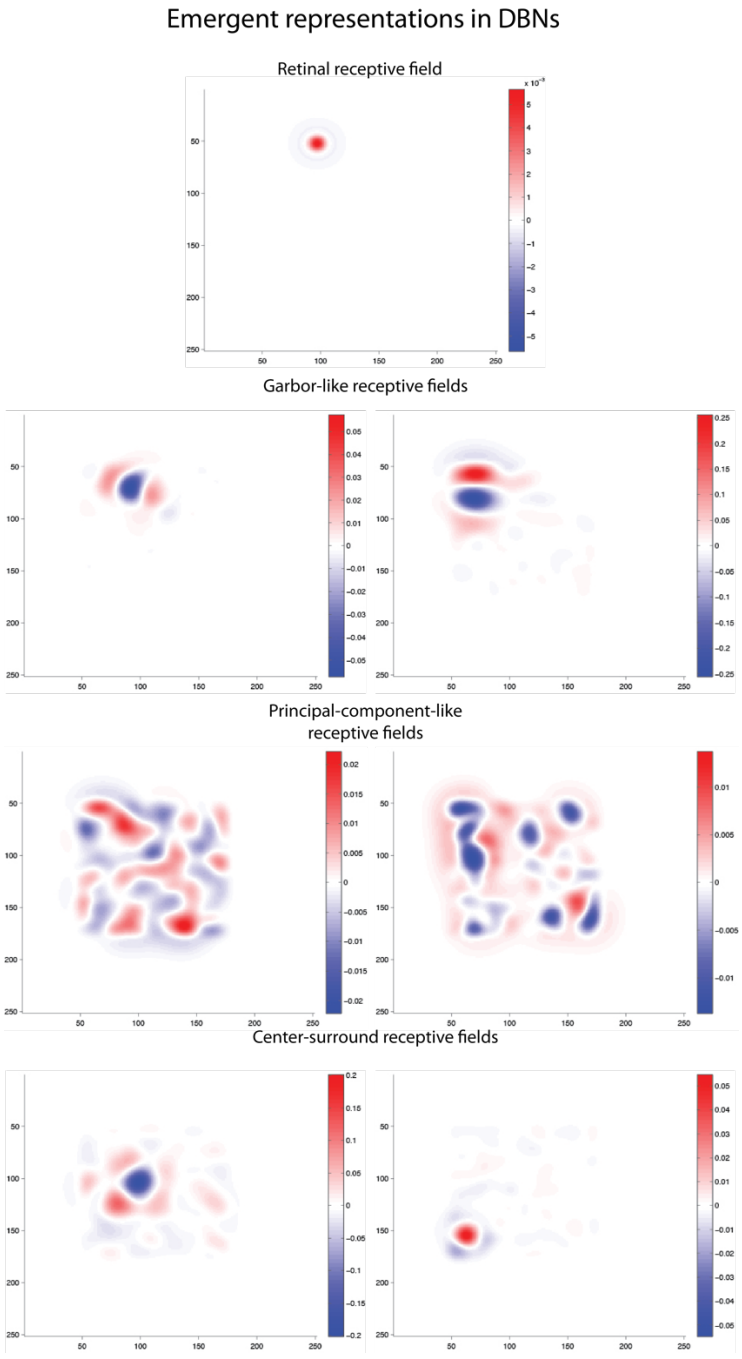


Figure 20: Emergent representations of visual stimuli in DBNs. Training DBNs on fake retinal outputs leads to hidden units with spatial receptive fields that can typically be characterized as one of three types. Some form Gabor-like filters that are similar to protoypical V1 responses. Others form diffuse receptive fields pooling from a large area of the visual scene that are reminiscent of the principle components of natural scenes. The third category has a similar center-surround structure to the original retinal responses, but covers a larger area.

*Discussion*

Here we have seen that training a Deep Belief Network on data approximating retinal output naturally leads to features that have been previously established as important for representing visual scenes. One class of responses, Gabor filters, are prototypical responses of neurons in V1, and have been previously shown to emerge from using ICA on natural images. Other more diffuse filters emerge as well, and are similar in structure to the principal components describing natural images [43]. These features occur naturally when scale invariance is present in the natural world, and here we observed that these features are recoverable after retinal filtering of a dynamic environment.

Although the larger receptive fields containing a similar center-surround structure to the original retinal input may be expected from the presence of a compressive stage, it very well may persist when compression is not a necessary first step. With larger datasets, we will be able to probe this question more deeply. Assuming that they continue to exist for DBNs without compressive projections, another important question would arise. Do these arise from training on an actual dataset of *real* retinal responses? One important feature our retinal model lacked was any kind of correlation between neurons. If these receptive fields cease to appear when the same DBN is trained on real retinal responses, it is likely that the correlations between neurons in real retinal output serve to reduce this particular type of redundancy.

A number of other studies have found various algorithms that lead to efficient representations of natural scenes recover similar visual features to the early visual

system, but here we wanted to ask a different question. We wanted to address whether or not learning based on input from other neural networks can explain the features that visual cortex attempts to represent. This question is important for generalizing coding strategies employed by the visual system to the rest of the brain. Part of the reason the visual system is naturally tractable is that it is relatively easy to characterize the responses of neurons in terms of the stimulus. Many other parts of the brain, such as those that deal with cost or planning [1], receive inputs that are much more challenging to characterize using our anthropocentric perspective of the world. However, this does not mean that the strategies employed to efficiently represent our environment differ dramatically (and if they did, it would be of great interest!). This exploratory study is an example of the sort of thinking that can help connect research in superficially distinct brain regions, and moving forward, could be quite impactful for all of neuroscience.

# 5. Modeling adaptive activity of cortical networks

*Cortical network dynamics*

Cortical networks have very interesting dynamic properties required for the kinds of stable activity patterns they exhibit. Balancing the inputs from excitatory and inhibitory neurons, for example, is of great interest for network stability. It has been observed that pharmacologically blocking inhibitory activity leads to epileptic activity in cortex [44]. It has also been observed that, in auditory cortex, the contribution of inhibitory inputs to pyramidal neurons almost exactly matches the excitatory inputs with a millisecond time delay, resulting in a temporal sharpening of the response [45]. Carefully modeling these phenomena can lead to important insights about the network structure and distill the essential components of what simple components allow the network to display the properties it has. In our lab, we have observed a number of interesting phenomena ranging from neural tone responses to differential responses based on how often a stimulus is presented. These experimental observations allowed us to, through the use of computational modeling, gain insight about underlying network parameters. More specifically, we will first present the basic equations that we use to describe the network activity that we observed experimentally. We will then move on to discuss the implications of experimental measurements due to optogenetic manipulation of PV interneurons and pyramidal neurons during tone response [3], and utilize these insights to model the observed data. We will then discuss results from a stimulus-specific adaptation experiment, in which different neural responses are observed in a stimulus that plays two tones regularly, tone A 80% of the time and tone B 20% of the time. When the tones are switched in proportion of presentations, it is observed that neural responses to

tone A are greater when it is infrequent than when it is frequent. This can contribute to novelty detection in the environment—an important cortical computation of obvious behavioral relevance. We will discuss and model experimental observations [4] that show the different role interneuronal subtypes contribute to this computation.

### *Wilson-Cowan dynamics model*

In order to simplify the activity of the network, we will approximate the population response of each neuronal subtype (Pyramidal, PV+, and SOM) using Wilson-Cowan dynamics. We allow the connections between the population of PVs and pyramidal neurons, and between the SOMs and pyramidal neurons in order to model the effects of each optogenetic manipulation. The equations describing the dynamics of the populations are:

$$\frac{dE}{dt} = \frac{1}{\tau_E}\left[-E(t) + (k-r)S\left(j_{light-E}(t) + j_{tone-E}(t) + S_{inh}\left(j_{IE}I(t)\right)\right)\right]$$

$$\frac{dI}{dt} = \frac{1}{\tau_I}\left[-I(t) + (k-r)S\left(j_{light-I}(t) + j_{tone-I}(t) + j_{EI}E(t)\right)\right]$$

where $E(t)$ represents the activity of the excitatory population, $I(t)$ represents the activity of the inhibitory population, $\tau_X$ are the synaptic timescales of excitatory and inhibitory networks (we take both to be 10ms), $k$ and $r$ correspond to the maximum and minimum "firing rates" (15 and 1, respectively), $j_{light-X}$ is used to model the optogenetic inputs and vary according to the experiment, $j_{tone-X}$ is used to model the inputs the neuron receives due to hearing the tone and varies according to the experiment, $j_{IE}$ and $j_{EI}$ are the synaptic transmission coefficients between excitatory and inhibitory populations, $S(x)$ is the transfer function between synaptic inputs and neural firing rate,

which for our purposes is linear for intermediate input values, but imposes minimum and maximum activation limits, and $S_{inh}(x)$ is a non-linear transfer function describing the input the excitatory population receives as a function of the firing rate of the inhibitory population. We will use several candidate non-linearities, and discuss them more below.

It will be useful to discuss a simple model of synaptic depletion for two reasons: (1) it provides a simple mechanism and mathematical description for a non-linearity in the transfer function discussed above and (2) in the case of stimulus specific adaptation, the inputs to auditory cortex are reduced with successive tone presentations and we need to simulate them dynamically. If the synapses have some finite resources, for example neurotransmitter, which is depleted at a rate proportional to the activity of the neuron and replenished at a rate proportional to how depleted the resource is, we may write

$$\frac{dg}{dt} = -\frac{gr}{T_d} + \frac{(g_0 - g)}{T_r}$$

where $r$ is the presynaptic firing rate, $g$ is the synaptic conductance, $g_0$ is the maximum conductance, and $T_d$ and $T_r$ are, respectively, the time constants for depletion and replenishment. The post-synaptic current is then given by the product, $gr$. The quasistatic solution can be written

$$gr = \frac{g_0 r}{1 + (T_r/T_d) * r}$$

which corresponds to a saturating non-linearity. There is a maximum output rate ($g_0 T_d/T_r$), and the second derivative is negative for positive firing rates.

We will also find it useful to compare facilitation to depression. The difference between the two is that the former has a *positive* second derivative for some region of

activation. In order to compare these two, we will use a sigmoid to model facilitation, and a hyperbolic tangent (which is a qualitatively similar function, but contains only the region with the *negative* second derivative) to model depression:

$$S_{fac}(r) = \frac{1}{1 + \exp[-p(r - \theta)]}$$

$$S_{dep}(r) = \frac{1 - \exp[-2r/s]}{1 + \exp[-2r/s]}$$

We don't worry about the fact that facilitation model does not have zero output for zero firing rates because this scenario is not realized in the data we model.

### *Modeling the change in tone-evoked responses to optogenetics*

As an important control for a variety of behavioral tasks, including emotional learning [3], it is important to understand the state of the excitatory-inhibitory network within the auditory cortex when the animal is exposed to tones. In order to do this we examined the tone-evoked responses of neurons in AC in the presence of three key optogenetic manipulations: activating PV interneurons with ChR2, suppressing PV interneurons with Arch, and activating Pyramidal neurons directly with ChR2. Because pyramidal neurons are more common in auditory cortex, results presented here primarily capture effects observed from them. Experimental results [3], shown below, indicating that manipulating PV interneurons changed the tone-evoked responses, measured as the difference between baseline firing rate and tone-evoked firing rate, but manipulating the pyramidal neurons directly did not. More specifically, optogenetically activating PV interneurons (Figure 21, panel **A**-**B**) increased the tone-evoked response, while optogenetically suppressing them (Figure 21, panel **C**-**D**) decreased the tone-evoked

response. Direct activation of pyramidal neurons (Figure 21, panel **E-F**) increased firing rate approximately the same amount under baseline and tone-evoked conditions, leading to no significant change in the tone-evoked response.

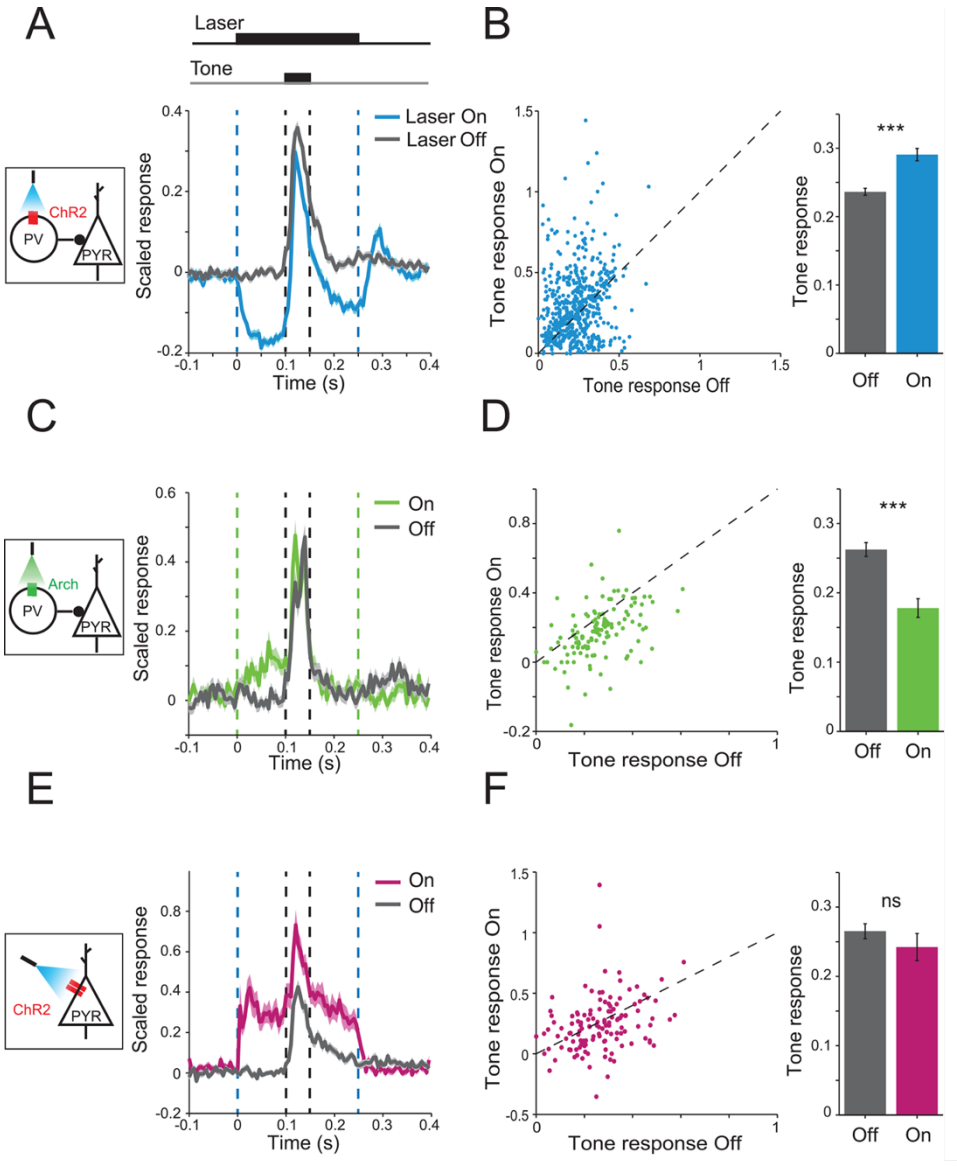


Figure 21: Measuring effects of optogenetic manipulations on tone-evoked responses. In panel **A**, we see that activating PV interneurons reduces the activity of Pyramidal neurons under baseline and tone-evoked conditions, but the change is smaller during the tone-evoked epoch. This is quantified in **B** by computing the difference in activity of the tone due to the background in both, the light-on and light-off conditions. The opposite results are found for suppressing PV interneurons (panels **C-D**)—that while there is generally an

increase in activity due to the manipulation, the tone-evoked response is reduced. Direct activation of Pyramidal neurons led to no significant change in the tone-evoked response (panels **E-F**).

The first significant piece of the model is that excitatory neurons themselves appear to exhibit the same tone-evoked responses even during optogenetic manipulations. The simplest explanation for this is that there is a linear response to the inputs. To understand the inhibitory manipulation, first recall the optogenetic manipulation evokes a smaller change in tone-evoked activity than baseline activity. Also, optogenetic manipulations effect a smaller change in the neural activity than the tone does (and can be thought of as a perturbation of the normal activity). The simplest model has symmetric inputs to both, the inhibitory and excitatory population (and there is evidence that many PVs have similar tuning properties to pyramidal neurons [45]). This suggests that the input from PV interneurons to pyramidal neurons is less affected by the optogenetic manipulation when the the PV neurons are most active (during tone presentation), which is a hallmark of a saturating non-linearity. We therefore decided to use the biologically-inspired quasistatic nonlinearity. We modeled tone input using a decaying exponential. The results of this model under the right choice of parameters [3] are presented in Figure 22.

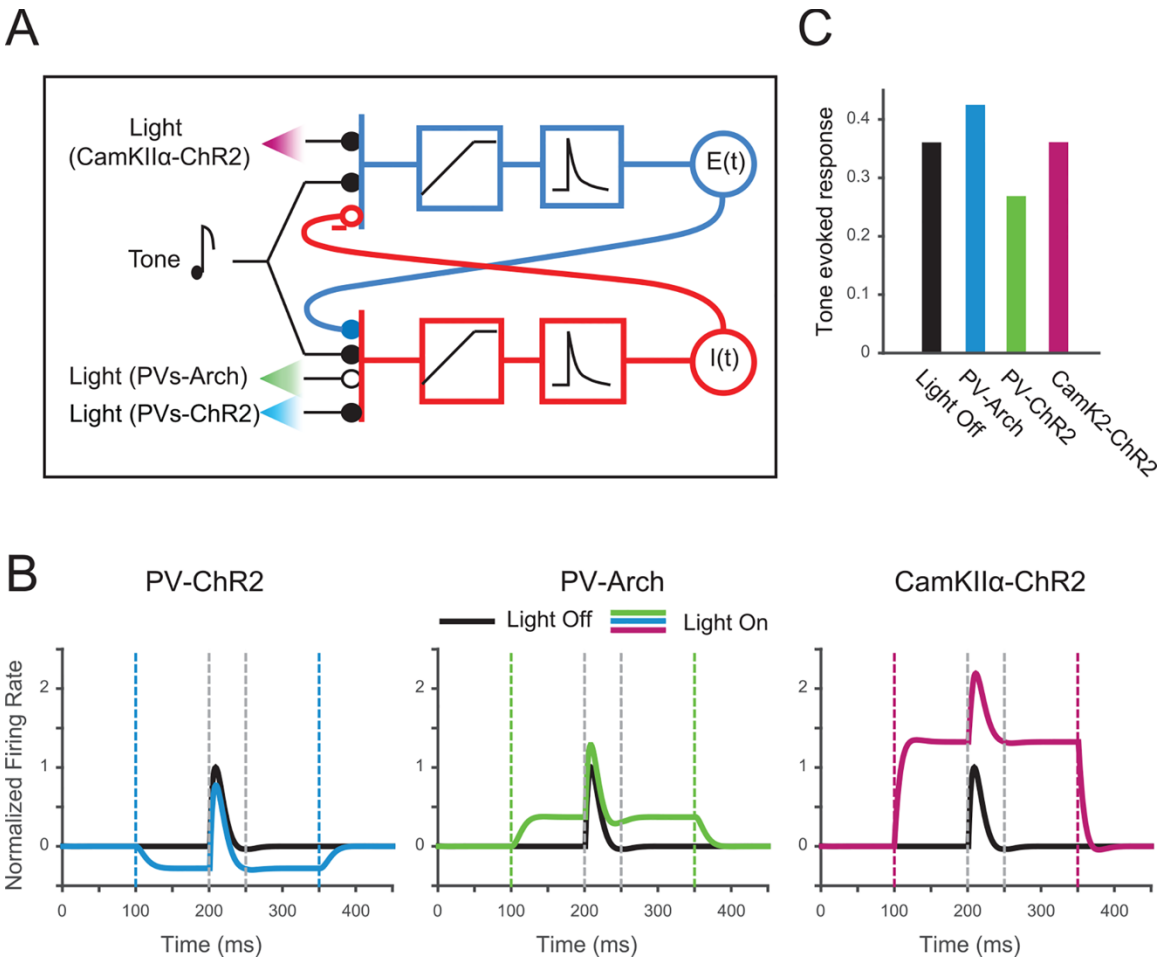


Figure 22: Modeling effects of optogenetic manipulations on tone-evoked responses. The model used is illustrated in panel **A**, which has tone inputs to both populations, as well as currents from the optogenetic manipulations. Inputs are summed and a non-linearity is applied. We observed that no non-linearity is necessary to account for the observations from activating pyramidal neurons, and a saturating non-linearity is the simplest way to account for observations for manipulating PV interneurons. Tone evoked firing rate traces are plotted in panel **B**. Tone-evoked responses are measured and plotted in panel **C**, and are observed to be consistent with the experimental findings.

*Modeling Stimulus Specific Adaptation*

A more sophisticated phenomenon we examined in detail is stimulus-specific adaptation. This phenomenon is inherently dynamic, as subsequent presentations of a stimulus reduce the neural response relative to its novel presentation. This kind of neural computation is more prominent in cortex than in earlier parts of the sensory periphery, and we want to understand how computations in cortex could contribute to the development of this representation. When a stimulus is presented frequently (standard), the neural response is smaller than when the stimulus is presented infrequently (deviant). Experiments [4] show that suppressing PV interneurons during standard tone presentation increases the tone-evoked activity by about the same amount as during the deviant tone, and this is more than the baseline increase (Figure 23, panels **A**-**C**). On the other hand, suppressing SOM interneuron activity affects the spontaneous activity and standard-tone-evoked activity the same amount, but causes no change in the response to deviant tones (Figure 23, panels **D-F**). This is of particular interest, because it suggests that SOM interneurons may contribute directly to the differential response to standard and deviant, while the PV interneurons may play a role more similar to gain control of the overall circuit. We will therefore try to understand what kind of circuit level mechanisms can explain these different phenomena.

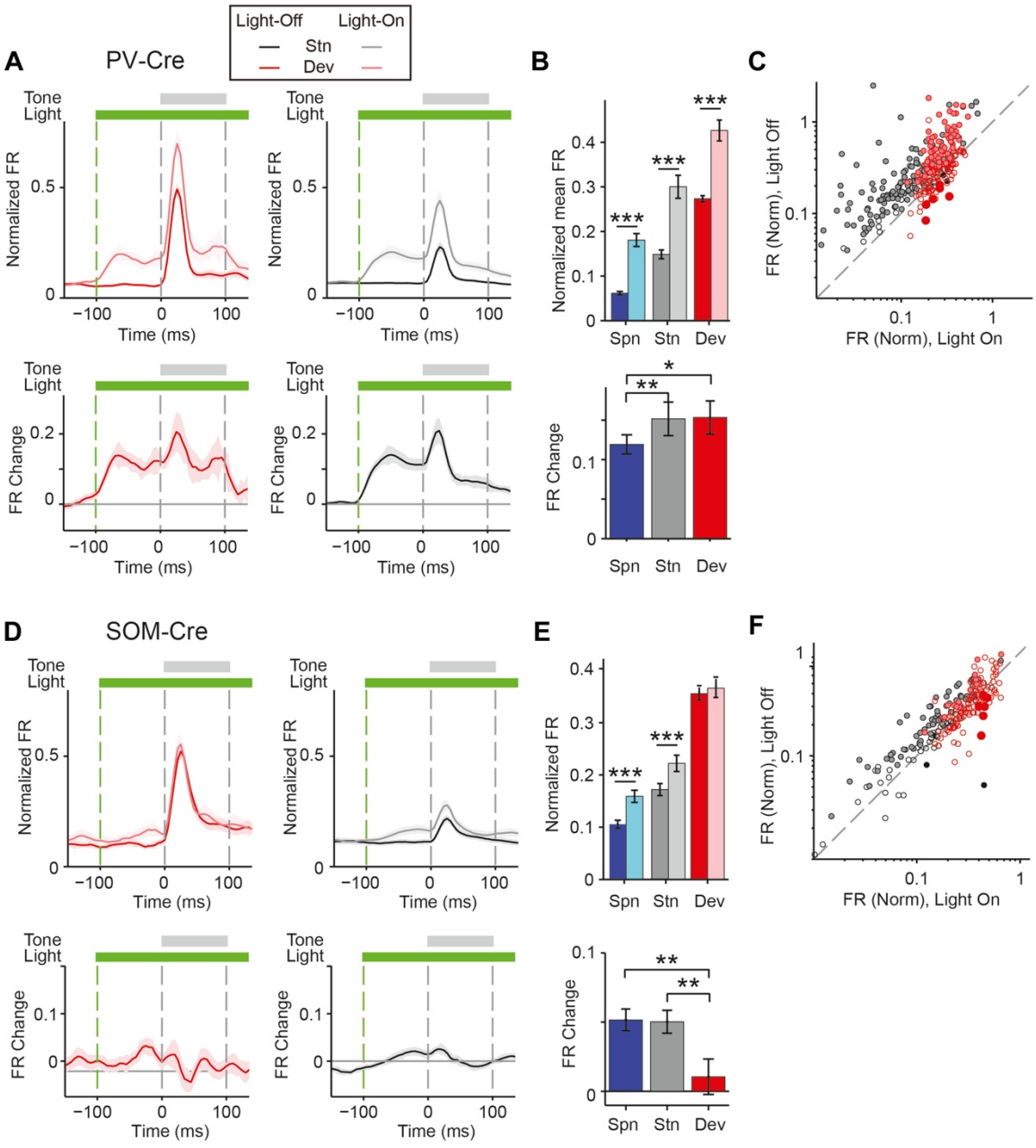


Figure 23: Measuring neural responses to standard and deviant tones. Optogenetically suppressing PV interneurons (panel **A**) leads to changes in the neural activity that differ for spontaneous response, standard response, and deviant response. For PV suppression, we see a larger increase in tone response than spontaneous activity (panel **B**), but the same change when the tone is the standard or the deviant. Individual neurons plotted in panel **C**. When suppressing SOM interneurons (panel **D**), we observe significant changes in spontaneous activity and activity in response to the standard tone, but not to the deviant tone (panel **E**). Individual points are plotted in panel **F**.

In order to model the difference in inputs to the standard and deviant tones, we pass unitary pulses through a depressing synapse with the full dynamic synaptic depression equation described above. The deviant response is calculated using the first input, while the standard response is calculated once the response stops changing with subsequent presentations. Because suppression of PV interneurons has the same effect on pyramidal neuron response to both, the standard and deviant tone, a simple explanation is that the tone-evoked responses lie in a linear portion of the PV-Exc transfer function. The reduced change to baseline activity suggests that for low response levels, its contribution is actually *increasing*. This suggests a facilitating non-linearity. Because suppression of SOM interneurons does not appear to affect deviant responses, a simple explanation is that the neurons have already saturated their capacity to influence the excitatory neurons. The equal magnitude effect on standard activity and baseline activity suggests that the neuron may be operating in a linear regime at these response levels. For these reasons, we will model the PV population's transfer function using a facilitating non-linearity (a sigmoid), and the SOM population using a depressing non-linearity (hyperbolic tangent) (see Figure 24, panel **A**). This model produces similar results to what are observed experimentally (see Figure 24, panels **B**-**C**).

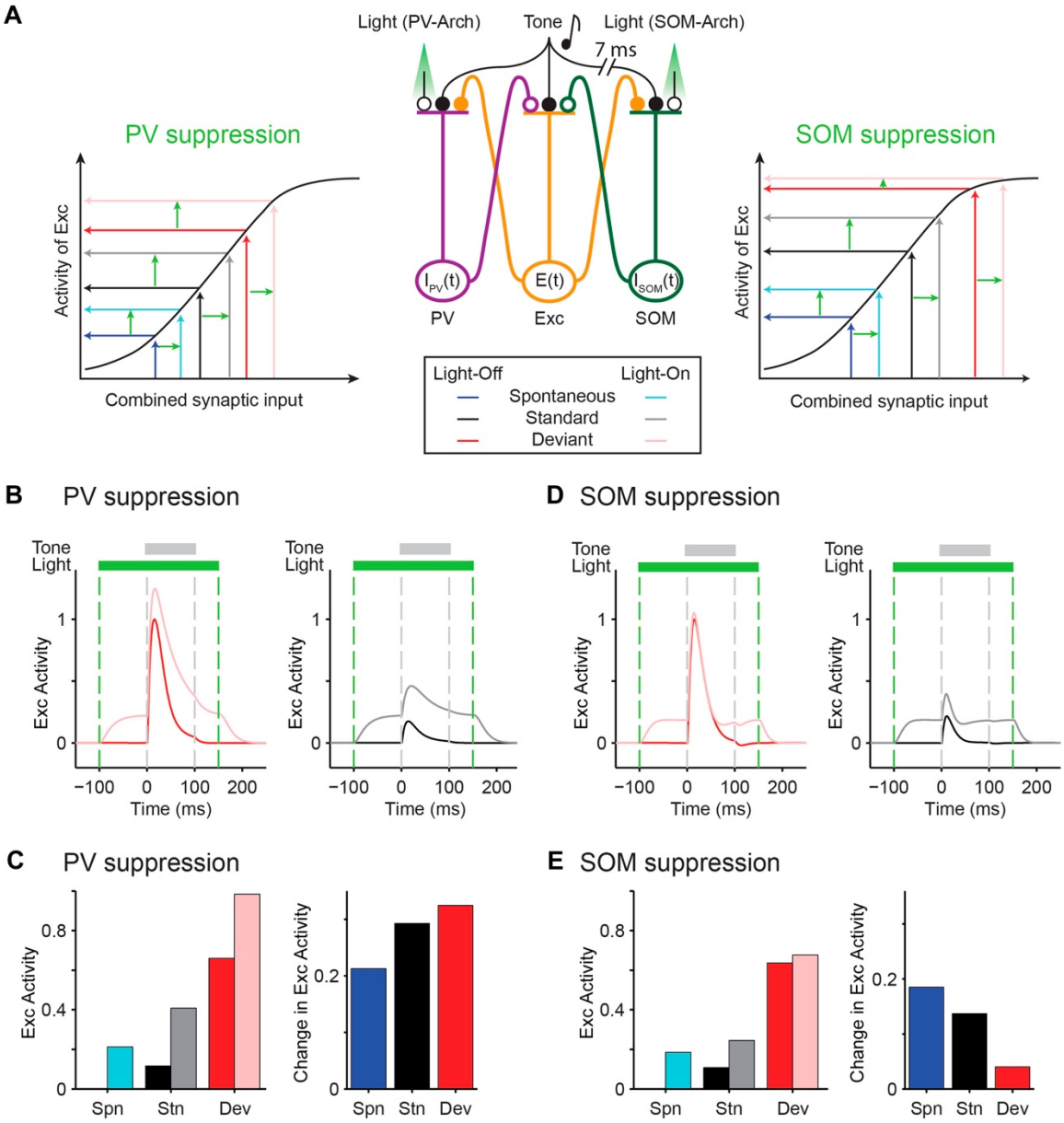


Figure 24: Modeling neural responses to standard and deviant tones. The model, depicted in panel **A**, center, contains excitatory neurons connected to either PV interneurons or SOM interneurons. To explain the PV interaction, we use a facilitating non-linearity, while for SOM interneurons, we use a depressing non-linearity. The tone response curves are plotted in panel **B** for PV interneurons and panel **D** for SOM interneurons, and the hallmark finding about the changes in excitatory population activity is plotted in panel **C** for PV interneurons and panel **E** for SOM interneurons.

## *Discussion*

Here we have seen that neural responses in auditory cortex to even surprisingly complex stimuli can be explained by fairly simple circuit architecture with simple, biologically plausible non-linearities. In one case, we saw that neural tone responses could be modulated by optogenetic manipulations, interestingly in a way that suggests that PV interneurons affect the size of the tone-evoked response of Pyramidal neurons, but not direct manipulation of the pyramidal neurons. We were able to capture these effects by using a model containing a linear contribution of both, the tone and the optogenetic manipulation, but with a non-linearity in the feedback the inhibitory population gives to the excitatory population. In another example, we looked at the contributions different interneurons made based on the frequency of a stimulus. We saw that SOM interneurons made no contribution to the tone-evoked response for rare tones, but a significant one for standard tones. We were able to capture this observation using a saturating non-linearity between SOM interneurons and pyramidal neurons. On the other hand, PV interneurons contributed equally to neural activity in response to both, the standard and deviant tones. This suggested a linear response, but the fact that the contribution to spontaneous activity was smaller implies that the best non-linearity to explain these phenomena was actually a *facilitating* one. We were then able to show using a rates model with Wilson-Cowan dynamics that these simple assumptions can account for the diverse experimentally observed results. It is also quite interesting that such rich adaptive behavior can be accounted for using only simple non-linearities. The model also makes explicit predictions about the activity level of the inhibitory

interneurons themselves—a prediction that can be tested by recording from more of these units directly.

# 6. Conclusions

In this dissertation, we have presented several lines of work that utilize well-formulated theoretical ideas to predict and understand a variety of properties of neural organization. In the second chapter, we saw a formulation of efficient coding was able to predict human sensitivity to visual textures based on natural image statistics. In the third chapter, we saw that the Fisher information in auditory cortex provided a strong indicator of the behavioral performance in a frequency discrimination task. In the fourth chapter, we saw that deep belief networks trained on fake retinal data exhibit cortical-like responses. Chapter 5 showed excitatory-inhibitory network responses in an adaptive environment can be explained with simple network dynamics and a single non-linearity. While each of these lines of work may seems superficially distinct, opportunistic application of theoretical ideas has proven to be fruitful in a field with such a broad scope of fascinating questions.

An important similarity between many of the ideas presented here is that they are inherently forward looking. Though examining fine detail of individual cell responses has led to many landmark results in neuroscience, as the recorded population sizes continue to grow, we need to approach data analysis in new ways. In chapter 2, we saw that the efficient coding principle can be applied to understand many facets of behavioral response, even when the information is guaranteed to be spread across many neurons. In chapter 3, we showed how a model that contains only a few dozen neurons could accurately predict how a mouse's frequency discrimination performance would change under optogenetic conditions. With a larger population, we would have been able to test

the *absolute* threshold for performance. In chapter 4, we proposed a model that allows for inputs from thousands of retinal cells, and showed the emergence of cortical-like responses. In each case, growing the neural population size is something that is handled gracefully. This is guaranteed to be important for analyzing future datasets, as neural recordings become possible and larger and larger scales. Instead of being paralyzed by larger data throughput, the predictions we make in these cases would actually be *refined*.

Another interesting commonality these lines of work is that, although by design they avoided having to address specific cortical representation issues, they still make predictions about resource distribution that will be empirically measurable with access to a significant fraction of the population. In the visual texture work, we applied the efficient coding hypothesis to predict the relative sensitivities to a variety of visual signals. The prediction for sensitivity was based on the gain of a filter, which must be encoded using cortical neurons. Although neurons may be responding diffusely to these higher order statistics, by knowing their responses to a variety of these stimuli, we can measure whether the neural population is itself as sensitive as we predict. Techniques presented in chapter 3 to predict cortical sensitivity to tones could be applied to this set of visual texture signals to test whether cortical sensitivities match the observed behavioral ones. This would create a closed-loop explanation, showing that natural image statistics predict the allocation of neural resources, which in turn explain behavioral sensitivity.

If our motivation for studying the brain is to unravel the mysteries of what makes us who we are, it is important to understand how the lessons we have learned can extend to teach us about other parts of the brain. While the primary motivation of studying sensory systems is that they are fundamentally tractable because of the level of control

we have over the inputs, each line of work presented here contributes to this broader goal in a unique way. In chapter 2, we used the efficient coding hypothesis to show not only that an organizing principle traditionally applied to the sensory periphery is useful in understanding cortical organization as well, but that the nature of cortical constraints may differ from those in the sensory periphery. In chapter 3, we used a generic tool to predict behavioral sensitivity based on neural responses that could apply to any sensory brain region. We also proposed future work that would probe how sensory information is deeply tied to behavior in a context-dependent manner. In chapter 4, we used machine learning techniques to understand how a brain region that sees nothing but neural inputs can organize to try to efficiently represent its inputs. In this case, the emergence of familiar receptive fields was of great interest not just as an explanation of observed activity in V1, but because the organization principle used could apply to *any* brain region whose inputs are other neurons (that is, *any* brain region in cortex). In chapter 5, we saw that complex novelty detection mechanisms can arise using very simple, biologically plausible network properties. The power of such simple non-linear transforms should not be underestimated when trying to understand the computation any brain region is responsible for.

As revolutionary new experimental techniques become available to probe larger and larger regions of the brain, we need to be ready with questions to ask and analysis techniques to address them. If we can do this as a field, the curiosities of the brain may cease to be mysteries.

# Bibliography


[1] J. J. Stott and A. D. Redish, "Representations of Value in the Brain: An Embarrassment of Riches?," *PLOS Biology,* 2015.

[2] A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian and G. Tkacik, "Variance predicts salience in central sensory processing," *eLife,* 2014.

[3] M. Aizenberg, L. Mwilambwe-Tshilobo, J. J. Briguglio, R. G. Natan and M. N. Geffen, Bidirectional Regulation of Innate and Learned Behaviors That Rely on Frequency Discrimination by Cortical Inhibitory Circuits, PLOS Biology, 2015.

[4] R. G. Natan, J. J. Briguglio, L. Mwilambwe-Tshilobo, S. Jones, M. Aizenberg, E. M. Goldberg and M. N. Geffen, Complementary control of sensory adaptation by two types of cortical interneurons, eLife, 2015.

[5] H. B. Barlow, "Possible Principles Underlying the Transformations of Sensory Messages," *Sensory Communication,* 1961.

[6] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network,* 2003.

[7] J. J. Atick and A. N. Redlich, "Towards a theory of early visual processing," *Neural Computation,* 1990.

[8] V. Balasubramanian and M. J. Berry, "A test of metabolically efficient coding in the retina," *Network,* 2002.

[9] E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, T. A. Machado, L. H. Jepson, K. Mathieson, D. E. Gunning, A. M. Litke, L. Paninski, E. J. Chichilnisky and E. P. Simoncelli, "Efficient Coding of Spatial Information in the Primate Retina," *Journal of Neuroscience,* 2012.

[10] B. A. Olshausen and D. J. Field, "Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?," *Vision Research,* 1997.

[11] M. S. Lewicki, "Efficient coding of natural sounds," *Nature: Neuroscience,* 2002.

[12] B. Rudy, G. Fishell, S. Lee and J. Hjerling-Leffler, "Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons," *Developmental Neurobiology,* 2011.

[13] L. Fenno, O. Yizhar and K. Deisseroth, "The development and application of optogenetics," *Annual Review of Neuroscience,* 2011.

[14] G. Tkacik, J. S. Prentice, J. D. Victor and V. Balasubramanian, "Local statistics in natural scenes predict the saliency of synthetic textures," *Proceedings of the National Academy of Sciences of USA,* 2010.

[15] J. H. van Hateren, "A theory of maximizing sensory information," *Biological Cybernetics,* 1992.

[16] J. D. Victor and M. M. Conte, "Local image statistics: maximum-entropy constructions and perceptual salience," *Journal of the Optical Society of America,* vol. 29, no. 7, pp. 1313-1345, 2012.

[17] G. Tkacik, P. Garrigan, C. Ratliff, G. Milcinski, J. M. Klein, L. H. Seyfarth, P. Sterling, D. H. Brainard and V. Balasubramanian, "Natural Images from the Birthplace of the Human Eye," *PLOS One,* 2011.

[18] D. L. Ruderman, "Origins of Scaling in Natural Images," *Vision Research,* vol. 37, no. 23, pp. 3385-3398, 1997.

[19] Y. Yu, A. M. Schmid and J. D. Victor, "Visual processing of informative multipoint correlations arises primarily in V2," *eLife,* 2015.

[20] J. D. Victor, D. J. Thengone and M. M. Conte, "Perception of second- and third-order orientation signals and their interactions," *Journal of Vision,* vol. 13, pp. 1-21, 2013.

[21] J. D. Victor, "Isolation of components due to intracortical processing in the visual evoked potential," *Proceedings of the National Academy of Sciences of USA,* vol. 83, pp. 7984-7988, 1986.

[22] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature: Neuroscience,* vol. 14, no. 9, pp. 1195-1201, 2011.

[23] J. D. Victor, S. M. Rizvi and M. M. Conte, "Simple combination rules for sensitivities to image statistics," in *Society for Neuroscience Meeting planner*, San Diego, 2016.

[24] Y. Gu, C. R. Fetsch, B. Adeyemo, G. C. Deangelis and D. E. Angelaki, "Decoding of MSTd population activity accounts for variations in the precision of heading perception," *Neuron,* vol. 66, pp. 596-609, 2010.

[25] A. D. S. Baia, M. W. Spitzer and T. T. Takahashi, "Prediction of auditory spatial acuity from neural images on the owl's auditory space map," *Nature,* vol. 424, pp. 771-774, 2003.

[26] S. K. Talwar and G. L. Gerstein, "Reorganization in awake rat auditory cortex by local stimulation and its effects on frequency-discrimination behavior," *Journal of Neurophysiology,* vol. 86, pp. 1555-1572, 2001.

[27] M. J. Tramo, G. D. Shah and L. D. Braida, "Functional role of auditory cortex in frequency processing and pitch perception," *Journal of Neurophysiology,* vol. 87, pp. 122-139, 2002.

[28] F. W. Ohl, W. Wetzel, T. Wagner, A. Rech and H. Scheich, "Bilateral ablation of auditory cortex in Mongolian gerbil affects discrimination of frequency modulated tones but not of pure tones," *Learning & Memory,* vol. 6, pp. 347-362, 1999.

[29] T. L. Gimenez, M. Lorenc and S. Jamarillo, "Adaptive categorization of sound frequency does not require the auditory cortex in rats," *Journal of Neurophysiology,* vol. 114, pp. 1137-11145, 2015.

[30] R. C. Froemke, I. Carcea, A. J. Barker, K. Yuan, B. A. Seybold, A. O. Martins, N. Zaikia, H. Bernstein, M. Wachs, P. A. Levis, D. B. Polley, M. M. Merzenich and C. E. Schreiner, "Long-term modification of cortical synapses improves sensory perception," *Nature Neuroscience,* vol. 16, pp. 79-88, 2013.

[31] M. P. Kilgard and M. M. Merzenich, "Order-sensitive plasticity in adult primary auditory cortex," *Proceedings of the National Academy of Sciences of the USA,* vol. 99, pp. 3205-3209, 2002.

[32] D. Polley, E. Steinberg and M. Merzenich, "Perceptual learning directs auditory cortical map reorganization through top-down influences," *Journal of Neuroscience,* vol. 67, pp. 1071-1091, 2006.

[33] J. Fritz, S. Shamma, M. Elhilali and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature Neuroscience,* vol. 6, pp. 1216-1223, 2003.

[34] J. Fritz, M. Elhilali and S. Shamma, "Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks," *Journal of Neuroscience,* vol. 25, pp. 7623-7635, 2005.

[35] R. L. Goris, J. A. Movshon and E. P. Simoncelli, "Partitioning neuronal variability," *Nature Neuroscience,* vol. 17, pp. 858-865, 2014.

[36] J. Wang, D. Caspary and R. J. Salvi, "GABA-A antagonist causes dramatic expansion of tuning in primary auditory cortex," *NeuroReport,* vol. 11, pp. 1137-1140, 2000.

[37] J. Wang, S. L. McFadden, D. Caspary and R. Salvi, "Gamma-aminobutyric acid circuits shape response properties of auditory cortex neurons," *Brain Research,* vol. 944, pp. 219-231, 2002.

[38] R. C. Froemke, "Plasticity of cortical excitatory-inhibitory balance," *Annual Review of Neuroscience,* vol. 38, pp. 195-219, 2015.

[39] M. Aizenberg and M. N. Geffen, "Bidirectional effects of auditory aversive learning on sensory acuity are mediated by the auditory cortex," *Nature neuroscience,* vol. 16, pp. 994-996, 2013.

[40] C. Savin, J. Prashant and J. Triesch, "Independent Component Analysis in Spiking Neurons," *PLOS Computational Biology,* 2010.

[41] H. Lee, C. Ekanadham and A. Ng, "Sparse deep belief net model for visual area V2," in *Neural Information Processing Systems 2007*, 2007.

[42] G. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation,* vol. 14, no. 8, pp. 1771-1800, 2002.

[43] A. Hyvarinen, J. Hurri and P. Hoyer, Natural Image Statistics: A Probabilistic Approach to Early Computational Vision, Springer-Verlag, 2009.

[44] M. A. Dichter and G. F. Ayala, "Cellular mechanisms of epilepsy: a status report," *Science,* vol. 237, no. 4811, pp. 157-164, 1987.

[45] M. Wehr and A. Zador, "Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex," *Nature,* vol. 426, pp. 442-446, 2003.