THE UNIVERSITY OF CHICAGO

RETINAL CODING: ENCODING AND DECODING IN NATURAL SCENES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON COMPUTATIONAL NEUROSCIENCE

BY
BENJAMIN HOSHAL

CHICAGO, ILLINOIS
DECEMBER 2023

For my family

If you want to go fast, go alone. If you want to go far, go together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I'd like to thank my family and friends who have supported me throughout this process. At times, the Ph.D can feel inherently isolating. Even so, it is impossible to complete alone. Without the support of those around me, none of this work would have been possible.

I'd like to thank Stephanie Palmer for her mentorship and guidance over the last five years. Throughout this time I've encountered both professional and personal adversity, and Stephanie always approached any problem with an abundance of compassion and kindness. Stephanie always encouraged me to explore a vast number of (often bad) ideas, and casual discussions surrounding these "dreamer" scientific efforts helped hone my own academic vision and curiosity.

Special thanks goes to Caroline Holmes, with whom I've worked closely alongside for the majority of the Ph.D. There are times throughout this process where we saw each other more than our own spouses while trying to hit deadlines or work through bugged models. As such, Caroline has certainly seen my grumpiest, most stubborn side and still decided to continue working with me. This feat alone deserves special recognition.

Lastly, I'd like thank my wife, Charlotte. This, like everything else we've accomplished, has been a team effort. From preparing my food during long meetings so I'd never miss a meal to encouraging me to get up and push forward even on the hardest days, your efforts have allowed us to finally cross this finish line. You always know when to give me a subtle push to help me achieve everything I've set out to do.

ABSTRACT

Everything that the brain sees must first be encoded by the retina, which maintains a reliable representation of the visual world in many different, complex natural scenes while also adapting to stimulus changes. Decomposing the population code into independent and cell-cell interactions reveals how broad scene structure is encoded in the adapted retinal output. By recording from the same retina while presenting many different natural movies, we see that the population structure, characterized by strong interactions, is consistent across both natural and synthetic stimuli. We show that these interaction contribute to encoding scene identity, and demonstrate that leveraging this underlying interaction network improves scene decoding. This population structure likely arises in part from shared bipolar cell input as well as from gap junctions between retinal ganglion cells and amacrine cells. Separately, we use a task-agnostic deep architecture, and encoder-decoder, to model the retinal encoding process and characterize its representation of 'time in the natural scene' in a compressed latent space. In this end-to-end training, an encoder learns a compressed latent representation from the retinal ganglion cell population, while a decoder samples from this latent space to to generate the appropriate future scene frame. By comparing latent representation of retinal activity from three natural movies, we find that the retina has a generalizable encoding for time in natural scenes, and that this encoding can be used to decode future frames with up to 17ms resolution. Lastly, we explore methods to efficiently scale small population models up to a large population using an aggregate approach.

CHAPTER 1

INTRODUCTION

Sensory perception is the foundation with which any organism interacts with its environment. From a plant sensing sunlight and shifting its leaves, to a mouse sensing a looming shadow and darting to avoid a swooping owl, to a baseball player tracking the motion of the ball to make contact with his bat, sensory perception allows organisms to encode various environmental cues and transmit that information to the rest of the brain in order to make decisions. In the visual system, sensory perception begins at the retina. Here, light sensation begins at the photoreceptor layer in the back of the eye and eventually transmits output downstream via the retinal ganglion cells (RGCs).

While the retina is mostly a feedforward network, the transformation of incoming light signals to outputs from retinal ganglion cells is anything but straightforward. The retina is not simply a camera that takes in a copy of the visual scene and sends it to cortex for processing. Instead, the retina performs a number of complex, nonlinear computations critical for perception. Retinal networks are flexible enough to encode a wide variety of complex stimulus features, such as object motion (Lettvin et al. 1959; Ölveczky, Baccus, and Meister 2003), motion reversals (Schwartz, Taylor, et al. 2007; Shah et al. 2020; E. Y. Chen et al. 2014), and omitted/occluded stimuli (Schwartz, Harris, et al. 2007; Ding et al. 2021). These early computations support efficient downstream readout by throwing away redundant information and preserving features that facilitate perception.

Early studies of retinal processing focused on the dynamics of single-cell responses to the presentation of different kinds of visual stimuli. Simple stimuli are shown to the retina, frequently in an *in vitro* preparation, while obtaining electrophysiological/patch-clamp recordings from single, well-isolated cells in order to characterize their spiking/voltage responses as a function of stimulus manipulations. These studies made critical steps facilitating our understanding of early visual perception, including the discovery of the center-surround receptive field (Rodieck and Stone 1965), the classification of many different RGC types (Hochstein

and Shapley 1976), and single cell adaptation to differing stimulus statistics such as contrast (Shapley and Victor 1978). While similar experiments have continued to advance our knowledge of retinal processing, they do not tell the full story of early visual perception.

One limitation of such experiments is the use of simple, easily-paramaterizable stimuli. These stimuli yield to the experimenter a tremendous amount of control for probing what drive single cells' responses, as changing the statistics and structure of the stimulus, and therefore the response of the cell in question, is straightforward. Organisms in the wild, however, do not encounter this kind of simple stimulus in their daily lives. These organisms are evolved to encode natural scenes with varying statistical structure that are characteristic of the spectrum of environments they encounter within their ecological niche. As such, natural scenes have been shown to drive a richer and more reliable code in the brain (Rikhye and Sur 2015; Froudarakis et al. 2014; Hasson, Malach, and Heeger 2010). While these scenes provide a more behaviorally relevant context with which to study visual perception, their use in experiments targeted to understanding the visual code comes with a significant cost. The natural environment has many complex spatio-temporal features that make neural encoding in the wild difficult to quantify and assess. Natural scenes vary in luminance over many orders of magnitude (Rodieck 1998) and variance (Ruderman and Bialek 1994) (Schwartz and Simoncelli 2001), and have complicated temporal and spatial structure (Dong and Atick 1995; Hateren and Ruderman 1998). Thus, probing visual perception with natural scenes comes with a tradeoff: the retinal responses will be more behaviorally-relevant, but at the (potentially steep) cost of losing the fine-tuned experimental control that synthetic stimuli offer.

Single-cell experiments yield in-depth characterizations of individual cells and cell types at the cost of ignoring how each cell in a population interacts. While single cells individually encode stimulus features, it is the response of the retinal population that drives perception. Many fundamental retinal encoding principles require a suite of at least a few retinal cells in consort, including object motion tracking (Lettvin et al. 1959; Ölveczky, Baccus, and

Meister 2003), motion anticipation (Berry et al. 1999), and latency coding (Gollisch and Meister 2008). For these reasons, it is crucial to study retinal encoding both at the single cell and population level.

Significant theoretical work has recently been devoted to studying population level responses (Kastner, Baccus, and Sharpee 2015; Maheswaranathan et al. 2018; Botella-Soler et al. 2018a; Molano-Mazon et al. 2018; Stringer et al. 2019). These works, along with ever improving experimental advances in recording from large number of populations (Marre et al. 2012; Berényi et al. 2014; Lopez et al. 2016; Steinmetz et al. 2021), create an opportunity to move to more complex, dynamic stimuli and analyze the population code in terms of the readout goals of the downstream networks. However, studying increasingly large population sizes requires careful consideration of new limitations. For any given population size N, there are 2^N possible states the population can take. This exponential increase in population states means in any given experiment some number of important encoding states may be undersampled or not present at all. The answer to this problem is **not** simply recording for a longer time. Recording times are limited by the amount of time retinal tissue can be kept alive and healthy *in vitro*, which lasts about four hours. Further, even if large scale advances in recording times solved a sampling problem, the fact remains this fully expressive code is potentially unreadable by downstream brain areas. These problems can be ameliorated by looking not just at the retinal population states, but focusing on the underlying statistics that govern those states. Understanding this latent population structure could provide valuable insights on population level coding principles.

This thesis explores population level retinal encoding and decoding in natural scenes primarily through the lens of these latent underlying principles that drive the population code. In Chapter 2, we use the responses of a salamander retina to repeated natural stimuli to infer a minimal model of the population response structure of its output retinal ganglion cells. Because of the diversity of subjects and locations, the chosen natural scenes are all ecologically relevant to the salamander and exhibit exhibit significantly different statistics.

Thus, playing a variety of scenes in a single experimental session allows for investigation into changes in the population structure across scenes.

We model the population response using maximum entropy models, which have a history of success using $O(N^2)$ parameters to capture the structure in the data, even higher-order features not explicitly constrained by the model (Schneidman et al. 2006; Pillow and Simoncelli 2006; Granot-Atedgi et al. 2013; Ganmor, Segev, and Schneidman 2011; Tkačik et al. 2014; Roudi, Nirenberg, and Latham 2009; Jaynes 1957; Tkačik et al. 2010). Often in neuroscience, maximum entropy models are constrained by the average firing rate of each cell and the correlation of each pair of cells, $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$. We implement a time-dependent variation of these models that is constrained by the time-varying firing rates of each unit in the population averaged across stimulus repetitions, along with pairwise correlations between cells. Using this model, we find a conserved population structure across natural scenes. We demonstrate that this conserved structure carries population level information about large scale scene statistics. Lastly, we uncover the aspects of the retinal network that carry this conserved structure.

In Chapter 3, we continue to explore the effects of the scene-invariant aspects of the population structure on scene decoding. We explore a graph based decoding paradigm that uses only the conserved population structure found in Chapter 2 and single cell PSTHs to obtain embeddings that capture scene identity. We obtain these embeddings using a graph neural network that learns to embed activity vectors via an unsupervised, contrastive learning approach, and takes advantage of pairwise noise correlations between units. This is a proxy for functional connectivity/interactions within the population. We show that the PSTHs and conserved population structure together are sufficient to obtain single-trial embeddings of activity during the presentation of natural scenes that are separable scene-by-scene. Further, the learned embedding space can support zero-shot incorporation of a novel natural scene not encountered during training. This provides evidence that leveraging a consistent, population-level structure, underpinned by consistent pairwise correlations between cells, may be critical

for scene decoding by downstream circuits, especially in a system where single cells are free to adapt to incoming stimuli.

Chapter 4 focuses again on compressing high dimensional retinal population activity into an interpretable latent space with the aim of decoding natural scenes. Here we designed a variational auto-encoder capable of accurately reconstructing future image frames from a low-dimensional bottleneck representation of input population retinal ganglion cell responses to natural scenes. By virtue of supporting high-fidelity reconstruction of future frames, the learned latent space of this network necessarily contains information relevant to scene decoding. We find that the learned compressed representation of the population responses contains features relating to both static and dynamic features of the natural scenes. This learned representation is generalizable to new stimuli 'for free', by virtue of being optimized for an ethologically-relevant/useful objective. Further, these static and dynamic features are synergistic with respect to encoding.

Chapter 5 departs from the direct study of population level retinal coding to demonstrate a promising method for building large N maximum entropy models. Developing approaches to build this kind of model constitute a critical area of need for neuroscience as it enters the age of high-recording throughput. However, as the population size N increases, maximum entropy models consistently underestimate the true entropy and correlations of the data (Macke, Murray, and Latham 2013; Granot-Atedgi et al. 2013). In this chapter, we focus on the strong stability of cell-cell couplings across multiple fits of the same coupling. This stability holds up to some scaling factor even when fitting increasingly large models. Since large N maximum entropy models have $O(N^2)$ parameters, large N models become increasingly difficult to fit well. It may be beneficial to instead fit a large number of smaller N models and use an aggregate approach to build a large N model from a collection of smaller models. We demonstrate, using the retinal ganglion cell dataset, that an aggregate approach to large N maximum entropy modeling yields similar results to fitting the larger model outright. We discuss where this strategy breaks down, and future plans to improve

this aggregate approach by testing on synthetic data.

CHAPTER 2

STIMULUS INVARIANT ASPECTS OF THE RETINAL CODE DRIVE DISCRIMINABILITY OF NATURAL SCENES

2.1 Abstract

Everything that the brain sees must first be encoded by the retina, which maintains a reliable representation of the visual world in many different, complex natural scenes while also adapting to stimulus changes. Decomposing the population code into independent and cell-cell interactions reveals how broad scene structure is encoded in the adapted retinal output. By recording from the same retina while presenting many different natural movies, we see that the population structure, characterized by strong interactions, is consistent across both natural and synthetic stimuli. We show that these interactions contribute to encoding scene identity. We also demonstrate that this structure likely arises in part from shared bipolar cell input as well as from gap junctions between retinal ganglion cells and amacrine cells.

2.2 Introduction

While single cells individually encode specific stimulus features (Horace B Barlow 1953; Hubel and Wiesel 1959; Hartline 1940), it is their aggregate response that drives our perception (Warland, Reinagel, and Meister 1997; Bialek et al. 1989; Gollisch and Meister 2010; Baccus et al. 2008). For this reason, it is important to understand not only how individual cells respond to stimuli, but also how cells influence each other within a population (Brivanlou, Warland, and Meister 1998; Shlens, Rieke, and Chichilnisky 2008; Pillow et al. 2008; Ramirez and Bialek 2021). Significant theoretical work has been devoted to understanding population responses (Kastner, Baccus, and Sharpee 2015; Maheswaranathan et al. 2018; Botella-Soler et al. 2018a; Molano-Mazon et al. 2018; Stringer et al. 2019), in tandem with experimental

innovations in recording from a large number of cells simultaneously (Marre et al. 2012; Berényi et al. 2014; Lopez et al. 2016; Steinmetz et al. 2021). This creates an opportunity to move to more complex, dynamic stimuli and analyze the population code in terms of the readout goals of the downstream networks.

The natural environment has many complex spatio-temporal features that make neural encoding in the wild difficult to quantify and assess. Natural scenes vary in luminance over many orders of magnitude (Rodieck 1998) and variance (Ruderman and Bialek 1994) (Schwartz and Simoncelli 2001), and have complicated temporal and spatial structure (Dong and Atick 1995; Hateren and Ruderman 1998). Visual systems adapt to these changes on many scales in time and space. Neural systems show near-perfect adaptation to these changes (Fairhall et al. 2001), so a question remains about how brains recover scene-specific information once in an adapted state. These complexities and open questions have led many studies to investigate animal behavior in natural settings (Joseph J Atick and A Norman Redlich 1992a; Nemenman et al. 2008; Jovancevic-Misic and Hayhoe 2009; Zimmermann et al. 2018). In this work, we quantify the structure of the neural code at the input end, and how it might support downstream readout that ultimately drives behavior in complex environments.

While natural scenes contain a multitude of higher order statistics, not all features are equally important. Even at the earliest stages of visual processing, the retina performs nonlinear computations to encode essential aspects of the visual scene. Retinal networks are flexible enough to encode a wide variety of complex stimulus features, such as object motion (Lettvin et al. 1959; Ölveczky, Baccus, and Meister 2003), motion reversals (Schwartz, Taylor, et al. 2007; Shah et al. 2020; E. Y. Chen et al. 2014), and omitted stimuli (Schwartz, Harris, et al. 2007). These early computations support efficient downstream readout by throwing away redundant information and preserving features that facilitate perception.

We use the responses of a salamander retina to natural stimuli to infer a minimal model of the population response structure of its output retinal ganglion cells. This structure is

conserved across scenes, and it has functional consequences; it helps the population carry information about large-scale scene statistics. Finally, we show that this functional role requires only a sparse set of connections, and that these sparse couplings appear to arise from both shared input (bipolar and amacrine cells) and direct connections (gap junctions).

2.3 Results

Probing multiple naturalistic, dynamic inputs to the retina

We make dense extracellular recordings from retinal ganglion cells (RGCs) in the larval tiger salamander (Fig. 2.1a) while presenting the retina with 20-second clips from five different movies (see Methods for details). Salamanders undergo metamorphosis, exposing them to both underwater and terrestrial environments while their retinal structure remains largely the same (Wong-Riley 1974; Burkhardt, Fahey, and Sikora 2006). Further, while salamanders are traditionally ambush predators, they still navigate through their environment, generating self-motion. The movies chosen represent a sampling of the wide variety of scenes that occur in the organism's ecological niche (Fig. 2.1b).

Because of the diversity of subjects and locations, the movies exhibit significantly different temporal and spatial correlation (Fig. 2.1d). Further, luminance correlations alone fail to capture behaviorally relevant features like motion, which arise from higher order structure. For example, object tracking in natural scenes reveal different shapes and timescales of the velocity autocorrelations for different scenes (Salibury and Palmer, n.d.). As animals navigate through different environments, they rapidly adapt to these changes in stimulus statistics (Fairhall et al. 2001; Kim and Rieke 2003; Rieke 2001). Thus, playing a variety of scenes in a single experimental session allows for investigation into changes in the population structure across scenes.

Pairwise couplings are consistent across movies

To fully describe a dynamic population code, we must enumerate 2^N possible states at each time point in the response. Even for modest N, a fully expressive code is both exper-

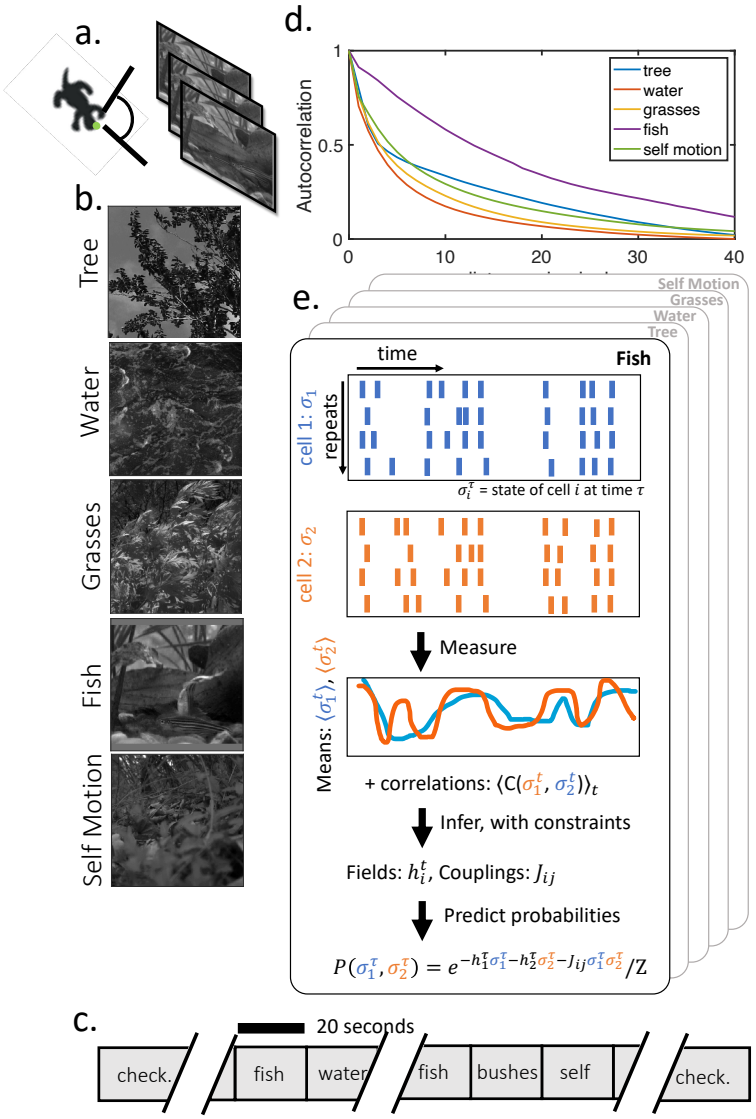

Figure 2.1: **Measuring retinal ganglion cell responses to natural scenes.** (a) Voltage responses were recorded from the retinal ganglion cell layer of a salamander retina stimulated by natural movies. (b) Example frames from each of five natural scenes, which show, respectively, trees blowing in the wind; flowing water; ferns and grasses in a breeze; fish swimming; and woodland underbrush as viewed by a moving camera. The bottom image shows the aggregation of the receptive fields of the recorded population of neurons. (c) In order to probe the statistics of responses, natural scenes were repeated a minimum of 80 times in pseudorandom order. A checkerboard stimulus lasting 25 minutes was shown before and after the natural scenes. (d) The five natural movies used have different statistics, including (shown here) different spatial autocorrelation functions. (e) To model responses to each of the movies, time-dependent maximum entropy models are fit to each of the five natural scenes.

imentally inaccessible and potentially unreadable by downstream networks. To summarize the population code, we use maximum entropy modeling, which has a history of success using $O(N^2)$ parameters to capture the structure in the data, even higher-order features not explicitly constrained by the model (Schneidman et al. 2006; Pillow and Simoncelli 2006; Granot-Atedgi et al. 2013; Ganmor, Segev, and Schneidman 2011; Tkačik et al. 2014; Roudi, Nirenberg, and Latham 2009; Jaynes 1957; Tkačik et al. 2010).

In applications of maximum entropy techniques in neuroscience, these models are constrained by the average firing rate of each cell and the correlation of each pair of cells (Schneidman et al. 2006), $\langle \sigma_i \rangle$ and $\langle \sigma_i \sigma_j \rangle$. We use a time-dependent maximum entropy model (Ferrari et al. 2018) that is also constrained by the time-varying firing rates averaged across repeated stimuli (see Methods for details). Our model takes the form

$$P(\vec{\sigma}^t) = \frac{1}{Z} e^{-\sum_i^N h_i^t \sigma_i^t - \sum_{i<j}^N J_{ij} \sigma_i^t \sigma_j^t}, \tag{2.1}$$

and our constraints are on $\langle \sigma_i^t \rangle_k$, which captures each cell's individual response to the stimulus at time t averaged over trials, k, as well as $\langle \sigma_i^t \sigma_j^t \rangle_{t,k}$, the correlations between cells. These two constraints map to two sets of parameters, the time-dependent fields h_i^t and the static couplings J_{ij}, respectively. Interactions between the time-dependent fields h_i^t absorb any stimulus-dependent correlations, leaving the couplings J_{ij} to capture the noise correlations. Given that our model accurately predicts population activity (Fig. 2.2a), and that the fields are explicitly constrained by stimulus-induced single-cell statistics, we consider the matrices J_{ij} to carry the essence of the intrinsic, non-independent population structure.

In all movies, the cells significantly increase their firing rates in the first 200 ms (Fig. 2.2b) following the switch to a new stimulus. This is followed by a rapid decay back to a baseline firing rate. This is likely due to a strong population response to abrupt changes in luminance within their receptive fields (Fig. 2.2b inset) (Jarsky et al. 2011; Demb 2008; Nakatani and Yau 1988). Despite this substantial adaptation to each movie, we find con-

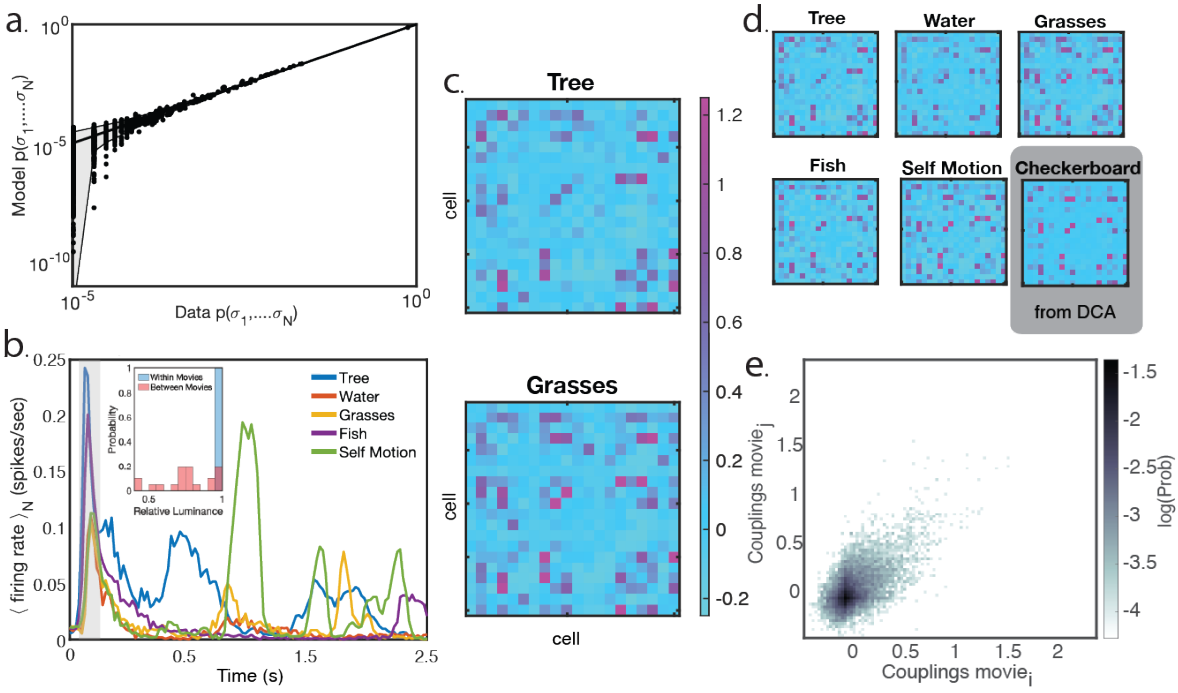

Figure 2.2: **Retinal ganglion cell population maintains consistent couplings across a variety of natural stimuli** (a) Probability of population states, as measured from data and compared to the model. Gray shading indicates expected sampling noise in the estimates of probability from data. (b) Average population firing rate as a function of time for the first two seconds of each natural movie. In the first 200ms (grey bar), the population firing rate peaks. This may be due to a change in luminance within the aggregate population receptive field between stimuli (subplot). (c) Couplings for an example 20 cell group from two of the movies, Tree and Grasses. The structure of the coupling matrices are consistent across scenes. (d) Couplings across all five natural scenes and the checkerboard white noise stimulus. In all cases, couplings are consistent across stimuli. Even for the checkerboard, where another modeling procedure (DCA) had to be used, similar couplings are found. (e) Couplings are quantitatively similar across movies ($R^2 = 0.74$), and for any choice of group of cells; here, we show the values of the couplings J_{ij} for all pairs of movies, for ten different groups of cells.

sistent J_{ij} matrices across movies, indicating that the noise correlations are conserved (Fig. 2.2c,d). Previous work has similarly found consistent couplings across visual inputs (Ferrari et al. 2018; Sorochynskyi et al. 2021; Simmons et al. 2013), but this is the first time this has been demonstrated across a range of naturalistic stimuli. These are entirely independently trained models, which separately learned the same couplings despite significantly varying scene statistics and population responses (Fig. 2.1). This strongly implies that this structure arises from the retina itself, rather than being inherited from the stimulus.

These consistent couplings are not unique to the particular 20-cell group analyzed in Fig. 2.2c and d. For a selection of randomly chosen groups, we plot the coupling J_{ij}^{α} between cells i and j in movie α against the couplings J_{ij}^{β}. We observe a strong correlation between the cell-cell interactions across movies (Fig. 2.2e). These couplings could arise in response to the shared long-timescale and length-scale correlations in natural scenes (Dong and Atick 1995), or they could be an anatomical property of the retina. In order to investigate this, we need to compare the conserved couplings we observed in response to the natural movies to those found in response to an entirely different, non-naturalistic stimulus.

Finally, we investigate the population structure in response to a white-noise checkerboard stimulus. Due to a lack of repeats, we used a different method, Direct Coupling Analysis (DCA) (Weigt et al. 2009), to infer these couplings (see Methods for details). Despite the many differences between this model and those above (stimulus, model details), we extract the same strong couplings as those found in response to the natural scenes. This strengthens the argument that the observed couplings are indicative of real biological interactions, not correlations inherited from the input.

Couplings allow for better decoding of scene identity

What could these static, sparse couplings be used for downstream of the retina? While single cells adapt to switches between scenes after about one second and then fluctuate in response to ongoing dynamics within a scene, the population structure remains constant. Surprisingly, this scene-invariant, static backbone of interactions supports faster and more

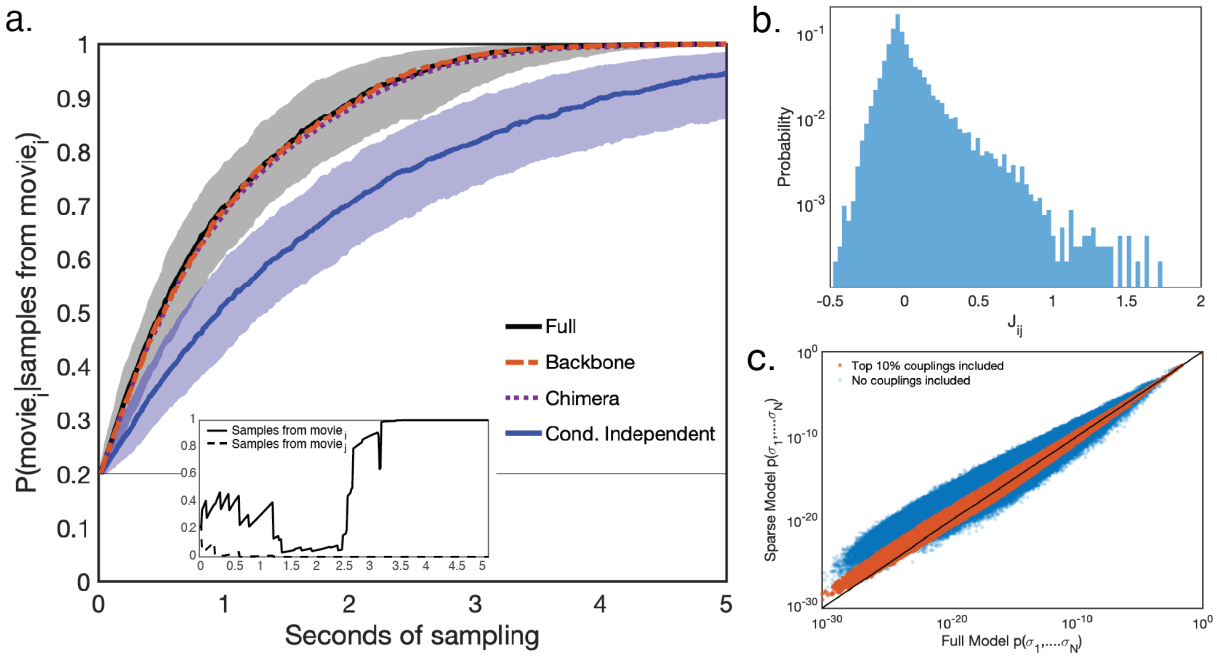

Figure 2.3: **The conserved coupling structure aids in decoding scene identity.** (a) Probability of a given scene, given a particular spike train. The coupling population structure gives dramatically faster scene decoding than an independent model. Additionally, both the chimera model, which swaps J_{ij} matrices, and the sparse backbone model, which preserves only 10% of couplings, carry nearly as much scene information as the full model. Inset: probability of scene given spikes for a single trial of generated samples. (b) The distribution of J_{ij}s across many 20-cell models; we find that this distribution is heavy tailed. (c) Probabilities of population states, for our full model, a model with all J_{ij}s set to zero, and a model with sparsified J_{ij}s. The backbone model captures the predictions of the full model, while the independent model fails spectacularly.

reliable readout of scene identity after this adaptation has occurred.

Interactions between cells in the retinal population have been observed in many other input contexts, and shape individual cell stimulus encoding (Tkačik et al. 2010). The sparse pairwise structure that we observe may combine with individual cell fluctuations to change the overall encoding map between scenes. This interaction may support decoding scene identity. The movies contain a suite of higher order features that make each one readily discriminable to the human eye, but may also impact the local, correlated retinal population code in a decodable way.

To quantitatively test whether the couplings affect discrimination between scenes, we take advantage of the fact that each of our movies comes from a significantly different environment. This means that the information about scene statistics can be approximated by information about scene identity. We quantify the ability of an ideal observer to correctly identify a scene based solely on access to retinal output, given by the posterior $P(\text{scene identity}|\text{spikes})$, as a function of number of samples of the retinal response. This decoding task is similar to the real problem solved by downstream brain areas when an organism moves between scenes as it navigates its natural environment, and must trigger different behaviors and priors in different niches.

The quantity $P(\text{scene identity}|\text{spikes})$ describes how likely any particular scene is given a particular series of spikes. We measure this quantity by generating independent samples from our time-averaged models for each movie, $P(\vec{\sigma}) = \frac{1}{T}\sum_t^T P(\vec{\sigma}^t|t)$, and calculating $P(\text{scene identity}|\text{spikes})$ as a function of the number of generated samples using Bayes' law,

$$P(\text{scene identity}|\text{spikes}) = \frac{P(\text{spikes}|\text{scene identity})P(\text{scene identity})}{P(\text{spikes})}.$$

The terms on the right hand side of this equation can all be generated from our model. Because each sample has a defined length of 1/60 s, we can then convert the number of

samples to a number of seconds of sampling.

By performing this analysis, we find that samples generated from just a 20-cell group carry enough information to correctly identify a scene within a few seconds (Fig. 2.3a). By contrast, spikes generated from a conditionally independent model, which is fit while constraining all J_{ij} to zero, take nearly twice as long to achieve the same scene discriminability. This is surprising, as it was not *a priori* obvious that couplings would contribute to decoding (Averbeck, Latham, and Pouget 2006).

While the population structure remains scene invariant, there are subtle changes in J_{ij} values across movies which might influence the neural code. Conversely, subtle changes in coupling strength might have minimal functional effect so long as the overall population structure remains consistent. To test whether small changes in coupling strength between movies affect scene discriminbability, we implement 'chimera' models. A chimera model for movie α uses the fields from that movie, $h_i^{t,\alpha}$, but replaces the couplings with those learned from a different movie, J_{ij}^{β}. This generates models that maintain the scene invariant coupling structure observed across movies while allowing individual coupling strengths to fluctuate. We find that spike trains generated from these models lead to similarly fast decoding as from the full model. This implies that fluctuations in the coupling values between scenes have little functional impact on scene readout. Instead, the scene-invariant population structure alone drives the improvement in scene discriminability.

We can then investigate whether the entire interacting population structure is important for decoding, or whether the J_{ij} interactions can be sparsified without sacrificing discriminability of scenes. The coupling distribution is heavy-tailed, with a sparse set of strong couplings (Fig. 2.3b) alongside many weak interactions. Previous work has conflicting reports on the relevance of weak couplings, where some show that weak couplings combine to have a large effect on population activity (Schneidman et al. 2006) while others suggest that ignoring weak interactions has minimal effect on population responses (Ganmor, Segev, and Schneidman 2011). To investigate the role of strong couplings, we sparsify the J_{ij} matrix,

leaving only the top 10% of couplings to shape population activity and re-train our model. We fit these 'backbone models' while constraining the weaker 90% of couplings from the full model to be zero. The backbone model is nearly as fast at identifying scenes as the full model, suggesting large-scale scene information is specifically preserved through the sparse, strong couplings rather than the combination of many weak couplings.

Additionally, upon comparing state probabilities, we find that the backbone model makes predictions that are very close to those of the full model, while the independent model fails significantly (Fig. 2.3c). These results suggest that the conserved population structure is dominated by a backbone of sparse strong couplings, and that these couplings play a functional role in preserving scene-level information.

Couplings arise from both gap junctions and shared bipolar cell input

We have found consistent population structure that is dominated by sparse couplings. This structure is hard-wired into the retina code and could arise from many different circuit properties. Pairwise retinal couplings could be correlated with shared upstream input from a bipolar cell (Fig. 2.4a, left) , direct gap junctions between retinal ganglion cells (Fig. 2.4a top), or gap junctions between RGCs and a third neuron such as an amacrine cell (Fig. 2.4a right).

Nonlinear summation of bipolar cell (BC) inputs has been shown to be an integral component of retinal computation (E. Y. Chen et al. 2013; Goldin et al. 2022; Gollisch 2013; Schwartz and Rieke 2011; Fairhall et al. 2006). The convergence of BCs onto RGCs is modeled using nonlinear summation in so-called cascade models. These models explain a wide array of complex retinal computations (e.g., motion onset (E. Y. Chen et al. 2013), omitted stimulus response (Schwartz, Harris, et al. 2007) , background vs object motion (Ölveczky, Baccus, and Meister 2003), reversal response (Schwartz, Taylor, et al. 2007; Shah et al. 2020; E. Y. Chen et al. 2014)). As a complement to that, BCs have diverging projections onto multiple RGCs on the retina (Asari and Meister 2012). While bipolar cells sample a smaller portion of the visual scene than RGCs, gap junctions networks between RGCs can expand

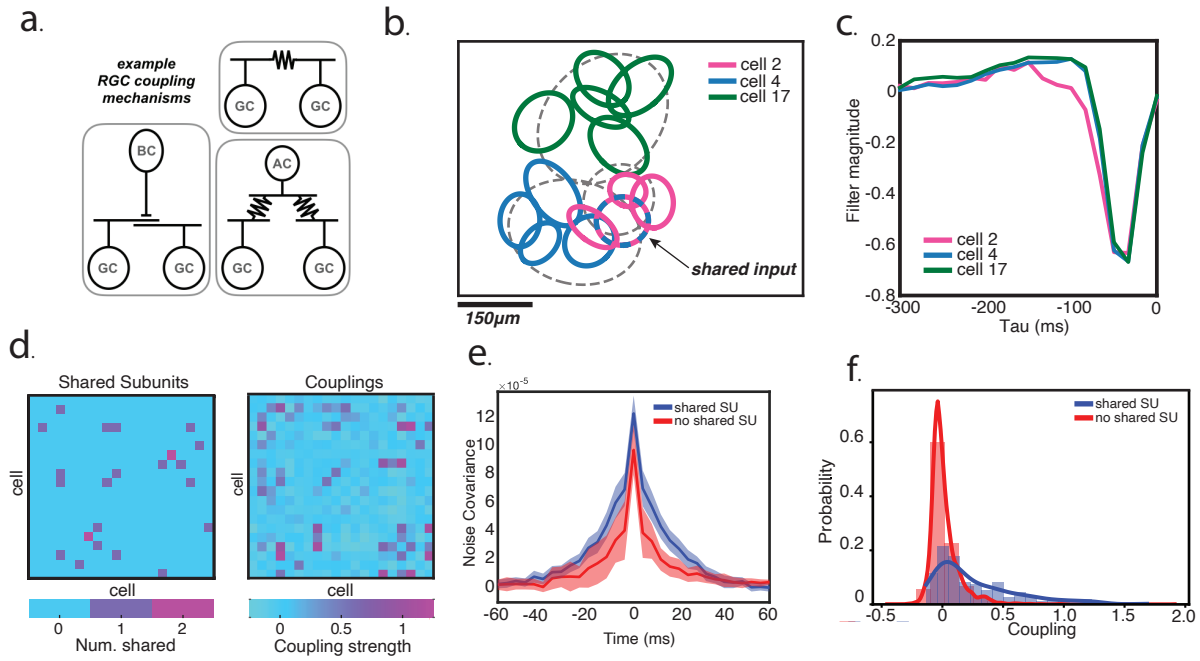

Figure 2.4: **Strong couplings result from both shared inputs and gap junctions.** (a) Three potential pairwise coupling sources. Left, shared input from an upstream bipolar cell. Middle, a direct gap junction between two retinal ganglion cells. Right, gap junction connections to a shared third neuron an amacrine cell. (b) ST-ICA modeled spatial subunits for three cells, two of whom share a subunit. (c) Time filters for the same cells shown in b. (d) The average noise covariance for highly coupled cells. Highly coupled cells that share a subunit have a broader noise covariance in time than coupled cells without a shared subunit. (e) A comparison between shared subunits and couplings indicates many highly coupled cells shared upstream input. (f) Distribution of coupling strengths, both for cells sharing subunits and those that do not.

the bipolar projective field out as far as ∼1mm (Asari and Meister 2014) This BC divergence could play a role in shaping the population code in the retina and needs to be further explored with naturalistic, dynamic inputs.

In order to detect putative bipolar cell inputs onto the RGCs in our dataset, we use spike triggered independent component analysis (ST-ICA) (Saleem, Krapp, and Schultz 2008) on the white noise checkerboard stimulus. ST-ICA models each RGC as the output of a temporal filter and spatial subunits. In a similar method, these subunits have been experimentally shown to map to bipolar cell inputs (J. K. Liu et al. 2017). We show an example of the spatial subunits and temporal filters for three fast-OFF RGCs in Fig. 2.4b and Fig. 2.4c, two pairs of which (blue, pink; green, pink) demonstrate strong couplings. One coupling pair (blue,pink) exhibits a highly overlapped spatial subunit that we classify as a shared upstream input (see Methods). In Fig. 2.4c, temporal filters for two cells (blue, green) show near identical characteristics. This may arise from a gap junction connecting them and might explain their strong coupling (Ferrari et al. 2018).

Across the population, many RGCs share spatial subunits. These subunits closely align with the observed coupling matrix from the stimulus dependent maximum entropy models (Fig. 2.4d), demonstrating that strong couplings may arise in part from shared upstream input. However, not all cells with strong couplings share upstream input, and the presence of a shared subunit alone does not guarantee the existence of a strong coupling. Thus, strong couplings might arise from multiple sources within the retinal cell population.

Previous work has suggested that gap junctions may underpin couplings between RGCs (Brivanlou, Warland, and Meister 1998; Ferrari et al. 2018). Here, we find evidence of gap junctions by inspecting the cross-covariance of responses after subtracting the trial-average (at zero lag this is the usual noise correlation). Pairwise noise correlations due to gap junctions can generally be split into two classes, direct RGC-RGC couplings and shared gap junctions with a third upstream neuron. The symmetric, medium width correlation we observe between some highly coupled cells without a shared subunit (Fig. 2.4d, red) likely arises

from this second class, as direct RGC-RGC gap junctions lead to transient noise correlations at a sub-millsecond timescale (Brivanlou, Warland, and Meister 1998). Furthermore, the broader noise correlations between RGCs with shared subunits demonstrates a timescale longer than can be explained from gap junctions alone and may indicate a coupling arising from shared upstream input (Fig. 2.4a, blue).

Shared input between RGCs, whatever the source, greatly increase the likelihood of a a strong coupling compared to RGC pairs without shared input (Fig. 2.4e). Of course, shared input does not *guarantee* a strong coupling between RGCs, but the long tail of strong coupling for pairs with shared input suggest that these mechanisms underpin our sparse network of strong interactions. We find that shared bipolar input and gap junctions work in consort to generate a sparse set of intrinsic correlations between RGCs.

2.4 Discussion

This work demonstrates that couplings between cells in a neural population are an important component of downstream readout of scene identity. While some studies show that independent models of the retinal code retain upwards of 90% of the response structure (Nirenberg et al. 2001), this is typically agnostic to the downstream readout goals of the organism. Without a defined goal, it is impossible to determine whether aspects of the response structure lost by an independent encoding scheme relay information meaningful to the organism. It is possible that an independent readout preserves the majority of information available in the retinal population while failing to effectively convey critical features of the visual scene. Natural scenes probe a behaviorally relevant context to assess the impact of noise correlations on neural coding. These movies, like other natural inputs, drive a richer and more reliable code in the brain (Rikhye and Sur 2015; Froudarakis et al. 2014; Hasson, Malach, and Heeger 2010). Comparing across movies reveals what the more subtle features in the neural code might be used for.

Our finding that sparse interactions sufficiently capture the functional impact of noise

correlations on neural encoding elaborates on previous work that argued for a dense network of weak couplings in the retinal code (Schneidman et al. 2006; Tkačik et al. 2014). In these early models, both the fields and interactions between cells were static. The fluctuating fields we included (following the time-dependent maximum entropy model work (Ferrari et al. 2018)) absorb much of that structure, and capture the independent component of stimulus-driven changes in the neural population response. The remaining sparse couplings are the key factor for efficient scene identification. A sparse backbone may be easier to implement and read out downstream. On the flip side, sparse codes might hamstring error correction (Puchalla et al. 2005; Ganmor, Segev, and Schneidman 2015), so future work should explore how these costs and benefits trade-off for behaviorally relevant inputs and tasks.

We find that the noise correlations have a large effect on scene decoding, which may arise from small effects aggregated over time. It is not clear from the analysis performed here what precisely gives rise to the beneficial impacts of noise correlations on decoding. One possible answer is that the noise correlations may reflect changes in scene correlation structure. This may help recover scene specific information that is otherwise lost to single-cell-level adaptation.

Unraveling how this sparse but strong structure in the code is mechanistically supported is an important next step in this work. In some ways, the circuit structure in the eye differs from that found in the cortex. The retina is not a recurrent neural network; RGCs do not have direct synaptic coupling, and the photoreceptor-to-RGC circuit is largely feed-forward. To create a population code with sparse interactions, the retina needs to be wired around these structural constraints. These sparse interactions seem to be the result of common bipolar inputs and gap junction coupling between RGCs. What we have observed is sparse, strong, functionally important, exclusively non-synaptic RGC-RGC couplings. Both gap junctions and common bipolar inputs lead to stronger coupling between cells, but our analysis is not sensitive enough to tease apart whether these two types of coupling sources are mutually exclusive. Exclusivity would be an efficient way to implement a sparse backbone of specific

cell-cell interactions. Future work to disentangle the circuit mechanisms giving rise to the sparse backbone might ultimately inform studies in cortex where gap junction coupling is also present (Friend and Gilula 1972; Peinado, Yuste, and Katz 1993; Y. Li et al. 2012).

2.5 Methods

Neural data

Voltage traces from the RGC layer of a larval tiger salamander retina were recorded following the methods outlined in (Marre et al. 2012). In brief, retina from a larval tiger salamander was isolated in darkness and pressed against a 252 channel multielectrode array. Recordings were taken during stimulus presentation and spike sorted using a mostly automated spike sorting algorithm. This technique captured a highly overlapping neural population of 93 cells that fully tiled a region of visual space. Spikes were binned at 60Hz for all analyses presented.

The binned data for both the checkerboard and movie stimuli can be found here: [LINK].

Stimuli during recording

A 30Hz white noise checkerboard stimulus was played for 30 minutes prior to and after the natural scene stimuli.

Five different natural scenes lasting 20s were played in a pseudorandom order, and each were shown a minimum of 80 times. Specifically, they were shown (in order, for the tree, water, grasses, fish, and self motion movies) 83, 80, 84, 91, and 85 times. All natural scenes except for the tree stimulus were displayed at 60Hz. The tree stimulus was shown at 30 Hz and is repeated twice during each 20s epoch.

These movies were collected [insert info on sources], and are distributed alongside the data.

Maximum entropy modeling

We followed the data-driven algorithm introduced in (Ferrari 2016) for our maximum entropy modeling. This data-driven algorithm is a quasi-Newton method that allows for

inference of model parameters $\mathbf{X}$, in our case time dependent fields h_i^t and couplings J_{ij}, without needing to compute the inverse model susceptibility matrix $\chi^{-1}[\mathbf{X}]$ at each time point during the learning dynamics.

As in (Ferrari et al. 2018) we learn a maximum entropy model with time-varying fields. Specifically, we learn a model of form

$$P(\sigma_1^t ... \sigma_N^t) \propto \exp\left(-\sum_i^N h_i^t \sigma_i^t - \sum_{i<j}^N J_{ij} \sigma_i^t \sigma_j^t \right). \tag{2.2}$$

The time dependent fields h_i^t capture the time-varying firing rate of cells σ_i in response to the stimulus. This means that stimulus dependent correlations between cells are absorbed into the fields, leaving the couplings J_{ij} to encapsulate the noise correlations between cells.

For several 20 cell groups, we validated model fits by comparing couplings in the first half and second half of each stimulus and ensuring couplings remained stable. To do this, we separately trained models on each half of the data.

All fits were done on groups of 20 cells, which were all subsets of the full population of 93 cells. These subsets were chosen at random

We additionally use sparse models to generate 2.3. For the independent model, we fit to the same fitting target, but constrained all couplings to be zero. For the backbone model, we first fit a full model with all couplings and fields. Then, we re-fit, constraining all but the top 10% of couplings from the first fitting to be zero.

DCA models

When we were looking for a model that would give us access to the couplings in the white noise stimulus, we were faced with the lack of stimulus repeats. The stimulus-dependant maximum entropy model fundamentally relies on measurements of noise correlations, and as such does not work without stimulus repeats. However, we also wished for a method that would allow us to infer couplings that are only representative of the noise correlations, not couplings that included both shared stimulus inputs and noise correlations. In general, this

is not possible, and standard maximum entropy model are not up to the task.

Here, however, we took advantage of a feature of this data: many of our cells have highly overlapping receptive fields. This means that two cells with similar RFs can be correlated two ways: because of a noise correlation, or because of shared stimulus drive. In particular, if several cells are driven in the same way by the stimulus, they will all have correlations with each other, and therefore many methods would infer 'loops' of couplings.

However, in the protein community, DCA (Direct Coupling Analysis) was developed as a maximum-entropy technique with an emphasis on ignoring indirect couplings (i.e., breaking loops), and a prior asking for a sparse coupling matrix. This means that in our case with several cells all driven in the same way by the stimulus, many of those correlations will be dropped in favor of a sparser explanation for population activity, and in particular one should expect the remaining couplings to be the strongest correlations - those where there is both a biological coupling between the cells and a shared receptive field. This means that we do not expect quantitative agreement between this method and stimulus-dependant maximum entropy, but that we can hope to find the same backbone of strong couplings.

We do not expect this method to be perfect or to be generally applicable. Here, however, the fact that we obtained highly similar couplings to those found from stimulus dependant maximum entropy (a task at which stimulus independent maximum entropy models fail entirely) is proof in and of itself that this method was reasonably successful in our context.

In order to fit the DCA mdoel, we followed methods discussed in (Weigt et al. 2009). We chose a gauge where neural silence/activity are described by {0,1} to more easily relate the DCA network to couplings fit from the time dependent models.

Decoding scene identity

For the scene identity decoding introduced in Fig 2.3, we used a Bayesian approach to measure $P(\text{movie}|\text{spikes})$,

$$P(\text{movie}|\text{spikes}) = \frac{P(\text{spikes}|\text{movie})P(\text{movie})}{P(\text{spikes})}. \tag{2.3}$$

We generate sample spike trains from the learned probability distributions for each of the five movies to test these probabilities. As our models of spike probabilities are equilibrium models, with no relationships between consecutive states, we do this by sampling from the distribution $P(\vec{\sigma}) = \frac{1}{T}\sum_t^T P(\vec{\sigma}^t|t)$ independently for each state in our simulated spike train.

For each sampled spike train, we can calculate $P(\text{spikes}|\text{movie}_\alpha)$, where α indexes movies, simply by plugging into the probability distributions that define our models. We set a uniform prior, $P(\text{movie}) = 1/5$. Finally, $P(\text{spikes}) = \sum_\alpha P(\text{movie}_\alpha)P(\text{spikes}|\text{movie}_\alpha)$.

For each combination of movie and model choice, we generate 1000 spike trains. In Fig 2.3a, we show the median performance for each model at decoding, and the quartiles for the full and independent models.

Spike Triggered ICA

To compute the Spike Triggered ICA (ST-ICA) we follow methods developed in (Saleem, Krapp, and Schultz 2008).

We first compute the spike triggered average for each cell in a natural cubic spline basis. This is a common method to reduce the number of parameters needed for the model and ensure that the resulting receptive fields are smooth in space and time. We choose the number of splines such that the log-likelihood on held out white noise data is maximized.

To further reduce the number of parameters, we assume our receptive field is rank 1, it can be separated into a spatial filter and a temporal filter. Following these assumptions we use SVD to find this rank 1 approximation. This provably minimizes reconstruction error under the Frobenius norm. We then crop the spatial dimensions for each cell to the regions containing the receptive fields and convolve the stimulus with the temporal filters, which leaves only spatial degrees of freedom.

For each cell, we then have a matrix of size N spikes by M features, where each feature is a spatial pixel convolved with time filter. We use Preconditioned ICA (Ablin, Cardoso, and Gramfort 2018), an algorithm for ICA that uses preconditioned L-BFGS, a low memory quasi newton optimization algorithm, for optimization to estimate 20 independent components.

Resulting components were considered proper subunit candidates based on the presence of significant spatial autocorrelations, following methods in (J. K. Liu et al. 2017).

With a list of candidate subunits for each cell we then computed the activation of that subunit by projecting the time convolved stimulus onto each filter identified by ST-ICA. Two units were considered the same following methods developed in (Jia et al. 2021).

CHAPTER 3

RETINAL GANGLION CELL POPULATION STRUCTURE ALLOWS ACCURATE DECODING OF NATURAL SCENES

3.1 Introduction

Sensory perception depends not only on the successful encoding of complex, varying stimuli, but also on a reliable, consistent method with which to read out that encoding downstream. The necessity for reliable readout puts significant strain on any encoding scheme. The encoding scheme must be flexible enough to extract information from a large variety of scenes while maintaining some fixed components that can be relied upon for readout. This problem becomes increasingly difficult in natural environments, where visual stimuli vary in luminance over many orders of magnitude (Rodieck 1998) and variance (Ruderman and Bialek 1994) (Schwartz and Simoncelli 2001), and have complicated temporal and spatial structure (Dong and Atick 1995; Hateren and Ruderman 1998). Retinal ganglion cells maintain their ability to encode these differing natural scenes at the individual level by adapting their response profiles to match the statistics of the environment (Fairhall et al. 2001). This flexibility at the single cell level, however, means that downstream readout becomes more difficult.

While single cells encode stimulus features, it is their population response that drives perception. For that reason, any static features in the population structure may be useful for decoding downstream. Previous work has shown that sparse, strong cell-cell interactions between neurons remain consistent across both natural and synthetic stimuli (Hoshal et al. 2023). These interactions, the noise correlations, have been shown to shape individual cell encoding (Tkačik et al. 2010) and therefore might be a valuable feature to leverage for downstream readout at the population level. Knowing how a network of cells must interact with each other, regardless of incoming stimuli, might make a downstream decoder robust to the large scale adaptation at the single-cell level.

How can stable pairwise interactions be leveraged to facilitate robust readout? Together,

individual cell responses to stimuli and the pairwise correlations between them imply a graph structure that might be well suited for graph-based decoding. To exemplify this, we imagine the RGC population as a graph $G = \{V, E\}$ with nodes V as the individual neurons and edges E the known cell-cell interactions between them. The individual nodes $v \in V$ represent each individual RGC's response to stimuli, while the edges $e \in E$ between them contain information about the fixed population level structure. Such graphs are the foundation for graph neural networks (GNNs), a type of neural network that takes graphs as inputs and learns an embedding space that represents both local graph structure and node features (Kipf and Welling 2016). GNNs have been employed for successful classification in a variety of contexts, from differentiating protein families (Zhang and Kabuka 2018) to determining loyalties within a social network (Zhuang and Ma 2018). These graph-based decoding strategies might extend to natural scene recognition using features of the retinal response as graph inputs.

To test this notion, we train a GNN on population retinal ganglion cell responses to multiple repeats of three different natural scenes and show that this GNN learns representations that support accurate classification of scene identity, even using single trial data. We further show that removing cell-cell interactions limits the cross-scene separation of representations learned by a multilayer perceptron, demonstrating the importance of these interactions for supporting the decoding of scene information on single trials. Finally, we show that the learned embedding space for this model is robust enough to support classification of single-trial data from a fourth movie not encountered during training.

3.2 Data and GNN architecture

Data

Voltage traces from the RGC layer of a larval tiger salamander retina were recorded following the methods outlined in (Marre et al. 2012). In brief, retina from a larval tiger salamander was isolated in darkness and pressed against a 252 channel multielectrode array.

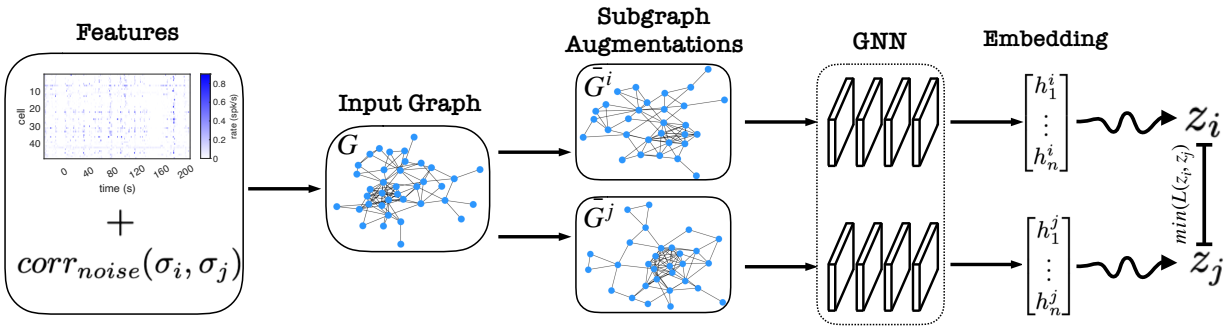

Figure 3.1: **Schematic of GraphCL**. First, the PSTH and pairwise noise correlations comprise the feature vectors for a graph $G = \{V, E\}$. Two subgraph augmentations $\bar{G}^i$ and $\bar{G}^j$ are generated from G to form a contrastive pair. These subgraphs are each independetly fed through a GNN to obtain embeddings h_i and h_j. A nonlinear transform is applied to the embeddings and agreement between the resulting z_i, z_j is maximized by minimizing the contrastive loss between them.

Multiple repeats (minimum 80) from five different natural movies are displayed in a pseudo-random order. Recordings were taken during stimulus presentation and spike sorted using a mostly automated spike sorting algorithm. This technique captured a highly overlapping neural population of 93 cells that fully tiled a region of visual space. Spikes were binned at 60Hz for all analyses presented.

GNN architecture

We modify a graph convolutional network from You et al. 2020, hereby called GraphCL. GraphCL is a 4-layer graph neural network designed to train using contrastive learning on input graphs. Contrastive learning in graphs is inspired by the same technique in image processing for convolutional networks (T. Chen et al. 2020). In principle, contrastive learning uses positive (alike) and negative (unalike) pairs of images to learn to build representations that encode the contrast between any pair of samples. This framework has been successful for unsupervised image classification. We modify GraphCL to have $dim = 80$ hidden units and output layers.

Training GraphCL

We form our input data by generating augmented subgraphs generated from neural responses to three different natural scenes. Each graph $G_{movie} = \{V, E\}$ takes as node features the PSTH for each neuron and as edge features the top 10% of noise correlations between all pairs of neurons. This creates a sparse graph with each node using as features a description of a the corresponding neuron's average activity over time for each stimulus. The edges remain fixed for all three graphs. Importantly, the graph inputs are unlabeled. The goal of GraphCL is this context is to generate an embedding with three emergent clusters, representing the input graphs from three different natural scenes.

Training GraphCL required the generation of two groups of subgraphs from each of these three graphs. First, we generated a group of subgraphs as part of a data augmentation procedure (a common technique for successfully training deep neural networks to build robust representations across deformations and transformations of their input spaces). Second, from

these subsampled graphs G we generate pairs of subgraphs to use for contrastive learning, $\bar{G^i}$ and $\bar{G^j}$. All subgraphs are generated with a dropout criterion of 0.2.

GraphCL learns this embedding via contrastive learning. For each augmented input graph G, GraphCL samples two distinct but related subgraphs of G, $\bar{G^i}$ and $\bar{G^j}$, and feeds them independently into the graph neural network based encoder to generate graph embeddings $h_{\bar{G^i}}$ and $h_{\bar{G^j}}$. These embeddings undergo a nonlinear transformation that maps them each to a latent space Z, where z_i and z_j are compared via contrastive loss. GraphCL learns by maximizing the agreement between the representations of these contrastive pairs. See Figure 3.1 for details.

3.3 Results

Learning representations using population structure captures natural scene identity from activity on single trials

While repetition-averaged PSTHs are known to support scene classification by themselves, classifying natural scenes from activity on single trials poses significant additional challenges. In particular, neural responses even to the same stimulus are stochastic and thus different trial-by-trial, and the probabalistic nature of neuronal firing means bit-flip errors are common over the course of a stimulus presentation. Although the PSTH gives a good estimate of the expected neural response over time, which is robust to trial-to-trial variability, the brain is constantly tasked with decoding in this extremely challenging single-trial regime.

To test whether taking advantage of cell-cell interactions can improve single-trial decoding, we fit and validate our GCN on input graphs that have information on the trial-averaged population response during stimulus presentation, via PSTHs, and the connectivity between neurons, based on the noise correlations. This gives the GCN a prior on the neural activity for each scene and allows it to use functional interactions to learn an embedding space. In testing, however, we feed in graphs with single trial data as feature vectors, instead of the PSTHs, and see where in the learned embedding space the GCN places that graph.

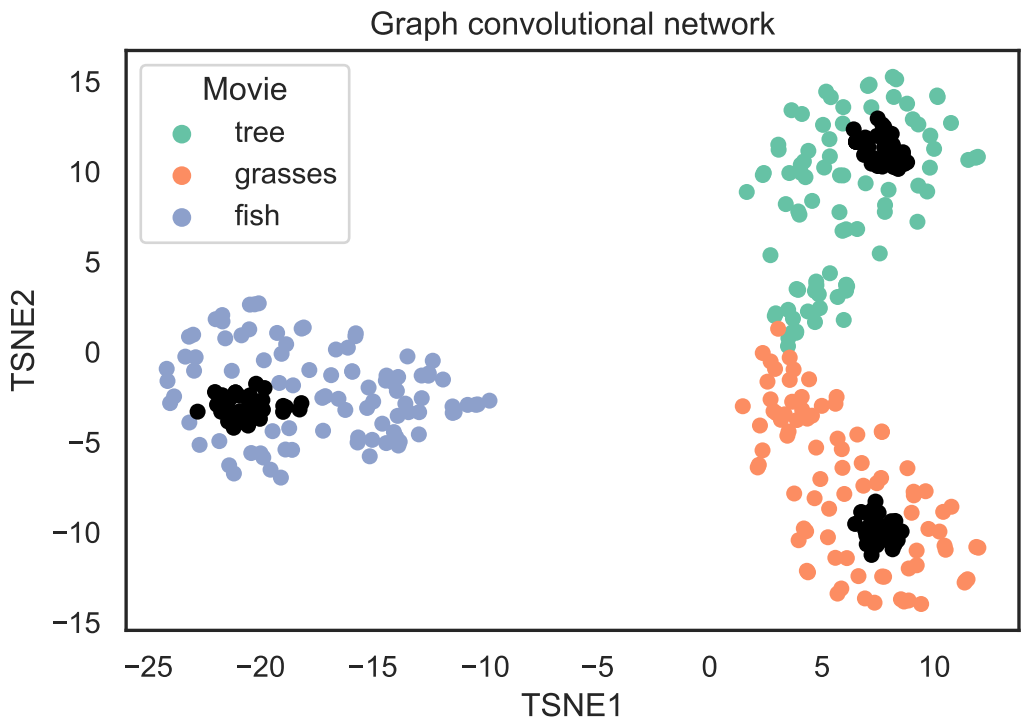

Figure 3.2: **GCN embedding space shows clustering of single trial activity for natural scenes.** A TSNE of the embedding space for the train graph convolutional network. The embedding space for validation graphs are shown in black. The single trial activity for each movie, shown in colors, populate the embedding space near one of the three clusters from the validated graphs from training on PSTHs.

We visualize the embedding space using t-distributed stochastic neighbor embedding (TSNE) on the last layer of the GCN (Fig 3.2). Embeddings from the validation set of graphs are shown in black. There is clear, emergent clustering of the three different movies, even though the learning process is unsupervised with respect to movie identity. This exemplifies GCN's ability to learn to discriminate the varying statistical structure of the neural response to each natural scene. The colored dots represent the test data for single trial activity of natural scenes. The clear separability of single trial data into three distinct classes shows the graph structure and PSTHs are sufficient for discriminating between diferent natural scenes at the level of neural activity on single trials.

To test whether the population PSTH alone is sufficient for single trial scene identification,

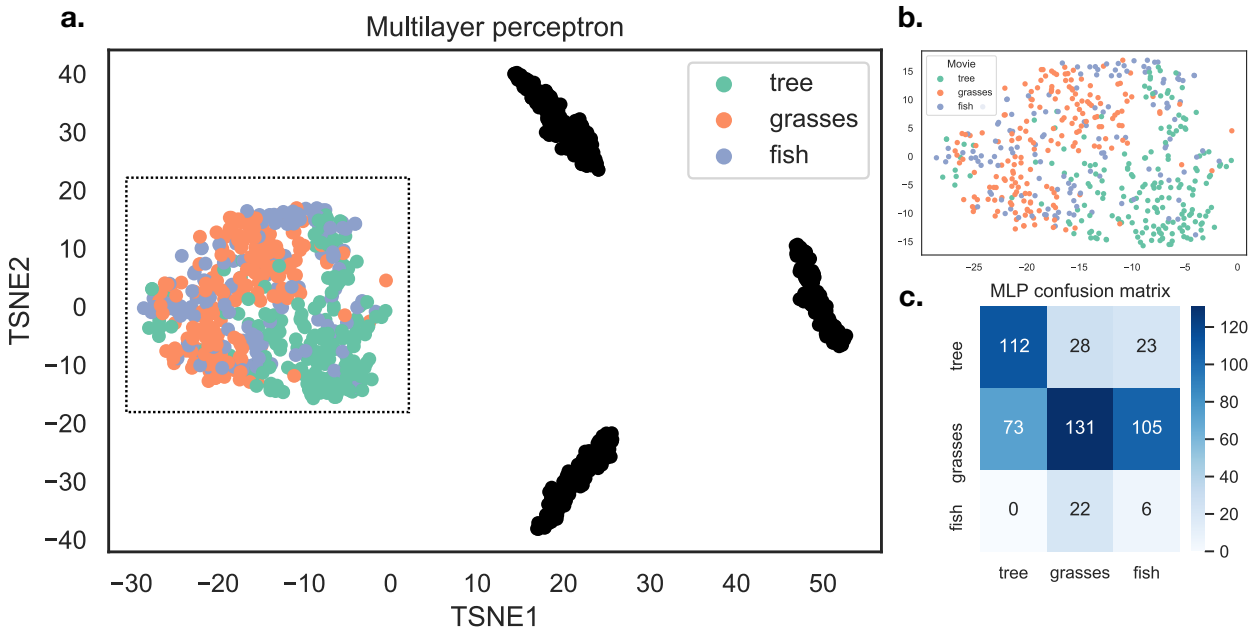

Figure 3.3: **MLP fails to classify single-trial activity** a) A TSNE of the final activation layer of the MLP trained and validated on labeled population PSTH activity. The validated data, in black, forms three distinct clusters that are decodable with 100% accuracy in validation. The single trial data,in color, lies far from the validation clusters and is not easily separable. b) A magnification of the boxed single trial data shown in a. The fish movie cannot be separated from the tree and grasses movies. c) The confusion matrix on the single-trial test data for the MLP.

we perform a similar analysis using a trained multilayer perceptron (MLP). An MLP is at minimum a three layer, fully connected feedfoward network known for its ability to classify data that is not linearly separable. We designed the MLP to have an equal number of layers and output dimensionality as our GCN implementation to control for any difference in cross-scene separability of representations due to parameter count. The MLP takes as input the neural population PSTHs, along with scene labels, but does not have access to the pairwise correlations between neurons. We perform a TSNE on the activations of the MLP's last hidden layer prior to classification (Fig 3.3a), observing that the MLP can easily separate the validation data into three distinct classes and obtains 100% classification accuracy. However, the MLP struggles to classify the single trial activity (Fig 3.3b). In particular the "fish" movie is indistinguishable from the "grasses" movie and the "tree" movie, as shown by the classifier's confusion matrix (Fig 3.3c). The failure of the MLP to classify single trial data based on PSTHs alone indicates that an independent readout mechanism may not be sufficient for single trial scene classification.

Separable zero-shot embedding of single-trial neural activity responding to a novel natural-scene stimulus

During the training of the GCN, we allowed the GCN to use PSTHs during three different natural scenes, which facilitated the learning of an embedding space that separates between each scene, even at the level of single trials. In principle, this setup relies on having a known prior (the PSTH) to perform decoding, i.e. the GCN might not be able to sensibly embed a novel stimulus. This is an ethologically unrealistic decoding scenario: when animals enter visually-novel settings, their visual system must still furnish them with information appropriate for coordinating action, even when the context is new. To probe whether incorporating neural population interaction structure can help to support stimulus decoding in this kind of setup, we evaluated our fit GCN on single trial neural responses recorded during the presentation of a held-out stimulus. If population interactions can support the discrimination of different scenes at the level of single trials, then the resulting embeddings

should cluster together, reflecting the mutual similarity of the activity state on all trials where this held-out scene was presented. Further, this cluster should segregate from the embedding clusters corresponding with the activity during the three movies used to fit the GCN. For this task, we chose to feed in an optic flow movie, which has wildly different statistical structure from the three previously included scenes (this is due to the camera moving in the optic flow movie, creating radial motion that is not present in the other movies with a fixed perspective). This ensures that the novel stimulus *should* be clustered away from the other movies, while also asking whether the learned embedding space is robust to a novel set of stimulus statistics that it likely did not encounter during training. In Fig 3.4, we again show a TSNE for the validation graph set, as well as the single-trial activity from the natural scenes. The optic flow movie trials cluster tightly together, showing that the learned embedding space is able to generalize to completely novel single trial data. Further, this cluster of optic flow trials is completely separable from the other three stimulus classes. This demonstrates that a graph based decoding scheme using only PSTHs and known consistent population structure is sufficient for generating representation spaces in which even novel scenes are distinguishable.

3.4 Discussion

In this chapter, we demonstrate that activity of retinal populations during the presentation of natural scenes can support the learning of embeddings (using an unsupervised, connectivity-informed technique) that are strongly mutually separable, even at the level of single trials. Even though single-cell responses can rapidly adapt to novel stimuli, the presence of consistent noise correlations is sufficient to separate between neural responses across a variety of different natural scenes. Questions remain as to whether these noise correlations are necessary for single-trial scene decoding. The failure of an MLP to do single-trial classification based on trial-averaged population neural activity does not imply cell-cell interactions are required, just that PSTHs alone are not sufficient. Some other feature of single cell response

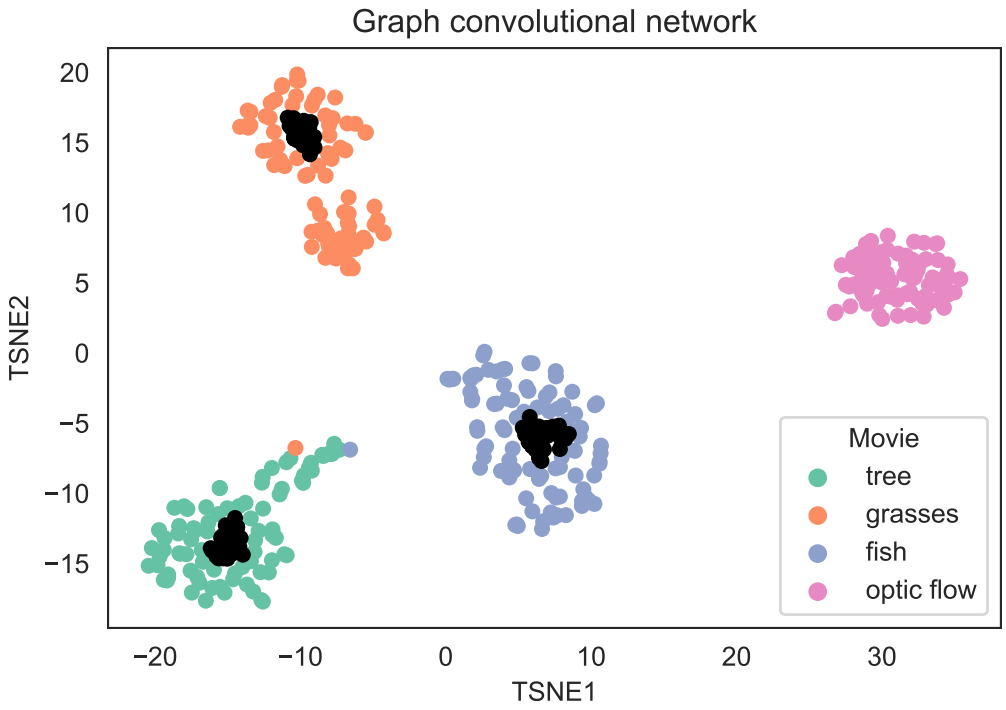

Figure 3.4: **Zero-shot classification of single-trial activity for a novel natural scene.** A TSNE of the embedding space from the GCN, as in 3.2, now included single trial activity for the optic flow stiumlus. The GCN was not trained on data from the optic flow stimulus, but still clusters the single trial together and separates it from the other natural scenes.

that we did not feed into our networks might rescue the MLP single trial decoding.

While our GCN uses only a sparse graph of known noise correlations for single-trial stimulus embedding, other edge structures may be better suited to this task. For example, we also tested a k-complete unweighted graph that had even better task performance. The aim of this paper is *not* to find the best graph based decoding mechanism for single trial readout. Instead, we ask whether a known feature of the RGC population structure is *sufficient* for such readout. As such, we decided to minimally transform our data, feeding in neural responses and edge information exactly as one might expect a downstream neural decoder to receive them.

Our GCN performs near-perfect single trial scene classification using a fairly high dimensional ($dim = 80$) embedding space. That the brain might support high-dimensional embedding spaces is certainly plausible, given the extremely high ambient dimensionality of even a single brain region in a small organism's brain (where $d_{ambient} = N$ neurons; for most brains/brain regions, $80 << N$). Even so, there are substantial theoretical and empirical reasons to expect much lower-dimensional embeddings. Theoretically, high-dimensional embeddings generally pose problems for encoding generalizability, which would be a problematic shortcoming for the retina (the bottleneck for all visual representations in the brain). Practically, correlated variability constrains the maximum possible intrinsic dimensionality that an embeddiing could realize (not all directions in the $N-$dimensional neural state space can be explored independently). In the context of the retina in particular, previous work has shown exactly this kind of result: retinal responses are compressible to a low-dimensional encoding space (e.g., 10 dimensions, as in (Wang et al. 2022)). Further work should explore whether embeddings learned via GCN that support separable representations at the level of single trials can happen in this kind of low-dimensional space which may be easier to read out.

Future work should expand on the results presented here in several key directions. First, choosing an appropriate null model for this setting is both important and technically difficult.

An ideal control, fully accounting for any possible difference in embedding model approach, parameter count, etc., would use the same GCN structure for learning the embedding space, but under different perturbations of the edge structure that governs message passing. However, our implementation of the GraphCL algorithm precluded the use of a no-edge model (learning the embedding space requires contrastive learning on augmented subgraphs, which cannot exist wiithout any edges), and is too flexible for an 'all-edges-equal' model to be a relevant comparison. In principle a node-dropping augmentation rather than a subgraph augmentation could ameliorate these shortcomings in this setting. Second, UMAP might be a more appropriate dimensionality reduction technique to visualize the embedding space as compared to t-SNE, as it better preserves the local and global structure of the embedding space.

CHAPTER 4

LEARNING LOW-DIMENSIONAL GENERALIZABLE NATURAL FEATURES FROM RETINA USING A U-NET

The following work was done along with Siwei Wang, who is the primary author on this paper. The work is presented here in its entirety for the sake of providing meaningful context to the reader. My contributions to this work include observing the benefits of simultaneously encoding static and dynamic feature representations in the retina.

4.1 Abstract

Much of sensory neuroscience focuses on presenting stimuli that are chosen by the experimenter because they are parametric and easy to sample and are thought to be behaviorally relevant to the organism. However, it is not generally known what these relevant features are in complex, natural scenes. This work focuses on using the retinal encoding of natural movies to determine the presumably behaviorally-relevant features that the brain represents. It is prohibitive to parameterize a natural movie and its respective retinal encoding fully. We use time within a natural movie as a proxy for the whole suite of features evolving across the scene. We then use a task-agnostic deep architecture, an encoder-decoder, to model the retinal encoding process and characterize its representation of "time in the natural scene" in a compressed latent space. In our end-to-end training, an encoder learns a compressed latent representation from a large population of salamander retinal ganglion cells responding to natural movies, while a decoder samples from this compressed latent space to generate the appropriate future movie frame. By comparing latent representations of retinal activity from three movies, we find that the retina has a generalizable encoding for time in the natural scene: the precise, low-dimensional representation of time learned from one movie can be used to represent time in a different movie, with up to 17 ms resolution. We then show that static textures and velocity features of a natural movie are synergistic. The retina

simultaneously encodes both to establish a generalizable, low-dimensional representation of time in the natural scene.

4.2 Introduction

The flexibility and computational power of convolutional neural networks (CNNs) has helped sensory neuroscience model the neural code for natural stimuli with rich feature repertoires. It has been shown that CNNs carryout encoding computations similar to those observed in the retina (McIntosh et al. 2016; Tanaka et al. 2019). A CNN also makes the inverse problem of decoding complex stimuli from the retinal response (Botella-Soler et al. 2018b) more tractable. Understanding decoding has its own significant merit because neural systems downstream of the retina can only 'see' the world through the retinal code. For any visual information to be used to guide behavior, it must first be decoded from retinal responses. Historically, the decoding problem for natural stimuli has been challenging because both the retinal response and the stimuli are high-dimensional. Although deep neural networks can capture the high dimensionality of neural inputs and responses (LeCun, Bengio, and Hinton 2015), they do so by projecting the neural code into another high-dimensional parameter space that is also hard to interpret. While these tools tell us that we *can* decode, our ability to understand the features of neural activity relevant for that decoding is limited. In this work, we propose a novel artificial neural architecture that can decode complex natural scenes from retinal responses with high fidelity while providing a low-dimensional latent space that is interpretable. Using this architecture, we obtain unique insights into what features are important for reading out the retina's population code, and why encoding these features might enable an animal to navigate a complex, dynamic natural environment.

We anchor our results on a unique dataset from the salamander retina. Nearly one hundred output ganglion cells were recorded simultaneously while several different natural movies were projected onto the photoreceptor layer. The relatively long (20 s) movie clips were played many times, in a pseudo-random order. During the lifespan of a salamander, it

goes through a transition from being aquatic to terrestrial. The sampled movies attempt to span these different motion environments. A movie of small fish in an aquarium with live plants was set up to match what a larval tiger salamander might see underwater while it hunts for food. A movie of leaves blowing in the wind resembles the scene a salamander may live in after it undergoes metamorphosis. Does a salamander retina re-use how it encodes features during the aquatic larval phase to represent features in a terrestrial scene? This motivates us to investigate whether the encoding of natural features from one particular movie is generalizable to a novel movie.

Natural movies contain complex spatio-temporal features on multiple scales. This makes enumerating all possible stimulus states in natural movies intractable. It is more feasible to investigate how the retina encodes time in the natural scene, as has been done in other studies (Xia et al. 2021). The salamander retina elicits precisely timed spikes (Berry, Warland, and Meister 1997). The idea that the retina may encode how features change over time to discriminate between frames has been explored before (Schwab et al.). It has an intricate connection to stimulus-dependent representational drift in sensory systems (Marks and Goard 2021). Previous work (Xia et al. 2021) reported that a low dimensional compressed representation of activity from a mouse V1 population can be used to discriminate frames that are 1s apart. Furthermore, the authors showed that if such an encoding of time in the natural scenes exists, it is likely to be low-dimensional (Xia et al. 2021). We train our deep neural network (an encoder-decoder in machine learning parlance) (Fig 4.1 and Section 4.3) for decoding. It reconstructs movie frames from retinal responses. This is different from previous works (McIntosh et al. 2016; Tanaka et al. 2019) that use CNNs as "in-silico retinas" to understand how specific stimuli drive the retina. During training, the encoder part of the deep neural network learns a continuous, compressed latent representation of the retinal responses from which the decoder part samples to reconstruct target movie frames. We enforce structural constraints (Hinton and Zemel 1993; Bengio, Courville, and Vincent 2013; Bowman et al. 2015; Kumar and Poole 2020; Kingma and Welling 2013) to obtain

meaningful, continuous latent space. We find that using this compressed representation, we can decode time in the natural scene up to single-frame resolution (17ms) (Fig 4.1). This learned, compressed representation allows for precise decoding in a novel set of natural scenes. We show that the retina is responding to spatio-temporal features that change over time, rather than having a clock. We divided these features into static (texture) and dynamic (optic flow–a vector field describing the motion between subsequent frames (Horn and Schunck 1981)) motifs. This allows us to construct two distributions which reflect how static features and dynamic features cluster to discriminate between frames. We find that these features are synergistic with respect to the encoding. By simultaneously encoding static and dynamic features, the retina establishes a generalizable, low-dimensional representation of time in natural scenes.

4.3 Data and encoder-decoder architecture

Data

Our dataset contains retinal recordings of 93 cells responding to repeated, 20 second presentations of three natural movies at 60 frames per second. There are 85-90 presentations of each movie interleaved in pseudo-random order. Spikes are binned at 17ms, to align with the movie frame duration. We compute each neuron's firing rate as a function of time within each movie, (the peri-stimulus time histogram, or PSTH) by averaging spikes across trials in these bins [1].

Encoder-decoder architecture

We use a U-net (Ronneberger, Fischer, and Brox 2015) as the backbone architecture. The U-net supplements an encoder of contracting layers with a nearly symmetric decoder of expansive layers (hence the U-shape, see Supplementary Information) (Fang et al. 2021; Howard et al. 2018). It is successful in domain-conversion problems, i.e., text-to-speech (R.

1. You can find our Pytorch implementation at https://github.com/sepalmer/VU-net

Li et al. 2021). We modify its skip connections to supply the decoder noisy features from the encoder. Because the intermediate features from the feedforward encoder have different resolutions at their convolutional layers, these skip connections enable the decoder to form a multi-scale and multi-level feature representation of the input. Our particular network has an encoder with the same architecture as ResNet18. We also initialize it with the ResNet18 (He et al. 2016) weights pre-trained by ImageNet (Krizhevsky, Sutskever, and Hinton 2012). Our decoder mirrors the feedforward architecture of the encoder (see Supplementary Information). In addition, we turn these skip connections into variational sampling layers and use them as compressed representations of the input retinal activity. The set of activation values in this latent space in response to the neural activity (input) from a given stimulus movie is referred to as Z_{movie}, where 'movie' is either 'fish', 'water', or 'leaf'). We constrain all variational sampling layers to have the same dimensionality, to simplify the training.

Knowing that the decoder reconstructs the movie frame from features across multiple spatial and temporal scales, we use the perceptual loss (Johnson, Alahi, and Fei-Fei 2016) as the objective function for the encoder-decoder. This loss function compares a reconstructed frame with its target frame with respect to learned features from a pretrained VGG, as opposed to their raw pixels. Throughout the paper, we use features from the pretrained VGG19 to represent movie frames.

4.4 Results

Retinal activity has a low intrinsic dimension

Fig 4.1A shows how we train an encoder-decoder for a specific movie (in this example the fish movie). The input is the retinal PSTHs from the 45 most reliably spiking cells in the population (see Supplementary Information). We take trial-averaged firing rates, and ignore 'noise' correlations between cells, to make an initial pass at this high-dimensional problem. Considering that many of the observed retinal computations happen on a timescale below 400ms (Baccus and Meister 2002), we restrict ourselves to 500 ms long snippets of the PSTHs.

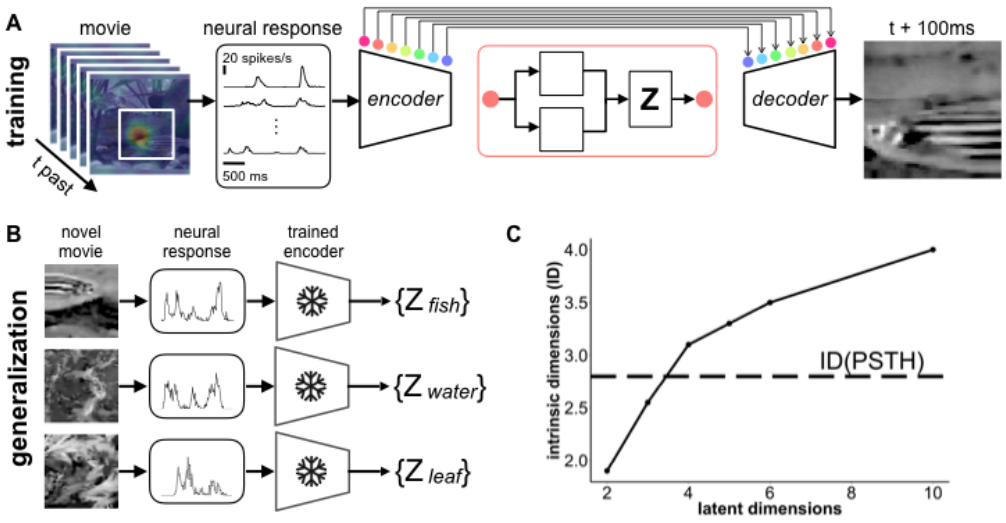

Figure 4.1: **Retinal activity has low intrinsic dimension** A) Encoder-decoder network trained to predict a movie frame 100ms in the future from a 500ms window of retinal activity in the past (the aggregated receptive field of the retina population is shown in the heat map from the left). The network learns a low dimensional variational representation, where μ and σ characterize the posterior distribution $p(x|z)$. The samples drawn from this latent space are referred to as Z in subsequent sections. All skip connections are trained to obtain a separate latent space for Z. Using a highly expressive encoder (Lin and Jegelka 2018), i.e., ResNet18, we empirically observe that the latent space learned for all skip connections are similar. B) We train the network on one movie (fish, for example), then the encoder weights are frozen. We obtain the Z's for retinal activity responding to different movies by passing test samples from all three movies through the trained encoder. C) The intrinsic dimensionality of the retinal activity (dashed line) and the latent dimensions for latent space with varying $\dim Z$. Note that intrinsic dimension measures the complexity of the retinal activity. We use it as a lower bound to constrain $\dim Z$ and add additional latent dimensions to improve reconstruction of target movie frames. We stop at $\dim Z = 10$ because we observe highly accurate reconstruction with $\dim Z = 10$ empirically.

To encourage the model to learn temporal structure within the movie, we ask the decoder to reconstruct a movie frame 100ms in the future, after the end of the 500ms snippet of neural response. We choose this particular Δt based on the timescale of predictive information in retinal populations (Palmer et al. 2015). We train one predictive encoder-decoder for a specific movie using 40,000 training samples (see Supplementary Information for details) with a 90%/10% training/validation split. The reconstruction is from an additional held-out test set of 10,000 samples (100 frames, 100 PSTH patterns in each frame, see Supplementary Information). We also train a static encoder-decoder that learns to reconstruct a frame centered within the 500ms window of neural activity (using 250ms before and after the target frame as the input, similar to (Botella-Soler et al. 2018b)). The predictive encoder-decoder achieves a reconstruction performance similar to the static one. We focus on the predictive encoder-decoder in our subsequent analysis because it may capture both static and temporal structures of a natural movie by design.

The latent space is a compressed representation, Z, of the mean firing rate patterns from the retinal population. We estimate the intrinsic dimension (Pope et al. 2021; MacKay and Ghahramani) of the retinal activity and use it to guide the selection of the dimensionality of the latent space. This intrinsic dimension is the number of variables needed to describe a data distribution (Levina and Bickel 2004). It is also a complexity measure of data because it determines the number of samples needed to characterize a data manifold (Narayanan and Niyogi 2009; Narayanan and Mitter 2010). We find that our estimate of the intrinsic dimensionality for all retinal activity in response to the three movies yields the same result as the estimate of the intrinsic dimensionality from the activity for each movie separately, i.e., including retinal activity from a different movie does not add complexity to the latent space representation of the retinal response. This suggests that features in natural movies may be encoded by the retina in a generalizable way across movies. We use the intrinsic dimension of the retinal activity as a lower bound to determine latent dimension, i.e., $\dim Z$ in the encoder-decoder. We then add additional latent dimensions to help encode factors

that may result from combinations of different intrinsic dimensions. We observe that the increase in the estimated intrinsic dimension of the latent space decreases after $\dim Z > 4$. Meanwhile, we empirically observe good reconstruction performance (the pixel MSE is about 0.02 averaged over 100 test frames of size 64X64 with pixel intensity $\in [0, 255]$) when we use latent dimension equal to 10 to reconstruct the held-out movie segments (see Supplementary Information). Thus, for further analyses, we use $\dim Z = 10$ unless otherwise stated. To address questions about whether features learned from retinal activity responding to one movie are generalizable to another novel movie, we generate "mismatched" Z's. These "mismatched" Z's are compressed representations of the retinal responses to Z_{water} and Z_{leaf} movies by passing those inputs through the encoder trained for the Z_{fish} movie (Fig 1B). The results presented below are similar regardless of which movie (water, leaf, fish) we train on and which other two movies are used to generate mismatched Z's.

The retina encodes complex, but interpretable spatio-temporal features in natural movies

The goal in this section is to qualitatively assess latent space stimulus features to show that the encoder-decoder contains features that are plausibly encoded by the retina. We visualize highly activated features from decoding layers of multiple spatial scales (threshold by top 1% activation intensity, see Supplementary Information for more visualization). We find both features that resemble background motion and features that resemble object motion (Fig 2) in a specific decoding layer. In particular, the object motion feature closely traces the movement of the fish in the target movie segment. The average activation of this feature along the x-axis shows that this feature may be responsible to decode both position and velocity of the fish movement in the target movie frames (top row). Because removing the latent activations silences both background and motion features, we also determined that this specific variational sampling layer generates these disentangled features. This observation agrees with previous experiments that the retina can encode features that evolve across time and space (Kühn and Gollisch 2016). This motivates us to investigate whether we can decode the "time in the natural movie", i.e., discriminate between different frames from this specific

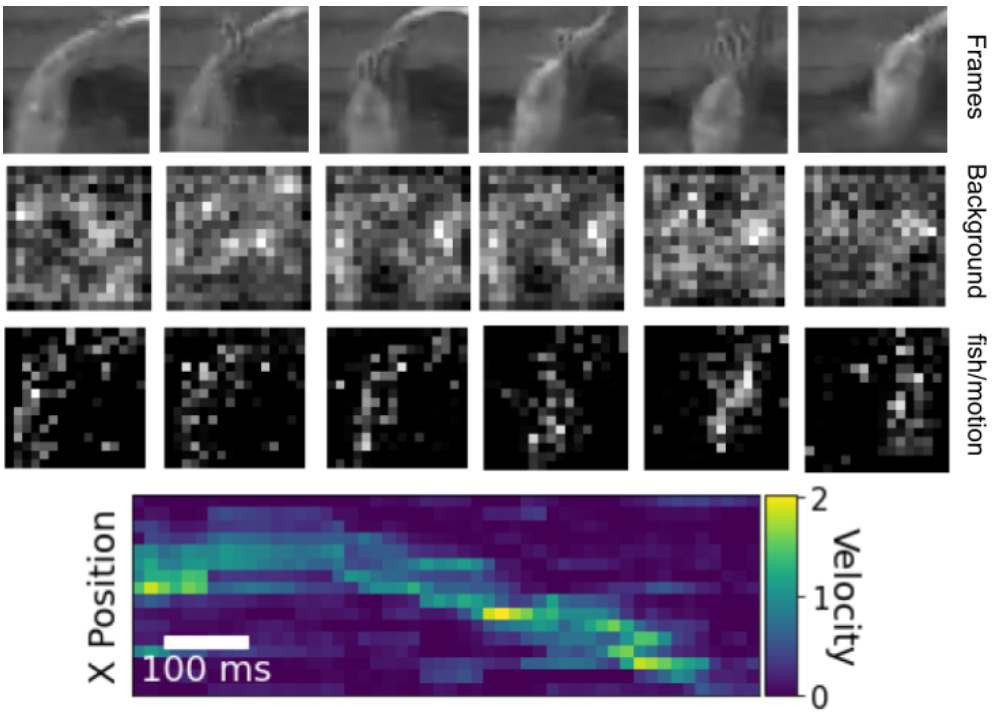

Figure 4.2: **Features from a decoder layer separately decode background and object motion.** 1st row (from top to bottom): target movie frames that are 100ms apart; 2nd row: an example feature inside the trained decoder and its activation for decoding frames in the top row. 3rd row: the spatial, temporal activation. 4th row: X-T receptive field (aggregated x-axis activation over time).

decoding layer.

Low dimensional, generalizable representation of time in multiple movies

In this section, we investigate whether we can discriminate frames in natural movies and thus decode time in natural scenes. We can also ask whether the feature space of the retinal population used for one movie can be used for the other two natural movies in our three-movie dataset. If this is the case, it would suggest that there is a general representation of spatio-temporal features in retinal activity that supports this decoding of time.

Inspired by previous work that the visual cortex may contain a low-dimensional representation of complex stimulus features evolving over time (Xia et al. 2021), we decode time in the natural scene with the particular latent space that corresponds to the decoding layer shown in Fig 4.2. We obtain the compressed representation of retinal activity on the held-

out test frames for the fish movie, i.e., Z_{fish}, as well as test frames for two other novel movies (e.g., leaf and water). This use of an encoder-decoder to generate a compressed representation of a data distribution has been investigated in detail in representation learning (Hinton and Zemel 1993). For each $Z_{fish,leaf,water}$ separately, we linearly decode the 1D frame label of held-out test frames from the corresponding Z. Fig 3A shows all decoding performance as a function of the number of dimensions allowed in Z (we trained a series of encoder-decoders with different $\dim Z$). Because all encoder-decoder models are trained with the fish movie, the decoding for the fish movie outperforms the other two. However, we observe that $\dim Z = 5$ is sufficient for the encoder-decoder to build a latent space Z which can decode >80% frames in all three movies. This is also the dimensionality needed for decoding in the natural scene reported in mouse visual cortex (Xia et al. 2021). Since these frames are 17ms apart, this decoding performance shows that salamander retina establishes a low dimensional representation of "time in the natural scene" with fine temporal resolution.

The high decoding performance from the mismatched Z's suggests that retinal activity may establish a general encoding for "time in the natural scene". How is this possible? The retina is presumably encoding general space-time features that are predictive of the future space-time features in natural scenes. These could be complex, but they seem to generalize across movies. We use information theory to directly evaluate the information that all three Zs contain that is relevant for decoding time. We first calculate the mutual information between latent representations of retinal activity responding to different movies (here,Z_{fish} is the learned compressed retinal responses to the fish movie, so we show the mutual information of fish vs. leaf $I(Z_{fish}; Z_{leaf})$ and fish vs. water, $I(Z_{fish}; Z_{water})$, respectively). We also observed that $I(Z_{fish}; Z_{leaf}, time) = I(Z_{fish}; Z_{leaf})$, i.e., including time does not add additional mutual information. This tells us the latent representation obtained from retinal activity for one movie encodes the generalizable "time in natural scenes" for a different movie, up to the full entropy of time itself. To show this, we use the

chain rule of mutual information and subtract mutual information that is independent of time, i.e., $I(Z_{fish}; Z_{leaf}|time) = I(Z_{fish}; Z_{leaf}, time) - I(Z_{leaf}; time) = I(Z_{fish}; Z_{leaf}) - I(Z_{leaf}; time)$, in Fig 4.3B. The cyan bar shows the MI between Z_{fish} and mismatched Z that is about time, $I(Z_{fish}; Z_{leaf}) - I(Z_{fish}; Z_{leaf}|time)$. It nearly saturates $H(time)$. We also observe that the mutual information between different movies is mostly about time. The difference between the pink and cyan bar is very small. This estimate confirms that the retina has a generalizable, precise representation of time in natural scenes that can discriminate consecutive frames that are only 17ms apart. Table 4.1 confirms that latent representations obtained from retinal responses to any one movie can be used to decode time for all three movies.

In the supplementary information, we also included simple visualizations and decoders to demonstrate that decoding "time in the natural scene" is challenging. First, "time in the natural scene" cannot be observed as a simple visual trend using the mean PSTH, pairwise frame-to-frame distance, or the 2D latent space trace. Second, we also included linear decoders trained on instantaneous PSTH's, raw PSTH's, shuffled PSTH's and three other dimensionality reductions (Isomap, 10D-PCA and 50D-PCA). We found that shuffled PSTH shows inferior performance compared to the one trained on raw PSTH's. This suggests that decoding time does not come from trivial gross changes in spiking statistics. We also observed that linear decoders fall short of the 10-D latent representation learned by the U-net in terms of generalizable performance, even when using the 50D-PCA. To learn a low dimensional feature space that can be applied to all three natural movies, our variational U-net carries out significant nonlinear transformation.

Synergistic features for encoding "time in the natural scene"

Fig 4.3 shows that the retina has a low-dimensional, generalizable representation for time in the natural scene. We next ask what features the retina uses for this low-dimensional, generalizable representation.

Although there are retinal circuits that encode object motion, most decoding work only

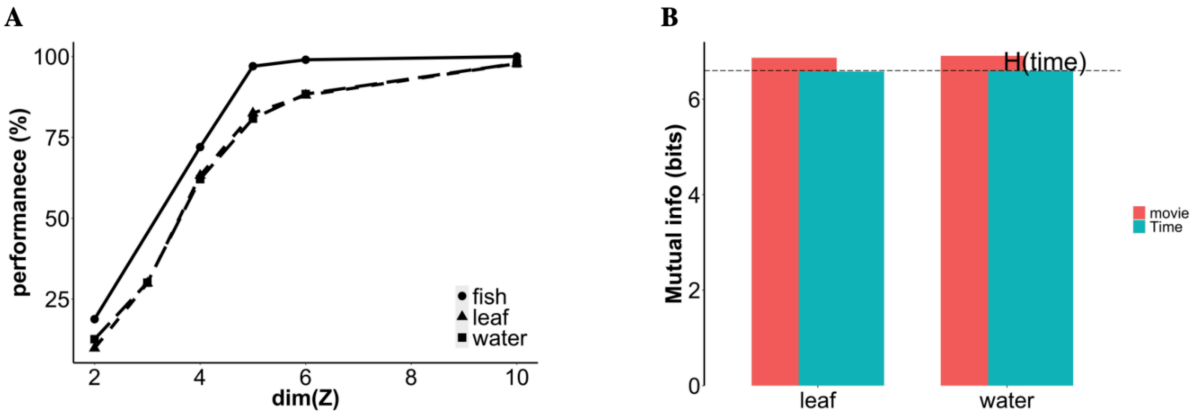

Figure 4.3: **A generalizable representation of time in natural scenes**A) Decoding performance of latent representations on test movie segments for all three movies. The encoder-decoder is trained in fish movie. The latent representations are obtained from retinal activity of all three movies (fish, leaf, water). B) Mutual information between movies vs. mutual information with respect to time. $H(time) = 6.6$ bits for the test movie clip of 100 frames long. Because we use the encoder-decoder trained with fish movie here, we show $I(Z_{fish}; Z_{leaf})$ and $I(Z_{fish}; Z_{water})$. (See the supplementary information)

	Water(5d)	Water(10d)	Leaf(5d)	Leaf(10d)
Fish	78.2%	96.9%	72.2%	98.0%
Leaf	79.5%	97.8%	84.4%	99.1%
Water	84.0%	97.4%	71.7%	97.9%

Table 4.1: Latent representations trained on any one movie can decode time in all three movies. Here we show $\dim Z = 5$ and $\dim Z = 10$, respectively.

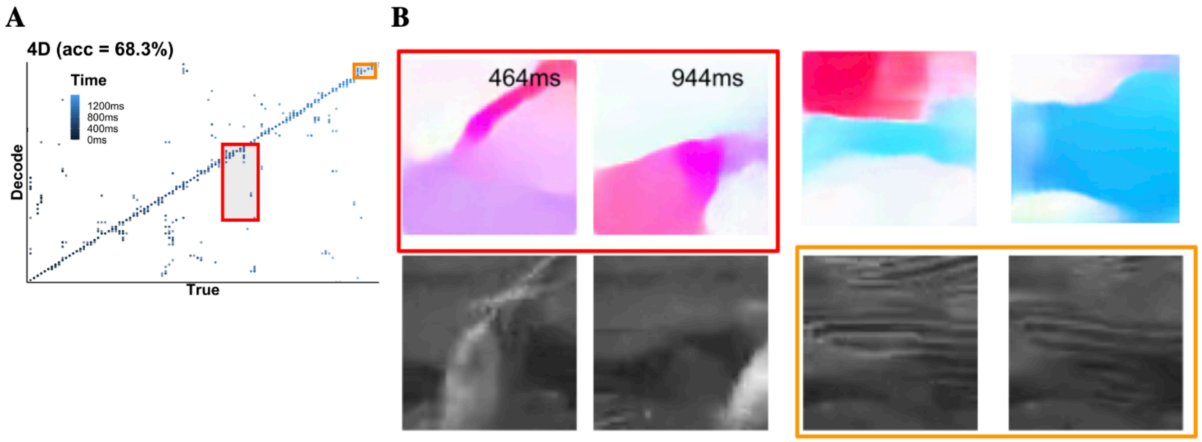

Figure 4.4: **Decoding performance and example decoding errors.** A) The scatter plot of decoding performance with a Z_{fish} of dim $Z = 4$. Correctly decoded samples appear along the diagonal line of true vs. pred (decode), incorrectly decoded samples show up in off-diagonal regions. With dim $Z > 4$, the decoding performance increases to $\sim 95\%$ (shown in Fig 4.3). B) Two examples of decoding errors (within the boxes shown in A). The two frames in the red box contain similar optic flow (dynamic) whereas the two frames in the orange box contain similar static textures.

uses natural images as their input stimuli (Brackbill et al. 2020). Our stimulus set allows us to explore how natural texture and motion might interact. In Fig 4.4, we show two example decoding errors (Fig 4.4B), one where the predicted and true frame have similar static features and one other where the predicted and true frames with similar optic flow (dynamic features). This indicates that our observed latent representation of "time in the natural scene" can be confused when either the static or dynamic structure between frames is similar. These examples are not sufficient to exclude the possibility that static textures may also be used to discriminate between different dynamic frames. To investigate this, we perform two parallel hierarchical clusterings: one on the frames themselves (static) and one on the optic flow frames (dynamic) of the test movie segment. If textures govern the discrimination between frames in both dynamic and static settings, then these clusterings should produce similar clusters. Before clustering, we convert all static frames and their corresponding optic flow frames to features that are the activation from the last ReLu layer

in a pre-traind VGG19 network (Kummerer et al. 2017). These activations are believed to mimic features used in human perception of generic natural stimuli (Johnson, Alahi, and Fei-Fei 2016). We observe that hierarchical clusterings on dynamic versus static motifs of all three movies yield different results (see Supplementary Information for details).

The difference between clusters of frames based on static versus dynamic features enables us to construct three distributions of "time in the natural scene" (Fig 4.5A). One uses the clustering based on the static features Y_{static}, another using the dynamic features $Y_{dynamic}$, and a third combining both sets of features Y_{joint}. By construction, we would like the joint distribution Y_{joint} to have an entropy as close as possible to the full entropy of time $H(time) = \log 2(100)$=6.6 bits. This construction narrows down our search of coding schemes to how the retina *combines* dynamic and static structures within natural movies to encode the "time in the natural scene".

Previous work showed that the retina performs efficient coding (Joseph J. Atick and A. Norman Redlich 1992b; H. B. Barlow 1961). Efficient coding predicts that redundancy should be minimized among different features of interest. Therefore, while we would like the joint distribution $H(Y_{joint})$ to contain most of the entropy of time, we also want to minimize the mutual information between the two components Y_{static} and $Y_{dynamic}$ (minimize their redundancy). We varied the threshold on the clustering hierarchies to coarse grain the distributions Y_{static} and $Y_{dynamic}$, and also to create the joint distribution Y_{joint}. Because we do not know *a priori* their relative contributions, we coarse-grain both clusterings with thresholds such that the coarse grained entropy of the two components are comparable, i.e., $E(Y_{static}) \sim E(Y_{dynamic})$. In Fig 4A, we show the coarse-grained joint distribution we use for subsequent analysis. It contains a small amount of redundancy, i.e., $I(Y_{dynamic}, Y_{static}) =$ 0.8 bits, while Y_{joint} includes most of the entropy for time ($H(Y_{joint}) = 5.05$ bits = 76% of $H(time) = 6.6$ bits). These two components are dominated by either static or dynamic features, respectively. In Supplementary Information, we discuss in detail how we find these distributions. A different threshold may either introduce a significantly higher redundancy

or sacrifice too much information from $H(time)$.

Using these distributions, now we can ask how the neural population encodes time through encoding the both the static and dynamic features of the natural scene. The latent activation of retinal inputs are compressed representations, so we frame this encoding problem using the information bottleneck method (Bialek and Tishby 1999). The information bottleneck method identifies whether a compressed representation T retains as much information as possible about the relevant variable Y while compressing away irrelevant components of the input X. In this context, the information bottleneck shows how much information a compressed representation needs, in order to encode a specific amount of information about the features of interest, Y.

The information bottleneck method minimizes the following objective function:

$$\mathcal{L}_{p(t|X),\beta} = I(X;T) - \beta I(Y;T) \tag{4.1}$$

To make a meaningful measurement of $I(X;T)$ and $I(Y;T)$, we first ensure that we have a meaningful latent representation Z. The challenge is to prevent the so-called "posterior collapse" (Bowman et al. 2015). This is a phenomenon previously reported in encoder-decoders with highly expressive architectures (like the ResNet18 network that we use here) (Lin and Jegelka 2018). These expressive architectures are capable of decoding complex features, e.g., our movie frames, without using Z. This results in the latent code Z only containing noise, i.e. $I(X;Z) = 0$. Here, we use a simple heuristic to circumvent this scenario. As discussed in (Razavi et al. 2019; Kumar and Poole 2020; Locatello et al. 2019), we can obtain a meaningful Z by making the posterior to have a small but nonzero noise. To be specific, we have $\sigma^2 > 0$ for $p(X|Z)$. This is also dubbed a 'committed rate' for the encoder. In our case, we choose $\log \sigma^2 = -1.0$ to further ensure numerical stability of the mutual information estimator that we use (Kolchinsky, Tracey, and Wolpert 2017).

The mutual information estimator also requires the prior of $p(Z)$ to have a factor-

ized marginal (each marginal is an independent Gaussian), $\mathcal{N}(0, I)$ (they are independent Gaussians). This is a typical constraint introduced in the original variational autoencoder (Kingma and Welling 2013). Combining this constraint on $p(Z)$ and the above constraint on $p(X|Z)$, we can approximate $I(X;Z)$ with the following estimator,

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X) \\ &\leq -\frac{1}{P}\sum_i \log \frac{1}{P}\sum_j \exp(-\frac{1}{2}\frac{||z_i - z_j||_2^2}{\sigma^2}) - \frac{D}{2}(1 + \log\sigma^2 + \log 2\pi) \end{aligned} \tag{4.2}$$

P is the number of test samples ($P = 10000$ in our case) and z_i, z_j are the latent activations for the i_{th} or j_{th} sample.

Because $P(Y)$ has a uniform distribution (100 retinal inputs per frame and there are 100 held-out frames), we can use the following estimator for $I(Y;Z)$ (we also validated this estimation with a widely used non-parametric estimator (Kraskov, Stögbauer, and Grassberger 2004), the difference is less than 0.1 in all our calculations):

$$\begin{aligned} I(Y;Z) &= H(Z) - H(Z|Y) \\ &\leq -\frac{1}{P}\sum_i \log \frac{1}{P}\sum_j \exp(-\frac{1}{2}\frac{||z_i - z_j||_2^2}{\sigma^2}) \\ &- \sum_{l=1}^{L} p_l \left[-\frac{1}{P}\sum_{i,Y_i=l} \log \frac{1}{P}\sum_{j,Y_j=l} \exp(-\frac{1}{2}\frac{||z_i - z_j||_2^2}{\sigma^2}) \right] \end{aligned} \tag{4.3}$$

Note that p_l is the number of test samples for the l_{th} frame. $p_l = 100$ in all of the test distributions.

Using $p(Z)$ with a factorized Gaussian marginal brings an additional benefit: it is the sufficient and necessary condition for the latent space to exhibit orthogonal symmetry (V.Skitovitch 1953; Darmois 1953; Lukacs and King 1954). This Darmois-Skitovitch characterization was first introduced to identify unique factors for independent component analysis (Hyvärinen and Pajunen 1999; Peters, Janzing, and Schölkopf 2017). Recent work

also showed that this factorized marginal Z is necessary to capture "ground truth factors of variation" (Bengio, Courville, and Vincent 2013) in the latent representation Z (Higgins et al. 2017; Kumar and Poole 2020).

In Fig 4.5B, we show the information plots for time, dynamic features, static features and the joint ($\{Y_{dynamic}, Y_{static}\}$) features. X is the retinal activity. Y are features of interest, which are time, $Y_{dynamic}$, Y_{static} and Y_{joint}, shown in different colors. Z are the latent activations of a series of encoder-decoders with different dim Z. Because the encoding of time reaches its full entropy of $H(time)$ (specifically, we observe this saturation at dim $Z > 5$), this shows that the compressed representations Z from the encoder-decoder are near-optimal. In the information bottleneck technique, the best any representation can do is to encode the full entropy of the relevant variable Y, i.e., here $H(Y) = H(time)$. This is similar to the near-optimality shown in previous work in the variational information bottleneck (Alemi et al. 2016; Kolchinsky, Tracey, and Wolpert 2017). In addition, we use many Z's spanning a range of dimensions to construct these information curves. We observe that all of these compressed representations encode comparable amounts of information about both dynamic and static features. This suggests that the dynamic features are as prominent as the static features in the retinal population's internal representation of complex natural scenes.

We highlight the benefit of simultaneous encoding of dynamic and static features in Fig 4.5C. By comparing the information about $Y_{joint} = \{Y_{static}, Y_{dynamic}\}$ and the sum of the information about these two components, we observe that there is synergy $= I(Y_{\text{joint}}; Z) - (I(Y_{\text{static}}; Z) + I(Y_{\text{dynamic}}); Z)$. By construction, the presence of synergistic information corresponds to latent representations with lower dimensions. This suggests why the retina can compress its encoding of both static and dynamic features. When it encodes both features simultaneously, the synergy between these features helps the retina to represent the full entropy of time itself in fewer latent dimensions.

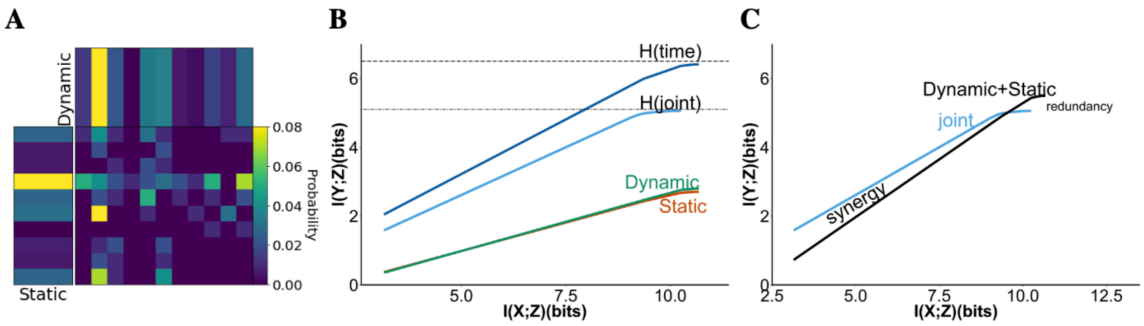

Figure 4.5: **Synergistic features for encoding time in natural scenes** A: Joint distribution of static and dynamic features. This joint distribution includes 76% of the entropy of time. Note that the mutual information between coarse-grained static and dynamic distributions is about 0.8 bits. Given that the entropy of static/dynamic features is around 2.8-2.9 bits, the amount of mutual information between them is relatively small. B: the information plane for fish data. Dark blue: the information curve for encoding time; Light blue: the information curve for encoding the joint distribution combining static and dynamic features; Red/Green: information curves for separated static (red) and dynamic (green) features. See Supplementary Information for information curves of the other two movies C: Blue: the information curve for encoding the joint distribution, the same as B; Black: the sum of information curves from Dynamic(Red)+Static(Green). There is a synergistic region between the information curve for the joint and the sum.

4.5 Discussion

This work uses a U-net-based deep learning architecture to reverse engineer a retinal encoding process for complex natural movies. Using the PSTHs of a large salamander retinal population, we identify stereotypical features that are generalizable across multiple natural movies. We find that the retina uses a transferable, low dimensional representation to encode a rich set of natural space-time features. The encoding obtained from one movie can be used to decode "time in the natural scene" for a different movie, despite differences in their particular spatio-temporal structures. We also discover that the retina encodes time through synergistic coding of both dynamic and static features.

Here, we only observed synergy within the feature space (using mean firing rates of retinal activity, we assume all cells are independent). We also decoded time in its simplest form by asking how well we discriminate between different frames. In future work, we would like to extend our analysis to temporal structure with proper predictive constraints, i.e., predicting a future at a longer Δt should be more challenging than predicting a smaller Δt (Tishby, Pereira, and Bialek 2000; Palmer et al. 2015). We are also aware that the synergy here is different from what can be observed between cells in the neural data. The synergy in the neural code may combine synergy in the feature space with synergy in the population code, itself (Schneidman, Bialek, and Berry 2003; Latham and Nirenberg 2005).

Our work is most similar to (R. Liu et al. 2021; Zhou and Wei 2020) when compared to other methods that also identify a latent representation between brain activity and external stimuli. They used a multilayer perceptron (MLP), a highly expressive feedforward encoder. MLP is fully-connected, so that its learned latent representation corresponds to a single global scale. Our U-net architecture, in contrast to the MLP, employs a ResNet as the encoder. The ResNet encoder attains the same performance as the MLP, but by cascading Resblocks from coarse-to-fine scales. This makes it possible for the U-net architecture to simultaneously learn compressed latent representation at various scales. Although we did not specifically explore this feature, it might be relevant for future research on understand-

ing brain dynamics in flexible natural environments. For example, there is a hierarchy of timescales both in natural scenes and output natural behaviors, ranging from hundreds of milliseconds to minutes (whisking to walking to making action plans (Recanatesi et al. 2022; Stern, Istrate, and Mazzucato 2021)). With additional constraints (Khemakhem et al. 2019), These variational sampling layers may learn hierarchically distinct latent representations for each timescale individually and comprehend how they might be coupled to create complicated behavioral outputs. Outside of neuroscience, This U-net is compatible to learn latent representations between other temporal sequences (e.g., text) and complex spatio-temporal signals (speech or video). Text-to-speech and video summarization are two possible applications. Combining latent representation at multiple scales may also reveal semantic relationships between complex features in general object recognition, e.g., how does a model combine local features (nose, eye) with global shape (e.g., body size) to discriminate between cats and dogs.

Our work shows that the retina leverages feature representations that are common across natural movies. This knowledge transfer differs from what is referred to as "transfer learning" in computer vision and machine learning. In computer vision, transfer learning refers to training a model with a much more complicated dataset (e.g., ImageNet with 1000 classes) and performing inference on a novel, but much smaller dataset (e.g., CIFAR10/100 or CelebA). Transfer learning presupposes that models trained on complex datasets contain sufficient variation to allow the learned features to be reused on new datasets. For the retina, evolutionary timescales underlie the "training from a complex dataset" stage. The retina is shaped in such a way that behaviorally significant components of all natural inputs in an organism's ecological niche are selectively encoded. This enables our training on one movie/retinal response dataset to reveal features transferable to another movie of a similar complexity or scale. Future studies may enable us to determine if such a generalizable feature representation is innate (sculpted only by evolution) or whether visual experience within a lifetime may refine it. This would depend on our ability to track changes in visual processing beyond the

retina (e.g., cortex) over the course of an animal's life (similar to fine-tuning in the transfer learning domain).

CHAPTER 5

LARGE N SCALING OF TIME-DEPENDENT MAXIMUM ENTROPY MODELS

5.1 Introduction

Neuroscience has experienced a profound revolution of measurement: new tools (2-photon imaging, high-density electrode arrays, etc.) have raised the throughput of neural recordings by several orders of magnitude, opening up neural population activity states as a fruitful object of scientific inquiry. This revolution has been accompanied by increasing interest in theoretical tools for understanding neural codes at the level of these large populations (Kastner, Baccus, and Sharpee 2015; Maheswaranathan et al. 2018; Botella-Soler et al. 2018a; Molano-Mazon et al. 2018; Stringer et al. 2019; Marre et al. 2012; Yuste 2015; Berényi et al. 2014; Lopez et al. 2016; Steinmetz et al. 2021). However, the promise of large populations comes with several unique kinds of peril. For any given population size N, the instantaneous activity vector can assume any of 2^N possible binary states; for large N, experiments conducted in finite time are overwhelmingly likely to undersample states that occur infrequently. On top of these sampling issues, whether all parts of these population states are or could be read out by downstream circuits remains unknown. To address these problems, efforts to develop efficient, effective, and principled models of the statistics underlying high-dimensional population activity state distributions have received substantial attention.

One approach to analyzing large populations is the use of maximum entropy models, which successfully use $O(N^2)$ parameters to capture population structure in neural data, even higher-order features not explicitly constrained by the model (Schneidman et al. 2006; Pillow and Simoncelli 2006; Granot-Atedgi et al. 2013; Ganmor, Segev, and Schneidman 2011; Tkačik et al. 2014; Roudi, Nirenberg, and Latham 2009; Jaynes 1957; Tkačik et al. 2010). Understanding the parameters that underlie these models can therefore yield valuable insights on the dynamics of governing large neural populations. However, fitting

maximum entropy models on large datasets has some limitations. With increasing population size, maximum entropy models tend to underestimate the true entropy of a given dataset (Macke, Murray, and Latham 2013). This underestimate grows for datasets with strong correlations. Further, for even an $N = 100$ sized maximum entropy model correlations in the data can be underestimated (Granot-Atedgi et al. 2013). This belies an opportunity to find an alternative approach to fitting large N populations using maximum entropy models.

For any sufficiently large network it is infeasible to read out the entire population activity for any successful decoding mechanism. Given a large N, the high dimensionality of the population limits the viability of population state-based decoding. Such a readout is susceptible to the noise inherent in any neural population. Further, neurons in any given population lack all-to-all connections between them. In reality, there may be "subnetworks" of highly connected neurons that exist within neural populations. Readouts based on these subnetworks and the interactions between them are lower dimensional, and would comprise a more plausible readout than one from the full population state. For these reasons, models that might emphasize or successfully identify subnetwork activity within a full population may be valuable for analysis of large populations.

Inspired by the success of aggregation in machine learning (Breiman 2001; Haussler, Kearns, and Schapire 1994; Monteith et al. 2011) – building a large collection of small models, each of which subsamples a region of the feature space, and combining them to bolster prediction power – here we investigated the effectiveness of aggregating small subnetwork maximum entropy models for approximating a full large-N model of population activity distributions. After a general introduction to the time-dependent maximum entropy approach and to the retinal ganglion cell dataset whose distribution we model, we describe our novel aggregation procedure, experiments comparing aggregate-model fits to fits of the full large-N model, and speculate on approaches for further improving the predictive power and interpretability of our approximation.

5.2 Time dependent maximum entropy modeling

Maximum entropy modeling has a history of success using $O(N^2)$ parameters to capture the structure in neural data, even higher-order features not explicitly constrained by the model (Schneidman et al. 2006; Pillow and Simoncelli 2006; Granot-Atedgi et al. 2013; Ganmor, Segev, and Schneidman 2011; Tkačik et al. 2014; Roudi, Nirenberg, and Latham 2009; Jaynes 1957; Tkačik et al. 2010). We use a time-dependent maximum entropy model that is constrained by the time-varying firing rates averaged across repeated stimuli and the pairwise correlations between cells. Our model takes the form

$$P(\vec{\sigma}^t) = \frac{1}{Z} e^{-\sum_i^N h_i^t \sigma_i^t - \sum_{i<j}^N J_{ij} \sigma_i^t \sigma_j^t}, \tag{5.1}$$

with constraints on $\langle \sigma_i^t \rangle_k$, which captures each cell's individual response to the stimulus at time t averaged over trials, k, and $\langle \sigma_i^t \sigma_j^t \rangle_{t,k}$, the correlations between cells. These two constraints map to two sets of parameters, the time-dependent fields h_i^t and the static couplings J_{ij}, respectively.

5.3 Data

Voltage traces from the RGC layer of a larval tiger salamander retina were recorded following the methods outlined in (Marre et al. 2012). In brief, retina from a larval tiger salamander was isolated in darkness and pressed against a 252 channel multielectrode array. Multiple repeats (minimum 80) from five different natural movies are displayed in a pseudorandom order. Recordings were taken during stimulus presentation and spike sorted using a mostly automated spike sorting algorithm. This technique captured a highly overlapping neural population of 93 cells that fully tiled a region of visual space. Spikes were binned at 60Hz for all analyses presented. The following work uses only one of the five movies for analysis.

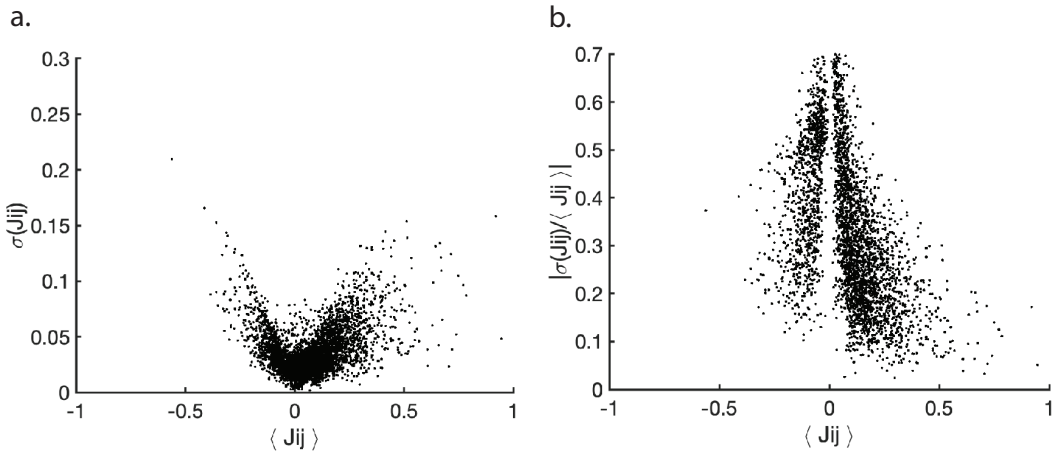

Figure 5.1: **Stability of couplings across multiple fits.** a) The average coupling weight and its standard deviation over all couplings in the dataset. b) The coefficient of variation for couplings against many views. The coefficient of variation explodes as the couplings approach 0, but variability is low particularly for strong, sparse couplings.

5.4 Aggregate approach

We build our large N models by first noticing an interesting property of time depending maximum entropy models. In retina, a learned coupling J_{ij} in a time dependent maximum entropy model largely remains stable regardless of the other neurons fit within any given group. For example, if a time dependent maximum entropy is fit on group A with participating cells $\{a, b, ..., k\}$, the learned coupling J_{ij} will be very similar to the learned couplings J_{ij} fit on group B with participants $\{i, j, l, m, n, ...t\}$. This stability might arise from the fact that learned time dependent maximum entropy couplings appear to arise from anatomical properties of the retina (Hoshal et al. 2023). This stability is not present in the standard Ising model (see Schneidman et al. 2006) that has been commonly used to interrogate this kind of neural data. The stability of couplings is demonstrated in Figure 5.1. In particular, the coefficient of variation decreases for the sparse, strong couplings.

We leverage this coupling property to build to large N by aggregating together a large number of smaller N models. We find that $N = 20$ cell groups consistently fit well with very low runtimes, making them a strong candidate for the small N fitting. We fit $m = 1000$ small N groups by randomly choosing $N = 20$ cells from the $N = 93$ cells in the dataset. Each $N = 20$ cell group yields $N * (N - 1)/2 = 190$ couplings, so fitting $k = 1000$ models means each of the $93 * 92/2 = 4278$ couplings have been fit an average of 44 times. We then take each coupling as the average coupling value over all J_{ij} fits and use these couplings to build an aggregate model. A similar approach is taken for fitting the fields h_i^t.

We also generate a sparse version of the aggregate model. The necessity of this sparse model arises from previous results in Chapter 2 showing the sparse couplings to be behaviorally relevant for stimulus discrimination in salamander retina. A sparse aggregate model then only contains the strong couplings that have a strong effect on the behavior of the time-dependent maximum entropy model. To build the sparse aggregate model, we take the same $m = 1000$ fits for the aggregate model, but keep only the top 10% of couplings in each $N = 20$ cell group. All other couplings are set to zero. We then aggregate the model in the same way as described above. In the process of taking $_{ij}\rangle$ for this sparse model, many zero values can be folded into a coupling that is "strong" in one group but outside of the top 10% of couplings in another. This dampens couplings that appear strong in any local $N = 20$ subgroup but on average are not the among the strongest couplings for the global $N = 93$ data.

5.5 Results

Comparing the aggregate approach to a full model

We evaluate the aggregate time dependent maximum entropy modeling approach by comparing it to a model fit on the entire $N = 93$ cell dataset. While similar couplings between models might indicate success of the aggregate approach, it is not a requirement. In particular, large N maximum entropy models fit on data with strong correlations have

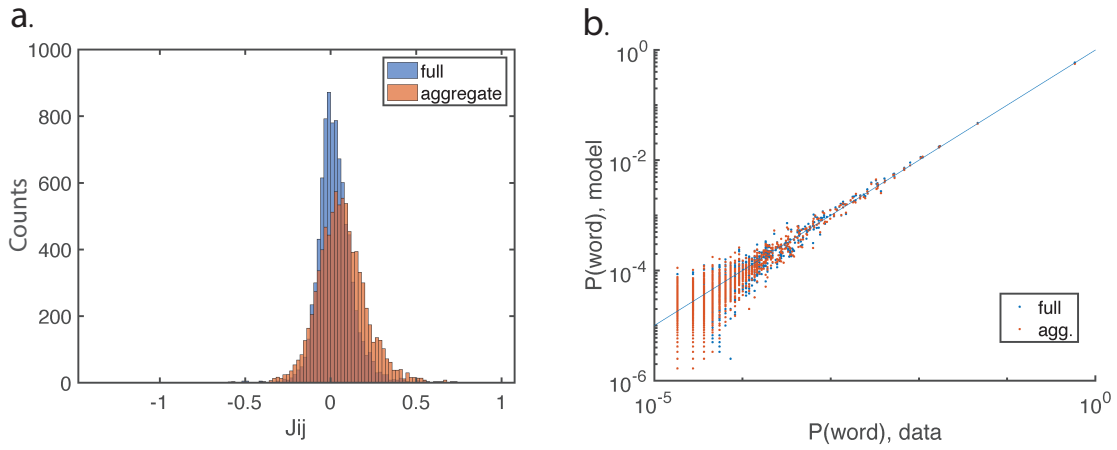

Figure 5.2: **Comparison between full fit and aggregate models** Large models are fit either directly ('full model') or by aggregating $N = 20$ cell groups ('aggregate model'). a) Coupling strengths in the full and aggregate models. The aggregate model has slightly higher mean and heavier tail than the full model. b) Word probability plots for both models plotted against data. Both models capture the observed word distribution, but the aggregate model performs slightly better in the low data regime.

been shown to underestimate the empirical entropy of the dataset (Macke, Murray, and Latham 2013). Further, stimulus dependent maximum entropy models, a close cousin of time dependent maximum entropy models, have been shown to slightly underestimate pairwise correlations for an $N = 100$ model (Granot-Atedgi et al. 2013). While this underestimate is small, it might have an effect on how well large N models predict population activity.

In Figure 5.2a, we compare the couplings fit from the aggregate model and the full fit of the population. There are two significant differences between the coupling distributions. Firstly, the aggregate model has an overall higher mean than the full model. As couplings J_{ij} in time dependent models capture the noise correlations between cells, this difference might come from small N models estimating a stronger noise correlation than larger N models. The larger mean arises from systemically higher coupling values in the small N models compared to fitting the model outright. Secondly, the aggregate model has a significantly heavier tail of coupling strengths compared to the full model. This is of particular interest as these sparse, strong couplings has been shown to improve discriminability between different natural scenes

(Hoshal et al. 2023). That means that the primary differences in coupling strength between these models arises from a behaviorally relevant set of couplings.

We then compare how well both the aggregate and full model capture the population responses present in the data. This is done by generating word-word probability plots, where each "word" in the data represents one of the 2^N possible states the neural population can take. We find that both the full and aggregate model well approximate the population responses in the data. Interestingly, as words in the data become less frequent (moving towards the bottom left corner of Figure 5.2b), the aggregate model performs slightly better than the full model, staying closer to the unity line until we approach the realm of data unreliability. This is where there are too few samples from the data to get a reliable estimate of the word probability. In Figure 5.2b, 10^{-5} represents only a single appearance of the word in the data.

A sparse aggregate model shows promise in capturing higher order interactions

In Figure 5.3, we evaluate the ability of the full and aggregate models to capture higher order interactions in the data. While time-dependent maximum entropy models are not constrained to capture higher order interactions, they have been shown to capture structure of neural data not explicitly constrained by the models (Schneidman et al. 2006; Pillow and Simoncelli 2006; Granot-Atedgi et al. 2013; Ganmor, Segev, and Schneidman 2011; Tkačik et al. 2014; Roudi, Nirenberg, and Latham 2009; Jaynes 1957; Tkačik et al. 2010). One such representation of higher-order statistics are triplet interactions, or how well our models capture how three cells interact with each other compared to the data 5.3. Interestingly, the full and aggregate model diverge in their descriptions of triplets, where the full model underestimates triplet interactions and the aggregate model largely overestimates them. The overestimate of triplets in the aggregate model might arise from the systemically high estimate of couplings in the small N models compared to the full model. If this is the case, a sparse aggregate model, which dampens some medium strength couplings in the aggregate model while not affecting the strongest couplings, might better predict triplet interactions.

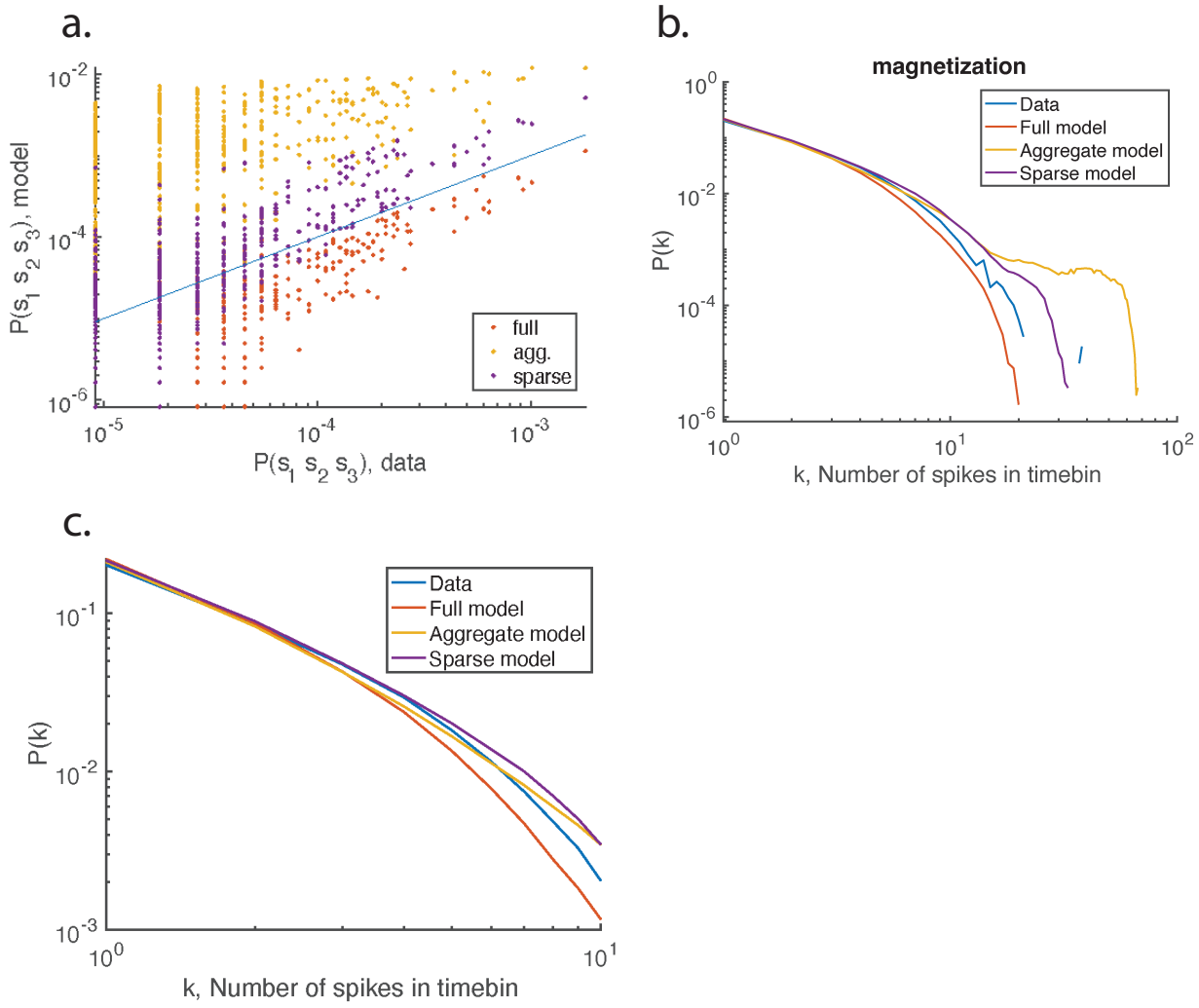

Figure 5.3: **Higher order interactions in time dependent maximum entropy models.** a) Triplet interactions for the full, aggregate, and spare aggregate models compared to the data. The full and aggregate models under- and overestimate the true triplet interactions, respectively. The sparse aggregate model lowers the overestimate of triplet interactions relative to the aggregate model. b) Magnetization for the full, aggregate, and sparse aggregate models. Models approximate the data well for lower spiking words. c) Magnetization up to $k = 10$ spikes.

Indeed, the sparse model still overestimates triplet interactions, but largely improves upon the estimate from the aggregate model.

Another estimate of higher order interactions is the magnetization, which measures the amount of spins that point in the same direction. In neuroscience, this translates to measuring the probability of having k-spike words in the data and comparing across different models. As neural data is inherently sparse, we expect magnetization to be most relevant in a smaller k-spike regime. Up to $k = 10$ spikes, both the full and aggregate models predict the probability of k-spike words in the data well, with the full model slightly underestimating and the aggregate model slightly overestimating $P(k)$ in the data. Like in the word-word probability plots in Figure 5.2b, these estimates begin to fail in the low data regime. Even so, the sparse aggregate model performs better in this low data regime relative to the aggregate model, indicating a its potential for fitting large N models.

5.6 Future directions

The current results in building the aggregate model indicate a promising baseline for future work. Many open questions surrounding the aggregate model revolve around the systemic high estimate of couplings J_{ij} relative to a fully fit large N model. As shown in Figure 5.4, the difference between the full model and aggregate couplings may be accounted for by some linear scaling factor. One way to determine that factor would be via the use of a synthetic dataset. Such a dataset would have ground truth coupling values with which to evaluate the full and aggregate models. Of particular interest would be determining whether the full model underestimates ground truth couplings or whether the aggregate model overestimates them. In either case, discovering a true scaling factor between the aggregate model and full model couplings provides a viable post-fitting correction term. Determining this scaling factor would also help in evaluating whether a correction for the aggregate model is even necessary.

The success of the sparse aggregate model in estimating triplet interactions relative to the

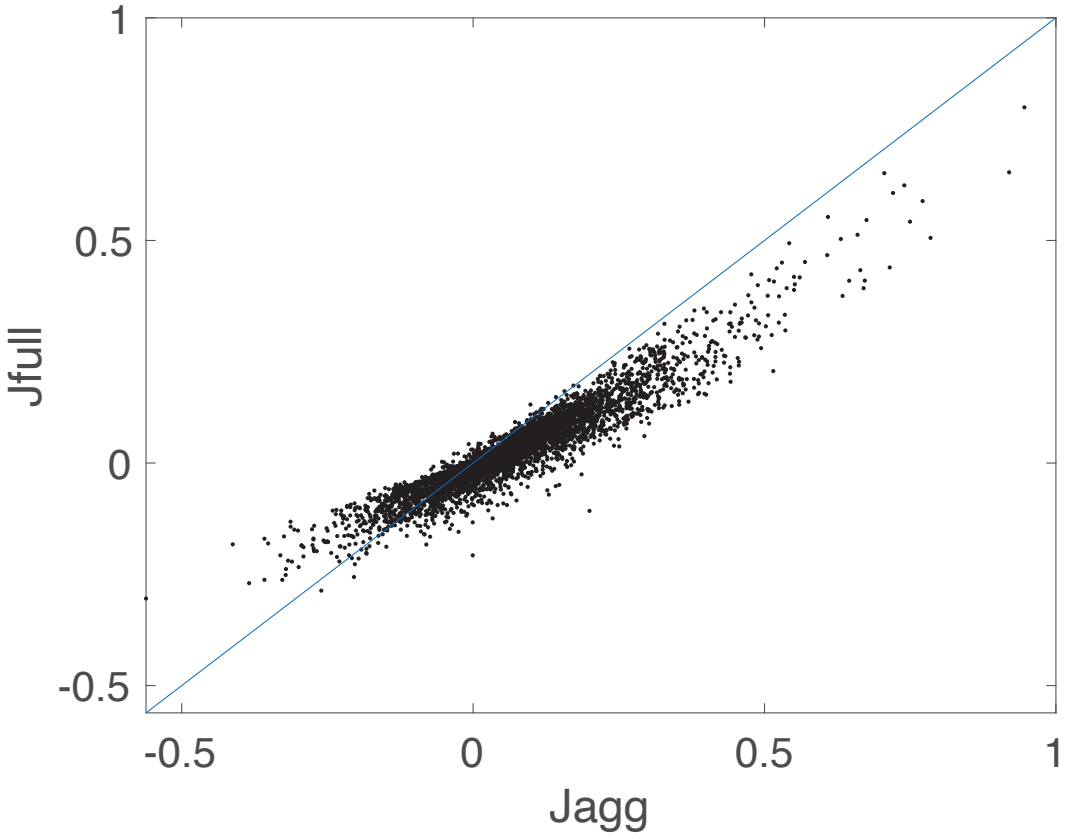

Figure 5.4: **Coupling comparison between full and aggregate models.** The comparison between full and aggregate coupling fits between models indicates a systemically higher estimate of couplings in the aggregate model. The appears to be a linear scaling factor between the full and aggregate model coupling estimates.

aggregate model provides another interesting avenue for further work. There are many viable sparsening schemes other than the one presented here, including sparsening the aggregate model after aggregating all small N models instead of doing so for each individual model fit. This would yield a true lower-parameter model, with the total number of couplings included being determined by the strictness of the sparsening criterion. Currently, sparsening is empirically set to be 10% of all couplings, but the use of a synthetic dataset, along with fitting many models with varying population size N, might yield a more theoretically based sparsening criterion.

The aggregate model as presented works well in retina, which is largely a feed forward network. It remains to be seen whether this approach can be taken for large N modeling in cortex, where there are large amounts of recurrence. One can imagine that this recurrence destabiliizes the discovered J_{ij} couplings for small N groups, as the coupling value may shift depending on whether or not the smaller N group contains all the cells participating in some recurrence loop. For example, if cells A, B, and C participate in a recurrance loop, estimating couplings by fitting the coupling J_{AB} in a group with all of A, B and C versus just A and B might yield different values. Testing the success of an aggregate method in cortex would provide clarity on whether this approach is robust to recurrent population structures.

5.7 Discussion

This work explores a novel aggregate approach to large N maximum entropy models. This leverages an interesting property of time-dependent maximum entropy models in retina: the fit couplings tend to remain very stable, even if they are fit multiple times within different subgroups of the population. This stability allows the aggregation of numerous models into a larger N model. We demonstrate that this model captures population activity roughly as well as the normal fitting procedure.

In general, maximum entropy models have had success fitting higher order interactions not explicitly constrained by the model (Schneidman et al. 2006; Pillow and Simoncelli 2006;

Granot-Atedgi et al. 2013; Ganmor, Segev, and Schneidman 2011; Tkačik et al. 2014; Roudi, Nirenberg, and Latham 2009; Jaynes 1957; Tkačik et al. 2010). At the population state level, this is true for both the aggregate and full fit models. Even so, some higher-order interactions, like triplets, are not well described. A sparse aggregate model shows promise in capturing these higher order interactions by emphasizing only the strongest couplings in the model. Future work should explore other sparsening schemes that might capture these higher order interactions.

Of particular interest is the potential for a linear scaling difference for couplings between the large N and small N models. The use of synthetic data with ground truth couplings may help uncover whether this scaling exists, and in which direction any correction to couplings needs to be made. A first order check might be the total coupling strength $||J||$ per number of spins included in the model. If this value remains consistent across different population sizes, it might explain how individual coupling values differ when fit in large N versus small N groups.

CHAPTER 6

CONCLUSION

In this thesis, we explore population based retinal coding strategies in natural scenes. We demonstrate the presence of a stimulus invariant population structure that remains static for both natural and synthetic stimuli. We show that leveraging this structure improves scene discriminability over an independent coding scheme. Further, this fixed structure depends largely on a few, sparse cell-cell interactions, which might be easier for downstream brain areas to read out. We explore readout mechanisms that can leverage this sparse population by implementing graph based neural networks. These networks are able to cluster different natural scenes in an unsupervised manner, and are able to discriminate even single trial neural activity from different natural scenes. They also learn an encoding space that can accommodate and separate single-trial neural responses from a natural scene not included during training. This kind of zero-shot encoding of novel stimuli is potentially ethologically relevant for behavior, as organisms will encounter novel environments through their lives that they must accurately perceive. We then explore what features of natural scenes the retina encodes by building an encoder-decoder tasked with reconstructing a future frame of a natural scene based on the population neural responses to that stimulus. By compressing the neural response into a low dimensional space and decoding a biologically relevant feature, we encourage the encoder-decoder to efficiently compress the relevant features encoded by the neural activity. We find that the retina learns a low-dimensional, generalizable representation of natural scenes. The retina responds to spatio-temporal features in natural scenes in order to do scene readout. These features can be split into static and dynamic motifs, which are synergistic with respect to the encoding. Lastly, we explore an aggregate approach to large N models of neural populations. We show that an aggregate model recapitulates population level neural activity, and that sparsening an aggregate model has promise for matching higher order interactions present in the data.

Chapter 2 demonstrates that couplings between cells in a neural population are an impor-

tant component of downstream readout of scene identity. It is possible that an independent readout preserves the majority of information available in the retinal population while failing to effectively convey critical features of the visual scene. Natural scenes probe a behaviorally relevant context to assess the impact of noise correlations on neural coding. These movies, like other natural inputs, drive a richer and more reliable code in the brain (Rikhye and Sur 2015; Froudarakis et al. 2014; Hasson, Malach, and Heeger 2010). Comparing across movies reveals what the more subtle features in the neural code might be used for. We find that sparse interactions sufficiently capture the functional impact of noise correlations. These sparse couplings are the key factor for efficient scene identification. A sparse backbone may be easier to implement and read out downstream. On the flip side, sparse codes might hamstring error correction (Puchalla et al. 2005; Ganmor, Segev, and Schneidman 2015), so future work should explore how these costs and benefits trade-off for behaviorally relevant inputs and tasks.

Noise correlations have a large effect on scene decoding, which may arise from small effects aggregated over time. It is not clear from the analysis performed here what precisely gives rise to the beneficial impacts of noise correlations on decoding. One possible answer is that the noise correlations may reflect changes in scene correlation structure. This may help recover scene specific information that is otherwise lost to single-cell-level adaptation.

Unraveling how this sparse but strong structure in the code is mechanistically supported is an important next step in this work. In some ways, the circuit structure in the eye differs from that found in the cortex. The retina is not a recurrent neural network; RGCs do not have direct synaptic coupling, and the photoreceptor-to-RGC circuit is largely feed-forward. To create a population code with sparse interactions, the retina needs to be wired around these structural constraints. These sparse interactions seem to be the result of common bipolar inputs and gap junction coupling between RGCs. What we have observed is sparse, strong, functionally important, exclusively non-synaptic RGC-RGC couplings. Both gap junctions and common bipolar inputs lead to stronger coupling between cells, but our analysis is not

sensitive enough to tease apart whether these two types of coupling sources are mutually exclusive. Exclusivity would be an efficient way to implement a sparse backbone of specific cell-cell interactions. Future work to disentangle the circuit mechanisms giving rise to the sparse backbone might ultimately inform studies in cortex where gap junction coupling is also present (Friend and Gilula 1972; Peinado, Yuste, and Katz 1993; Y. Li et al. 2012).

In Chapter 3, we demonstrate that activity of retinal populations during the presentation of natural scenes, informed by functional connections between the cells, can support the emergent learning of an embedding space that distinguishes between different scenes, and is robust to noise on single trials. This is facilitated by noise correlations between cells that are consistent across wildly different scenes, and which are thought to reflect a static underlying connectivity structure. Although our model could use these cell-cell interactions to support the unsupervised learning of this embedding space, future work will be required to validate the necessity of these interactions for single-trial scene decoding.

The graph-based representation learning scheme we used relied on a relatively high-dimensional ($dim = 80$) embedding space for separating between activity vectors corresponding with each scene. However, an embedding space of this dimensionality may not the basis for the retina's encoding of natural scenes. Although the possible dimensionality of retinal encodings could scale linearly with the number of RGCs N, in reality this scaling is limited by the pairwise correlations between cells, and higher-dimensional encoding schemes are known to present problems for robust, generalizable readout that lower-dimensional schemes ameliorate. Retinal responses in particular are known to be compressible to a low-dimensional encoding space (e.g., 10 dimensions, as in (Wang et al. 2022)). Resolving how graph-informed representations separate between scenes on the basis of single-trial activity vectors even when using lower-dimensional representations constitutes an important extension for additional investigation.

Future work should expand on the results presented here in several key directions. First, choosing an appropriate null model for this setting is both important and technically difficult.

An ideal control, fully accounting for any possible difference in embedding model approach, parameter count, etc., would use the same GCN structure for learning the embedding space, but under different perturbations of the edge structure that governs message passing. However, our implementation of the GraphCL algorithm precluded the use of a no-edge model (learning the embedding space requires contrastive learning on augmented subgraphs, which cannot exist wiithout any edges), and is too flexible for an 'all-edges-equal' model to be a relevant comparison. In principle a node-dropping augmentation rather than a subgraph augmentation could ameliorate these shortcomings in our setting.

Chapter 4 uses a U-net-based deep learning architecture to reverse engineer a retinal encoding process for complex natural movies. Using the PSTHs of a large salamander retinal population, we identify stereotypical features that are generalizable across multiple natural movies. We find that the retina uses a transferable, low dimensional representation to encode a rich set of natural space-time features. The encoding obtained from one movie can be used to decode "time in the natural scene" for a different movie, despite differences in their particular spatio-temporal structures. We also discover that the retina encodes time through synergistic coding of both dynamic and static features.

Here, we only observed synergy within the feature space (using mean firing rates of retinal activity, we assume all cells are independent). We also decoded time in its simplest form by asking how well we discriminate between different frames. In future work, we would like to extend our analysis to temporal structure with proper predictive constraints, i.e., predicting a future at a longer Δt should be more challenging than predicting a smaller Δt (Tishby, Pereira, and Bialek 2000; Palmer et al. 2015). We are also aware that the synergy here is different from what can be observed between cells in the neural data. The synergy in the neural code may combine synergy in the feature space with synergy in the population code, itself (Schneidman, Bialek, and Berry 2003; Latham and Nirenberg 2005).

Our work is most similar to (R. Liu et al. 2021; Zhou and Wei 2020) when compared to other methods that also identify a latent representation between brain activity and ex-

ternal stimuli. They used a multilayer perceptron (MLP), a highly expressive feedforward encoder. MLP is fully-connected, so that its learned latent representation corresponds to a single global scale. Our U-net architecture, in contrast to the MLP, employs a ResNet as the encoder. The ResNet encoder attains the same performance as the MLP, but by cascading Resblocks from coarse-to-fine scales. This makes it possible for the U-net architecture to simultaneously learn compressed latent representation at various scales. Although we did not specifically explore this feature, it might be relevant for future research on understanding brain dynamics in flexible natural environments. For example, there is a hierarchy of timescales both in natural scenes and output natural behaviors, ranging from hundreds of milliseconds to minutes (whisking to walking to making action plans (Recanatesi et al. 2022; Stern, Istrate, and Mazzucato 2021)). With additional constraints (Khemakhem et al. 2019), These variational sampling layers may learn hierarchically distinct latent representations for each timescale individually and comprehend how they might be coupled to create complicated behavioral outputs. Outside of neuroscience, This U-net is compatible to learn latent representations between other temporal sequences (e.g., text) and complex spatio-temporal signals (speech or video). Text-to-speech and video summarization are two possible applications. Combining latent representation at multiple scales may also reveal semantic relationships between complex features in general object recognition, e.g., how does a model combine local features (nose, eye) with global shape (e.g., body size) to discriminate between cats and dogs.

Our work shows that the retina leverages feature representations that are common across natural movies. This knowledge transfer differs from what is referred to as "transfer learning" in computer vision and machine learning. In computer vision, transfer learning refers to training a model with a much more complicated dataset (e.g., ImageNet with 1000 classes) and performing inference on a novel, but much smaller dataset (e.g., CIFAR10/100 or CelebA). Transfer learning presupposes that models trained on complex datasets contain sufficient variation to allow the learned features to be reused on new datasets. For the retina, evolu-

tionary timescales underlie the "training from a complex dataset" stage. The retina is shaped in such a way that behaviorally significant components of all natural inputs in an organism's ecological niche are selectively encoded. This enables our training on one movie/retinal response dataset to reveal features transferable to another movie of a similar complexity or scale. Future studies may enable us to determine if such a generalizable feature representation is innate (sculpted only by evolution) or whether visual experience within a lifetime may refine it. This would depend on our ability to track changes in visual processing beyond the retina (e.g., cortex) over the course of an animal's life (similar to fine-tuning in the transfer learning domain).

Chapter 5 investigates aggregation-based strategies for extending maximum entropy models to the large-N regime. The approach we introduce takes advantage of the tendency of cell-cell couplings to be consistent, likely reflective of anatomical connections that are consistent across time/stimulus presentations. This consistency manifests even when viewing subsets of the whole simultaneously recorded population, suggesting that it may be amenable to use in a repeatedly-subsample-then-aggregate model combination scheme. We used this idea to build many small N models into one large N model, while eschewing the direct fit of the large N model itself entirely. We further improve this process by developing a sparse-aggregate extension model. This leverages the asymmetrical distribution of coupling prominence/importance discovered in Chapter 2 as the rationale for focusing on a sparse subnetwork. This sparsening step improves on the tendency of aggregate models to mis-estimate some high-order interactions between cells.

Of particular interest is the potential for a linear scaling difference for couplings between the large N and small N models. The use of synthetic data with ground truth couplings may help uncover whether this scaling exists, and in which direction any correction to couplings needs to be made. A first order check might be the total coupling strength $||J||$ per number of spins included in the model. If this value remains consistent across different population sizes, it might explain how individual coupling values differ when fit in large N versus small

N groups.

The projects that together comprise this thesis coalesce around several high-level themes of retinal coding. First, they underline the importance of population structure, on top of any independent coding scheme adopted by the population, for information coding. These benefits can arise through any of several avenues – covariation between pairs of cells, higher-order relationships, etc. – but in general promise a richer code for downstream circuits elsewhere in the brain to mine for the elaboration of higher-order representations. Second, the results discussed here highlight the importance of static structure for readout. This theme is especially important in the retina, where strong input adaptation is a feature rather than a bug, even in response to different stimuli with highly varying underlying statistics. Third, in the presence of bandwidth bottlenecks like the retina, information encoding should be as parsimonious as possible. Not everything about a stimulus can be encoded, and an effective initial processor should elide any information that downstream circuits will not need to support survival-relevant behaviors. Fourth, the retina's encoding generalizes across stimulus contexts. This is a desirable feature for a sensory bottleneck to have. Given that all visual information that the brain can make use of is processed through the retina, the retina must be robust to whatever level of change in stimulus statistics the visual world might manifest in order for the organism to perform any visually-informed behavior. Even if an organism's visual processing circuits beyond the retina were highly sophisticated, if the retinal encoding were too context-specific, the organism would face a profound adaptive disadvantage, failing to make sense of the environment outside the center of its niche.

These projects have together serve to contextualize the retinal population coding of natural scenes. This constitutes a specific instance of the central problem of systems/computational neuroscience – to understand how populations of cells make use of their available bandwidth to signal information to their downstream partners. Core to building this understanding is a careful consideration of (a) what information those downstream partners need access to (what must be decoded?), which must reference the behavioral goals and ecological niche

that the animal occupies; and (b) given what the code must convey, how the code should be formatted.

REFERENCES

Ablin, Pierre, Jean-Francois Cardoso, and Alexandre Gramfort. 2018. "Faster independent component analysis by preconditioning with Hessian approximations." *IEEE Transactions on Signal Processing* 66 (15): 4040–4049.

Alemi, Alexander A., Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2016. "Deep Variational Information Bottleneck." *Proceedings of the International Conference on Learning Representations (ICLR) 2017* (December). arXiv: 1612.00410 [cs.LG].

Asari, Hiroki, and Markus Meister. 2012. "Divergence of visual channels in the inner retina." *Nature neuroscience* 15 (11): 1581–1589.

Asari, Hiroki, and Markus Meister. 2014. "The projective field of retinal bipolar cells and its modulation by visual context." *Neuron* 81 (3): 641–652.

Atick, Joseph J, and A Norman Redlich. 1992a. "What does the retina know about natural scenes?" *Neural computation* 4 (2): 196–210.

Atick, Joseph J., and A. Norman Redlich. 1992b. "What Does the Retina Know about Natural Scenes?" 4:196–210. ISSN: 0899-7667. https://doi.org/10.1162/neco.1992.4.2.196.

Averbeck, Bruno B, Peter E Latham, and Alexandre Pouget. 2006. "Neural correlations, population coding and computation." *Nature reviews neuroscience* 7 (5): 358–366.

Baccus, Stephen A, Bence P Ölveczky, Mihai Manu, and Markus Meister. 2008. "A retinal circuit that computes object motion." *Journal of Neuroscience* 28 (27): 6807–6817.

Baccus, Stephen A., and Markus Meister. 2002. "Fast and slow contrast adaptation in retinal circuitry." Ppublish, *Neuron* 36 (5): 909–919. ISSN: 0896-6273. https://doi.org/10.1016/s0896-6273(02)01050-4.

Barlow, H. B. 1961. "Possible Principles Underlying the Transformations of Sensory Messages." *MIT Press,* https://doi.org/10.7551/mitpress/9780262518420.003.0013.

Barlow, Horace B. 1953. "Summation and inhibition in the frog's retina." *The Journal of physiology* 119 (1): 69.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation learning: a review and new perspectives." Ppublish, *IEEE transactions on pattern analysis and machine intelligence* 35 (8): 1798–1828. ISSN: 1939-3539. https://doi.org/10.1109/TPAMI.2013.50.

Berényi, Antal, Zoltán Somogyvári, Anett J Nagy, Lisa Roux, John D Long, Shigeyoshi Fujisawa, Eran Stark, Anthony Leonardo, Timothy D Harris, and György Buzsáki. 2014. "Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals." *Journal of neurophysiology* 111 (5): 1132–1149.

Berry, Michael J, Iman H Brivanlou, Thomas A Jordan, and Markus Meister. 1999. "Anticipation of moving stimuli by the retina." *Nature* 398 (6725): 334–338.

Berry, Michael J., David K. Warland, and Markus Meister. 1997. "The structure and precision of retinal spike trains." *Proceedings of the National Academy of Sciences* 94, no. 10 (May): 5411–5416. https://doi.org/10.1073/pnas.94.10.5411.

Bialek, William, Fred Rieke, Robert van Steveninck, and David Warland. 1989. "Reading a neural code." *Advances in neural information processing systems* 2.

Bialek, William, and Naftali Tishby. 1999. "Predictive Information" (February). arXiv: cond-mat/9902341 [cond-mat.stat-mech].

Botella-Soler, Vicente, Stéphane Deny, Georg Martius, Olivier Marre, and Gašper Tkačik. 2018a. "Nonlinear decoding of a complex movie from the mammalian retina." *PLoS computational biology* 14 (5): e1006057.

Botella-Soler, Vicente, Stéphane Deny, Georg Martius, Olivier Marre, and Gašper Tkačik. 2018b. "Nonlinear decoding of a complex movie from the mammalian retina." Edited by Aldo A Faisal. *PLOS Computational Biology* 14, no. 5 (May): e1006057. https://doi.org/10.1371/journal.pcbi.1006057.

Bowman, Samuel R., Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. 2015. "Generating Sentences from a Continuous Space." *SIGNLL Conference on Computational Natural Language Learning (CONLL), 2016* (November). arXiv: 1511.06349 [cs.LG].

Brackbill, Nora, Colleen Rhoades, Alexandra Kling, Nishal P Shah, Alexander Sher, Alan M Litke, and EJ Chichilnisky. 2020. "Reconstruction of natural images from responses of primate retinal ganglion cells." *eLife* 9 (November). https://doi.org/10.7554/elife.58516.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45:5–32.

Brivanlou, Iman H, David K Warland, and Markus Meister. 1998. "Mechanisms of concerted firing among retinal ganglion cells." *Neuron* 20 (3): 527–539.

Burkhardt, Dwight A, Patrick K Fahey, and Michael A Sikora. 2006. "Natural images and contrast encoding in bipolar cells in the retina of the land-and aquatic-phase tiger salamander." *Visual neuroscience* 23 (1): 35–47.

Chen, Eric Y, Janice Chou, Jeongsook Park, Greg Schwartz, and Michael J Berry. 2014. "The neural circuit mechanisms underlying the retinal response to motion reversal." *Journal of Neuroscience* 34 (47): 15557–15575.

Chen, Eric Y, Olivier Marre, Clark Fisher, Greg Schwartz, Joshua Levy, Rava Azeredo da Silveira, and Michael J Berry. 2013. "Alert response to motion onset in the retina." *Journal of Neuroscience* 33 (1): 120–132.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. "A simple framework for contrastive learning of visual representations." In *International conference on machine learning,* 1597–1607. PMLR.

Darmois, G. 1953. "Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire." *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 21 (1/2): 2. https://doi.org/10.2307/1401511.

Demb, Jonathan B. 2008. "Functional circuitry of visual adaptation in the retina." *The Journal of physiology* 586 (18): 4377–4384.

Ding, Jennifer, Albert Chen, Janet Chung, Hector Acaron Ledesma, Mofei Wu, David M Berson, Stephanie E Palmer, and Wei Wei. 2021. "Spatially displaced excitation contributes to the encoding of interrupted motion by a retinal direction-selective circuit." *Elife* 10:e68181.

Dong, Dawei W, and Joseph J Atick. 1995. "Statistics of natural time-varying images." *Network: Computation in Neural Systems* 6 (3): 345.

Fairhall, Adrienne L, C Andrew Burlingame, Ramesh Narasimhan, Robert A Harris, Jason L Puchalla, and Michael J Berry. 2006. "Selectivity for multiple stimulus features in retinal ganglion cells." *Journal of neurophysiology* 96 (5): 2724–2738.

Fairhall, Adrienne L, Geoffrey D Lewen, William Bialek, and Robert R de Ruyter van Steveninck. 2001. "Efficiency and ambiguity in an adaptive neural code." *Nature* 412 (6849): 787–792.

Fang, Linjing, Fred Monroe, Sammy Weiser Novak, Lyndsey Kirk, Cara R. Schiavon, Seungyoon B. Yu, Tong Zhang, et al. 2021. "Deep learning-based point-scanning super-resolution imaging." Ppublish, *Nature methods* 18 (4): 406–416. ISSN: 1548-7105. https://doi.org/10.1038/s41592-021-01080-z.

Ferrari, Ulisse. 2016. "Learning maximum entropy models from finite-size data sets: A fast data-driven algorithm allows sampling from the posterior distribution." *Physical Review E* 94 (2): 023301.

Ferrari, Ulisse, Stéphane Deny, Matthew Chalk, Ga šper Tka čik, Olivier Marre, and Thierry Mora. 2018. "Separating intrinsic interactions from extrinsic correlations in a network of sensory neurons." *Phys. Rev. E* 98 (4): 042410. https://doi.org/10.1103/PhysRevE.98.042410. https://link.aps.org/doi/10.1103/PhysRevE.98.042410.

Friend, Daniel S, and Norton B Gilula. 1972. "Variations in tight and gap junctions in mammalian tissues." *The Journal of cell biology* 53 (3): 758–776.

Froudarakis, Emmanouil, Philipp Berens, Alexander S Ecker, R James Cotton, Fabian H Sinz, Dimitri Yatsenko, Peter Saggau, Matthias Bethge, and Andreas S Tolias. 2014. "Population code in mouse V1 facilitates readout of natural scenes through increased sparseness." *Nature neuroscience* 17 (6): 851–857.

Ganmor, Elad, Ronen Segev, and Elad Schneidman. 2011. "Sparse low-order interaction network underlies a highly correlated and learnable neural population code." *Proceedings of the National Academy of sciences* 108 (23): 9679–9684.

Ganmor, Elad, Ronen Segev, and Elad Schneidman. 2015. "A thesaurus for a neural population code." *Elife* 4:e06134.

Goldin, Matías A, Baptiste Lefebvre, Samuele Virgili, Mathieu Kim Pham Van Cang, Alexander Ecker, Thierry Mora, Ulisse Ferrari, and Olivier Marre. 2022. "Context-dependent selectivity to natural images in the retina." *Nature Communications* 13 (1): 5556.

Gollisch, Tim. 2013. "Features and functions of nonlinear spatial integration by retinal ganglion cells." *Journal of Physiology-Paris* 107 (5): 338–348.

Gollisch, Tim, and Markus Meister. 2008. "Rapid neural coding in the retina with relative spike latencies." *science* 319 (5866): 1108–1111.

Gollisch, Tim, and Markus Meister. 2010. "Eye smarter than scientists believed: neural computations in circuits of the retina." *Neuron* 65 (2): 150–164.

Granot-Atedgi, Einat, Gašper Tkačik, Ronen Segev, and Elad Schneidman. 2013. "Stimulus-dependent maximum entropy models of neural population codes." *PLoS computational biology* 9 (3): e1002922.

Hartline, Haldan K. 1940. "The receptive fields of optic nerve fibers." *American Journal of Physiology-Legacy Content* 130 (4): 690–699.

Hasson, Uri, Rafael Malach, and David J Heeger. 2010. "Reliability of cortical activity during natural stimulation." *Trends in cognitive sciences* 14 (1): 40–48.

Hateren, J Hans van, and Dan L Ruderman. 1998. "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265 (1412): 2315–2320.

Haussler, David, Michael Kearns, and Robert E Schapire. 1994. "Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension." *Machine learning* 14:83–113.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, June. https://doi.org/10.1109/cvpr.2016.90.

Higgins, Irina, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework." In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.* OpenReview.net. https://openreview.net/forum?id=Sy2fzU9gl.

Hinton, Geoffrey E, and Richard Zemel. 1993. "Autoencoders, Minimum Description Length and Helmholtz Free Energy." In *Advances in Neural Information Processing Systems,* edited by J. Cowan, G. Tesauro, and J. Alspector, vol. 6. Morgan-Kaufmann. https://proceedings.neurips.cc/paper/1993/file/9e3cfc48eccf81a0d57663e129aef3cb-Paper.pdf.

Hochstein, S, and RM Shapley. 1976. "Quantitative analysis of retinal ganglion cell classifications." *The Journal of physiology* 262 (2): 237–264.

Horn, Berthold K.P., and Brian G. Schunck. 1981. "Determining optical flow." *Artificial Intelligence* 17, nos. 1-3 (August): 185–203. https://doi.org/10.1016/0004-3702(81)90024-2.

Hoshal, Benjamin, Caroline M Holmes, Kyle Bojanek, Jared M Salisbury, Michael J Berry, Olivier Marre, and Stephanie Palmer. 2023. "Stimulus invariant aspects of the retinal code drive discriminability of natural scenes." *bioRxiv,* 2023–08.

Howard, Jeremy, et al. 2018. *fastai.* https://github.com/fastai/fastai.

Hubel, David H, and Torsten N Wiesel. 1959. "Receptive fields of single neurones in the cat's striate cortex." *The Journal of physiology* 148 (3): 574.

Hyvärinen, Aapo, and Petteri Pajunen. 1999. "Nonlinear independent component analysis: Existence and uniqueness results." *Neural Networks* 12, no. 3 (April): 429–439. https://doi.org/10.1016/s0893-6080(98)00140-3.

Jarsky, Tim, Mark Cembrowski, Stephen M Logan, William L Kath, Hermann Riecke, Jonathan B Demb, and Joshua H Singer. 2011. "A synaptic mechanism for retinal adaptation to luminance and contrast." *Journal of Neuroscience* 31 (30): 11003–11015.

Jaynes, Edwin T. 1957. "Information theory and statistical mechanics." *Physical review* 106 (4): 620.

Jia, Shanshan, Zhaofei Yu, Arno Onken, Yonghong Tian, Tiejun Huang, and Jian K Liu. 2021. "Neural system identification with spike-triggered non-negative matrix factorization." *IEEE Transactions on Cybernetics.*

Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. 2016. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution." *European Conference on Computer Vision (ECCV)* (March): 694–711. arXiv: 1603.08155 [cs.CV].

Jovancevic-Misic, Jelena, and Mary Hayhoe. 2009. "Adaptive gaze control in natural environments." *Journal of Neuroscience* 29 (19): 6234–6238.

Kastner, David B, Stephen A Baccus, and Tatyana O Sharpee. 2015. "Critical and maximally informative encoding between neural populations in the retina." *Proceedings of the National Academy of Sciences* 112 (8): 2533–2538.

Khemakhem, Ilyes, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. 2019. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework." *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pages 2207-2217, year 2020* (July). arXiv: 1907.04809 [stat.ML].

Kim, Kerry J, and Fred Rieke. 2003. "Slow Na+ Inactivation and Variance Adaptation in Salamander Retinal Ganglion Cells." *Journal of Neuroscience* 23 (4): 1506–1516. ISSN: 0270-6474. https://doi.org/10.1523/JNEUROSCI.23-04-01506.2003. eprint: https://www.jneurosci.org/content/23/4/1506.full.pdf. https://www.jneurosci.org/content/23/4/1506.

Kingma, Diederik P, and Max Welling. 2013. "Auto-Encoding Variational Bayes" (December). arXiv: 1312.6114 [stat.ML].

Kipf, Thomas N, and Max Welling. 2016. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907.*

Kolchinsky, Artemy, Brendan D. Tracey, and David H. Wolpert. 2017. "Nonlinear Information Bottleneck" (May). arXiv: 1705.02436 [cs.IT].

Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger. 2004. "Estimating mutual information." Ppublish, *Physical review. E, Statistical, nonlinear, and soft matter physics* 69 (6 Pt 2): 066138. ISSN: 1539-3755. https://doi.org/10.1103/PhysRevE.69.066138.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems,* edited by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kühn, Norma Krystyna, and Tim Gollisch. 2016. "Joint Encoding of Object Motion and Motion Direction in the Salamander Retina." Ppublish, *The Journal of neuroscience : the official journal of the Society for Neuroscience* 36 (48): 12203–12216. ISSN: 1529-2401. https://doi.org/10.1523/JNEUROSCI.1971-16.2016.

Kumar, Abhishek, and Ben Poole. 2020. "On Implicit Regularization in $\beta -VAEs$." In *International Conference on Machine Learning,* 5480–5490. PMLR.

Kummerer, Matthias, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. 2017. "Understanding Low- and High-Level Contributions to Fixation Prediction." In *2017 IEEE International Conference on Computer Vision (ICCV).* IEEE, October. https://doi.org/10.1109/iccv.2017.513.

Latham, Peter E., and Sheila Nirenberg. 2005. "Synergy, redundancy, and independence in population codes, revisited." Ppublish, *The Journal of neuroscience : the official journal of the Society for Neuroscience* 25 (21): 5195–5206. ISSN: 1529-2401. https://doi.org/10.1523/JNEUROSCI.5319-04.2005.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." Ppublish, *Nature* 521 (7553): 436–444. ISSN: 1476-4687. https://doi.org/10.1038/nature14539.

Lettvin, Jerome Y, Humberto R Maturana, Warren S McCulloch, and Walter H Pitts. 1959. "What the frog's eye tells the frog's brain." *Proceedings of the IRE* 47 (11): 1940–1951.

Levina, Elizaveta, and Peter Bickel. 2004. "Maximum Likelihood Estimation of Intrinsic Dimension." In *Advances in Neural Information Processing Systems,* edited by L. Saul, Y. Weiss, and L. Bottou, vol. 17. MIT Press. https://proceedings.neurips.cc/paper/2004/file/74934548253bcab8490ebd74afed7031-Paper.pdf.

Li, Rui, Dong Pu, Minnie Huang, and Bill Huang. 2021. "Unet-TTS: Improving Unseen Speaker and Style Transfer in One-shot Voice Cloning." *ICASSP2022* (September). arXiv: 2109.11115 [cs.SD].

Li, Ye, Hui Lu, Pei-lin Cheng, Shaoyu Ge, Huatai Xu, Song-Hai Shi, and Yang Dan. 2012. "Clonally related visual cortical neurons show similar stimulus feature selectivity." *Nature* 486 (7401): 118–121.

Lin, Hongzhou, and Stefanie Jegelka. 2018. "ResNet with one-neuron hidden layers is a Universal Approximator" (June). arXiv: 1806.10909 [cs.LG].

Liu, Jian K, Helene M Schreyer, Arno Onken, Fernando Rozenblit, Mohammad H Khani, Vidhyasankar Krishnamoorthy, Stefano Panzeri, and Tim Gollisch. 2017. "Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization." *Nature communications* 8 (1): 1–14.

Liu, Ran, Mehdi Azabou, Max Dabagia, Chi-Heng Lin, Mohammad Gheshlaghi Azar, Keith B. Hengen, Michal Valko, and Eva L. Dyer. 2021. "Drop, Swap, and Generate: A Self-Supervised Approach for Generating Neural Activity." *Annual Conference on Neural Information Processing Systems (NeurIPS)* (November). arXiv: 2111.02338 [cs.LG].

Locatello, Francesco, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. 2019. "Disentangling Factors of Variation Using Few Labels." *Eighth International Conference on Learning Representations - ICLR 2020* (May). arXiv: 1905.01258 [cs.LG].

Lopez, Carolina Mora, Srinjoy Mitra, Jan Putzeys, Bogdan Raducanu, Marco Ballini, Alexandru Andrei, Simone Severi, Marleen Welkenhuysen, Chris Van Hoof, Silke Musa, et al. 2016. "22.7 A 966-electrode neural probe with 384 configurable channels in 0.13 μm SOI CMOS." In *2016 IEEE International Solid-State Circuits Conference (ISSCC),* 392–393. IEEE.

Lukacs, Eugene, and Edgar King. 1954. "A Property of the Normal Distribution." *The Annals of Mathematical Statistics* 25, no. 2 (June): 389–394. https://doi.org/10.1214/aoms/1177728796.

MacKay, David J.C., and Zoubin Ghahramani. *Comments on 'Maximum Likelihood Estimation of Intrinsic Dimension' by E. Levina and P. Bickel (2004).*

Macke, Jakob H., Iain Murray, and Peter E. Latham. 2013. "Estimation Bias in Maximum Entropy Models." *Entropy* 15 (8): 3109–3129. ISSN: 1099-4300. https://doi.org/10.3390/e15083109. https://www.mdpi.com/1099-4300/15/8/3109.

Maheswaranathan, Niru, Lane T McIntosh, David B Kastner, Josh B Melander, Luke Brezovec, Aran Nayebi, Julia Wang, Surya Ganguli, Stephen A Baccus, and Stanford University. 2018. "Deep learning models reveal internal structure and diverse computations in the retina under natural scenes." *BioRxiv,* 340943.

Marks, Tyler D., and Michael J. Goard. 2021. "Stimulus-dependent representational drift in primary visual cortex." Epublish, *Nature communications* 12 (1): 5169. ISSN: 2041-1723. https://doi.org/10.1038/s41467-021-25436-3.

Marre, Olivier, Dario Amodei, Nikhil Deshmukh, Kolia Sadeghi, Frederick Soo, Timothy E Holy, and Michael J Berry. 2012. "Mapping a complete neural population in the retina." *Journal of Neuroscience* 32 (43): 14859–14873.

McIntosh, Lane, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen Baccus. 2016. "Deep Learning Models of the Retinal Response to Natural Scenes." In *Advances in Neural Information Processing Systems,* edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/a1d33d0dfec820b41b54430b50e96b5c-Paper.pdf.

Molano-Mazon, Manuel, Arno Onken, Eugenio Piasini, and Stefano Panzeri. 2018. "Synthesizing realistic neural population activity patterns using generative adversarial networks." *arXiv preprint arXiv:1803.00338.*

Monteith, Kristine, James L Carroll, Kevin Seppi, and Tony Martinez. 2011. "Turning Bayesian model averaging into Bayesian model combination." In *The 2011 international joint conference on neural networks,* 2657–2663. IEEE.

Nakatani, K, and K-W Yau. 1988. "Calcium and light adaptation in retinal rods and cones." *Nature* 334 (6177): 69–71.

Narayanan, Hariharan, and Sanjoy Mitter. 2010. "Sample Complexity of Testing the Manifold Hypothesis." In *Advances in Neural Information Processing Systems,* edited by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, vol. 23. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2010/file/8a1e808b55fde9455cb3d8857ed88389-Paper.pdf.

Narayanan, Hariharan, and Partha Niyogi. 2009. "On the Sample Complexity of Learning Smooth Cuts on a Manifold." In *COLT.*

Nemenman, Ilya, Geoffrey D Lewen, William Bialek, and Rob R de Ruyter van Steveninck. 2008. "Neural coding of natural stimuli: information at sub-millisecond resolution." *PLoS computational biology* 4 (3): e1000025.

Nirenberg, Sheila, Steve M Carcieri, Adam L Jacobs, and Peter E Latham. 2001. "Retinal ganglion cells act largely as independent encoders." *Nature* 411 (6838): 698–701.

Ölveczky, Bence P, Stephen A Baccus, and Markus Meister. 2003. "Segregation of object and background motion in the retina." *Nature* 423 (6938): 401–408.

Palmer, Stephanie E., Olivier Marre, Michael J. Berry, and William Bialek. 2015. "Predictive information in a sensory population." *Proceedings of the National Academy of Sciences* 112, no. 22 (May): 6908–6913. https://doi.org/10.1073/pnas.1506855112.

Peinado, Alejandro, Rafael Yuste, and Lawrence C Katz. 1993. "Gap junctional communication and the development of local circuits in neocortex." *Cerebral Cortex* 3 (5): 488–498.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms.* Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press. ISBN: 978-0-262-03731-0. https://mitpress.mit.edu/books/elements-causal-inference.

Pillow, Jonathan W, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. 2008. "Spatio-temporal correlations and visual signalling in a complete neuronal population." *Nature* 454 (7207): 995–999.

Pillow, Jonathan W, and Eero P Simoncelli. 2006. "Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis." *Journal of vision* 6 (4): 9–9.

Pope, Phillip, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. 2021. "The Intrinsic Dimension of Images and Its Impact on Learning" (April). arXiv: 2104.08894 [cs.CV].

Puchalla, Jason L, Elad Schneidman, Robert A Harris, and Michael J Berry. 2005. "Redundancy in the population code of the retina." *Neuron* 46 (3): 493–504.

Ramirez, Luisa, and William Bialek. 2021. "Compression as a path to simplification: Models of collective neural activity." *arXiv preprint arXiv:2112.14334*.

Razavi, Ali, Aäron van den Oord, Ben Poole, and Oriol Vinyals. 2019. "Preventing Posterior Collapse with delta-VAEs" (January). arXiv: 1901.03416 [cs.LG].

Recanatesi, Stefano, Ulises Pereira-Obilinovic, Masayoshi Murakami, Zachary Mainen, and Luca Mazzucato. 2022. "Metastable attractors explain the variable timing of stable behavioral action sequences." Ppublish, *Neuron* 110 (1): 139–153.e9. ISSN: 1097-4199. https://doi.org/10.1016/j.neuron.2021.10.011.

Rieke, Fred. 2001. "Temporal contrast adaptation in salamander bipolar cells." *Journal of neuroscience* 21 (23): 9445–9454.

Rikhye, Rajeev V, and Mriganka Sur. 2015. "Spatial correlations in natural scenes modulate response reliability in mouse visual cortex." *Journal of Neuroscience* 35 (43): 14661–14680.

Rodieck, Robert W. 1998. *The first steps in seeing.* Sinauer Associates.

Rodieck, Robert W, and Jonathan Stone. 1965. "Analysis of receptive fields of cat retinal ganglion cells." *Journal of neurophysiology* 28 (5): 833–849.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation" (May). arXiv: 1505.04597 [cs.CV].

Roudi, Yasser, Sheila Nirenberg, and Peter E Latham. 2009. "Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't." *PLoS computational biology* 5 (5): e1000380.

Ruderman, Daniel L, and William Bialek. 1994. "Statistics of natural images: Scaling in the woods." *Physical review letters* 73 (6): 814.

Saleem, Aman, Holger Krapp, and Simon R. Schultz. 2008. "Receptive field characterization by spike-triggered independent component analysis." *Journal of Vision* 8 (13).

Salibury, Jared, and Stephanie E. Palmer. n.d. "A dynamic scale-mixture model of motion in natural scenes."

Schneidman, Elad, Michael J Berry, Ronen Segev, and William Bialek. 2006. "Weak pairwise correlations imply strongly correlated network states in a neural population." *Nature* 440 (7087): 1007–1012.

Schneidman, Elad, William Bialek, and Michael J. Berry. 2003. "Synergy, redundancy, and independence in population codes." Ppublish, *The Journal of neuroscience : the official journal of the Society for Neuroscience* 23 (37): 11539–11553. ISSN: 1529-2401.

Schwab, David, Stephanie Palmer, Thierry Mora, and Olivier Marre. *Decoding and encoding retinal ganglion cell responses with deep neural networks.* NeurIPS2016 workshop: Brains and Bits.

Schwartz, Greg, Rob Harris, David Shrom, and Michael J Berry. 2007. "Detection and prediction of periodic patterns by the retina." *Nature neuroscience* 10 (5): 552–554.

Schwartz, Greg, and Fred Rieke. 2011. "Nonlinear spatial encoding by retinal ganglion cells: when 1+ 1≠ 2." *Journal of General Physiology* 138 (3): 283–290.

Schwartz, Greg, Sam Taylor, Clark Fisher, Rob Harris, and Michael J Berry. 2007. "Synchronized firing among retinal ganglion cells signals motion reversal." *Neuron* 55 (6): 958–969.

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

Shah, Nishal P, Nora Brackbill, Colleen Rhoades, Alexandra Kling, Georges Goetz, Alan M Litke, Alexander Sher, Eero P Simoncelli, and EJ Chichilnisky. 2020. "Inference of nonlinear receptive field subunits with spike-triggered clustering." Edited by Tatyana O Sharpee and Joshua I Gold. *eLife* 9 (March): e45743. ISSN: 2050-084X. https://doi.org/10.7554/eLife.45743. https://doi.org/10.7554/eLife.45743.

Shapley, Robert M, and Jonathan D Victor. 1978. "The effect of contrast on the transfer properties of cat retinal ganglion cells." *The Journal of physiology* 285 (1): 275–298.

Shlens, Jonathon, Fred Rieke, and EJ Chichilnisky. 2008. "Synchronized firing in the retina." *Current opinion in neurobiology* 18 (4): 396–402.

Simmons, Kristina D, Jason S Prentice, Gašper Tkačik, Jan Homann, Heather K Yee, Stephanie E Palmer, Philip C Nelson, and Vijay Balasubramanian. 2013. "Transformation of stimulus correlations by the retina." *PLoS computational biology* 9 (12): e1003344.

Sorochynskyi, Oleksandr, Stéphane Deny, Olivier Marre, and Ulisse Ferrari. 2021. "Predicting synchronous firing of large neural populations from sequential recordings." *PLoS computational biology* 17 (1): e1008501.

Steinmetz, Nicholas A, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. 2021. "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings." *Science* 372 (6539): eabf4588.

Stern, Merav, Nicolae Istrate, and Luca Mazzucato. 2021. "A reservoir of timescales in random neural networks" (October). https://doi.org/10.1101/2021.10.11.463861.

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

Tanaka, Hidenori, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen Baccus, and Surya Ganguli. 2019. "From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction." In *Advances in Neural Information Processing Systems,* edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/eeaebbffb5d29ff62799637fc51adb7b-Paper.pdf.

Tishby, Naftali, Fernando C. Pereira, and William Bialek. 2000. "The information bottleneck method" (April). arXiv: physics/0004057 [physics.data-an].

Tkačik, Gašper, Olivier Marre, Dario Amodei, Elad Schneidman, William Bialek, and Michael J Berry. 2014. "Searching for collective behavior in a large network of sensory neurons." *PLoS computational biology* 10 (1): e1003408.

Tkačik, Gašper, Jason S Prentice, Vijay Balasubramanian, and Elad Schneidman. 2010. "Optimal population coding by noisy spiking neurons." *Proceedings of the National Academy of Sciences* 107 (32): 14419–14424.

V.Skitovitch. 1953. "On a property of the normal distribution." *DAN SSSR* 89:217–219.

Wang, Siwei, Benjamin Hoshal, Elizabeth de Laittre, Thierry Mora, Michael Berry, and Stephanie Palmer. 2022. "Learning low-dimensional generalizable natural features from retina using a U-net." *Advances in Neural Information Processing Systems* 35:11355–11368.

Warland, David K, Pamela Reinagel, and Markus Meister. 1997. "Decoding visual information from a population of retinal ganglion cells." *Journal of neurophysiology* 78 (5): 2336–2350.

Weigt, Martin, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. 2009. "Identification of direct residue contacts in protein–protein interaction by message passing." *Proceedings of the National Academy of Sciences* 106 (1): 67–72.

Wong-Riley, MTT. 1974. "Synaptic organization of the inner plexiform layer in the retina of the tiger salamander." *Journal of neurocytology* 3 (1): 1–33.

Xia, Ji, Tyler D. Marks, Michael J. Goard, and Ralf Wessel. 2021. "Stable representation of a naturalistic movie emerges from episodic activity with gain variability." *Nature Communications* 12, no. 1 (August). https://doi.org/10.1038/s41467-021-25437-2.

You, Yuning, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. "Graph contrastive learning with augmentations." *Advances in neural information processing systems* 33:5812–5823.

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

Zhang, Da, and Mansur R. Kabuka. 2018. "Protein Family Classification with Multi-Layer Graph Convolutional Networks." In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 2390–2393. https://doi.org/10.1109/BIBM.2018.8621520.

Zhou, Ding, and Xue-Xin Wei. 2020. "Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE." *NeurIPS 2020* (November). arXiv: 2011.04798 [stat.ML].

Zhuang, Chenyi, and Qiang Ma. 2018. "Dual Graph Convolutional Networks for Graph-Based Semi-Supervised Classification." In *Proceedings of the 2018 World Wide Web Conference,* 499–508. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee. ISBN: 9781450356398. https://doi.org/10.1145/3178876.3186116. https://doi.org/10.1145/3178876.3186116.

Zimmermann, Maxime JY, Noora E Nevala, Takeshi Yoshimatsu, Daniel Osorio, Dan-Eric Nilsson, Philipp Berens, and Tom Baden. 2018. "Zebrafish differentially process color across visual space to match natural scenes." *Current Biology* 28 (13).