Title

Modeling Visual Cortical Development

Author

Ligeralde, Andrew Christian de la Cruz

Publication Date

2023

Peer reviewed|Thesis/dissertation

Modeling Visual Cortical Development

by

Andrew Christian de la Cruz Ligeralde

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael R. DeWeese, Chair
Professor Bruno A. Olshausen
Professor Michael A. Silver
Professor Frederic E. Theunissen

Fall 2023

Modeling Visual Cortical Development

Abstract

Modeling Visual Cortical Development

by

Andrew Christian de la Cruz Ligeralde

Doctor of Philosophy in Biophysics

University of California, Berkeley

Professor Michael R. DeWeese, Chair

Representation is a critical component of visual neuroscience. While there is an extensive body of literature on the nature of visual representations, we lack a set of guiding principles for understanding how representations are learned during development. Our analysis here focuses on this question at a computational level. The first set of results addresses how representations are learned under the assumption of a sparse prior on the data. It is well known that sparse coding models trained on natural images learn basis functions whose shapes resemble the receptive fields (RFs) of simple cells in the primary visual cortex (V1). However, it is unclear whether certain types of basis functions emerge more quickly than others, or whether they develop simultaneously. We train an overcomplete sparse coding model (Sparsenet) on natural images and find that there is a spectral bias in the order of development of its basis functions, with basis functions tuned to lower spatial frequencies emerging earlier and higher spatial frequency basis functions emerging later. We observe the same trend in a biologically plausible sparse coding model (SAILnet) that uses leaky integrate-and-fire neurons and synaptically local learning rules, suggesting that this result is a general feature of sparse coding. These results are consistent with recent experimental evidence that the distribution of optimal stimuli for driving neurons to fire shifts towards higher frequencies during normal development in mouse V1. We find that the input data statistics can fully account for the spectral bias in sparse coding, and propose that visual experience is sufficient to drive the spectral bias in receptive field development. Our analysis of sparse coding models during training yields experimentally testable predictions for V1 development.

In the next set of results, we investigate the potential for innately generated neural activity to drive the development of efficient representation in the visual cortex. Prior to the onset of vision, neurons in the developing mammalian retina spontaneously fire in correlated activity patterns known as retinal waves. Experimental evidence suggests that retinal waves strongly influence the emergence of sensory representations before visual

experience. We model this early stage of functional development by using movies of neurally active developing retinas as pre-training data for neural networks. Specifically, we use unsupervised learning to train models on movies of retinal waves, then evaluate its performance on image classification tasks. We find that pretraining on retinal waves significantly improves performance on tasks that test object invariance to spatial translation, while slightly improving performance on more complex tasks like image classification. Notably, these performance boosts are realized on held-out natural images even though the pre-training procedure does not include any natural image data. We then propose a geometrical explanation for the increase in network performance, namely that the spatiotemporal characteristics of retinal waves facilitate the formation of separable feature representations. In particular, we demonstrate that networks pre-trained on retinal waves are more effective at separating image manifolds than randomly initialized networks, especially for manifolds defined by sets of spatial translations. These findings indicate that the broad spatiotemporal properties of retinal waves prepare networks for higher order feature extraction.

To

Lola Viring

Contents

List of Figures

Acknowledgments

First, to my academic mentors: Mrs. Ford taught the first science class I ever liked, and was the reason I decided to pursue chemistry in undergrad. Amina Qutub gave me my first undergraduate research opportunity, which is what inspired me to pursue neuroscience in graduate school. Markita Landry went above and beyond as my first rotation mentor, letting me work on a project that eventually led to my graduate research fellowship. Charles Frye's patience and knack for coming up with intuitive explanations made my transition to computational neuroscience as smooth as it could have possibly been. SueYeon Chung gave me the tools, knowledge, and resources to pursue the question that forms the basis for the third chapter of this thesis, which wouldn't exist without her and the massive help of Yilun Kuang, Teddy Yerxa, Miah Pitcher, and Marla Feller. My thesis committee, Bruno Olshausen, Frederic Theunissen, and Michael Silver, always had valuable questions that shaped my research direction and managed to squeeze every bit of value out of our yearly meetings. Finally, my graduate mentor, Mike DeWeese, took me on as an unconventional fourth rotation student. I wrote about Mike in my graduate school application, and am still in awe that I had the opportunity to work with him. His optimism and sharp scientific insight allowed me to grow in ways I never thought possible.

Next, to my friends who supported my graduate school journey, in the Bay and beyond, I probably spent way too much time hanging out with you during my Ph.D., and I don't regret a second of it. I want to specifically thank Ameesh and Alex, who ever since that one conversation at lunch in the Jones Commons, have continued to inspire me to do research. Elaine, who kept me from dropping out on so many occasions. Elena, who at times felt like she was the only person I could relate to in Berkeley. And my brother and roommate Andy, who's stuck with me since freshman year of college — there's no one else in the world I would live with for an extended period of time.

Finally, to my family: I'm incredibly fortunate that I was born into a family afflicted by a total obsession with higher education. As with many educated immigrant families, meeting them consists of the typical job, school, college major line of questioning, the answers to which are either met with a round of approving nods or a colder, but still respectful chorus of disappointed *hm*'s from onlooking aunties and uncles. I fear sometimes that to outsiders, this can read as elitist and haughty. But in reality, my family are exceedingly kind and warm, as anyone who meets them will tell you; these questions act as a sort of expectation-setting filter that can be easily overridden (in either direction). The other caveat to this is that as much as they're obsessed with education, they're entirely uninterested in a person's social and economic status. While they acknowledge the economic value of education in a purely pragmatic sense — it was a window of opportunity for most of them, after all — they never pushed the view of education as purely a means to financial gain. Learning was taught to us as something inherently valuable.

My parents, who themselves hold Ph.D.s, chose their career paths in service of their home country of the Philippines. My dad, an economist, studied economic policies addressing poverty as a way to uplift a struggling nation ravaged by the Marcos regime. My

mom, a teacher turned political scientist, would traverse Manila in jeepneys to teach college courses in the morning and do volunteer work in the slums in the evening. When they came to America to start a family, their singular focus was to make sure that my brother and I had every opportunity they didn't have, at the glaring expense of their own lifestyles. I often think about an article written by a teacher at a low-resource high school in the Bay Area begging Stephen Curry *not* to come visit his students. The reason was many of his low-income students had been fed media narratives that being drafted in the NBA is a viable way out of poverty, when in reality, it's an exceedingly unlikely event, precipitated by the coincidence of all-consuming work ethic, an outlier level of natural ability, and in Steph's case, extraordinarily lucky parentage in the form of former NBA player Dell Curry, who gave him a massive leg up amidst many aspiring athletes. I'm by no means comparing myself to Steph, but I am comparing my parents to his. They wholly supported every single pursuit I wanted to try, academic or not, from jazz piano to basketball, and even baseball, no matter how painful it was for them to watch. Being raised by my parents guaranteed that my early years in school were incredibly repetitive and boring, and my later years in college and beyond were endlessly fascinating. I owe them everything.

My Lola Viring passed when I was 9 years old. I knew her mostly through stories, but her legacy and mark on this family are immeasurable. Her small home-catering business sent her children to private school, and in turn lifted her family out of poverty, enabling her children to live full, rich lives. She raised them to be unselfish, caring, relentless individuals, my mom being one of six. Auntie Myra and Uncle Bob gave me my first piano keyboard when I was 5 years old and since then have attended nearly every major milestone in me and my brother's lives, down to middle school graduation. They're our stand in grandparents and constant sources of inspiration. Uncle Bertie searched for my first apartment in Berkeley. Before that, he was my remote math tutor, and before that, he led the way for his younger siblings as the eldest of six. He is a model older brother and one of the kindest and most generous people I know. Auntie Liza, Uncle Vic, and Uncle Cholo are my home away from home, and let me talk to them when there are things that I don't yet want to share with my parents. Ate Nini, Kuya Ninong, Kuya Anton, and Kuya Misha are the older siblings I never had, and I look up to them to this day. The Nackleys, whose visits to the Bay saved me from eating eggs for a week straight, and Isabel, who was always inviting me to things and pulling me out of spirals during stressful stretches of grad school. The Chans, who lent me their old BMW when I lived in Houston, which enabled me to continue doing my research there, continue to be endlessly supportive and kind long after I left Texas. And finally, my younger brother Robby saved me from being an only child, for which I'm eternally grateful. Despite my influence, he turned out to be kind, smart, tenacious, funny, and moral to a fault. He constantly inspires me to be better.

Chapter 1

Vision as Representation Learning

1.1 Overview

At every turn, the *E. coli* bacterium is faced with a binary choice: "run" or "tumble". During "run", the bacterium rotates its flagella counterclockwise, forming a bundle that propels it in a straight line. During "tumble", the bacterium rotates its flagella clockwise, which unravels its bundle and causes it to randomly change direction [1]. This choice is informed by two sensory mechanisms: one which records the current concentration of food in its surroundings, and another which records the concentration of food moments earlier. If the bacterium senses a positive gradient of an enticing chemical, it will continue swimming in a straight line. If not, it will tumble until it does [2].

Many times during graduate school, I felt like the *E. coli*. In the literal sense, I was constantly looking for free food. More figuratively, the research process often resembles the *E. coli*'s random walk, especially in the early days. I'm not sure whether my sense for detecting positive gradients improved over time, or whether my graduation timeline compelled me to commit to more audacious leaps forward in my last couple years. Either way, this thesis represents the sum total of my random walk. The purpose of this overview is to describe the path that led me to Modeling Visual Cortical Development — for one, to give this thesis some narrative structure — but also to be forthcoming about the nonlinear nature of finding a good research question, a process which is typically omitted (for good reason) in scientific writing.

My first real exposure to neuroscience was as an undergrad working in Amina Qutub's lab, where I was tasked with developing a protocol to grow neural progenitor cells suspended in a three-dimensional matrix. The idea was to more closely mimic how they might form connections during development, so we could image them and analyze their network properties. For a whole summer, I tried what must've been ten iterations of a protocol that all inevitably ended in mass cell death, before they even formed any synapses. It was maybe at this point that I decided I would pivot to computation.

I was lucky that one of my early tumbles in graduate school led me to a project that,

while not a chapter in this thesis, informed my research immensely. The broader goal of this project, led by Charles Frye, was to understand why it's so easy to optimize neural networks without frequently landing in bad local minima. Prior work proposed the explanation, substantiated by numerical evidence, that there simply are no bad local minima. We studied the algorithms used to gather this evidence — critical-point finding algorithms — and found that they often fail on neural networks, which previous work hadn't considered [3]. Through this project, I dove into machine learning and optimization theory, learned to write good research code, and most importantly, developed instincts for casting complex problems as simpler, illustrative toy problems. In a much broader sense, I began thinking more about how complex systems — from the artificial neural networks we studied, to biological neural networks in the brain — actually learn to do the tasks they're good at solving.

Without a well-defined research question in mind, it took me many months of reading and playing around with different existing models for neuroscience — including the sparse coding models I discuss in Chapter 2 — before I came across an article called "Could a Neuroscientist Understand a Microprocessor?" [4] that helped me crystallize the question I was really after. I realized what I wanted to ask essentially fell under Marr's famous three levels of analysis: 1) If brain development is optimizing to solve a particular computational problem, what is it? 2) What algorithms does it employ to solve this problem? And 3) What are the substrates the brain uses to execute these algorithms? [5].

Brain development is a huge area of research. But at the time, I decided to focus on the development of the visual system, out of a self-imposed pragmatism that I felt the need to adopt in order to finish my Ph.D. in a reasonable time. After all, it seemed like vision had been widely studied as a computational problem, with image classification models achieving state of the art performance [6] and deep networks being extensively applied to understand visual cortical responses [7, 8]. Vision presented what I felt at the time was a tractable path forward: a testbed for asking new questions, but within a well-developed area of research.

Vision was the medium through which I became interested in *representation* — how sensory neurons, and the brain in general, represents information — a topic which concerns the majority of the work in this thesis. The first project in this thesis, which constitutes Chapter 2, started out with a simple question: which receptive fields develop first, second, third, and so on and so forth in a sparse coding model? To orient this question in terms of the three levels of analysis, this makes an assumption about the answer to 1) and asks specifically about 2). That is, we assume the problem neurons in V1 are learning to solve during development is to efficiently represent visual stimuli [9]. We then ask, at a descriptive level, how is it doing this, and how does it prioritize which features to learn first?

In Chapter 3, we explore how the substrates of biological development (Question 3) lead to efficient representations (Questions 1 and 2). Specifically, we look at how retinal waves — spontaneous neural activity patterns that occur during development — influence the geometry and efficiency of representations. At a computational level, we take

biologically plausible "training data" and explore the scope of visual functions that can be learned.

Butchering a quote from a movie that came out at the time of writing this: computation can only take you so far. Given infinite time and resources, a hypothetical Chapter 4 would get at testing the hypotheses for development we pose in Chapters 2 and 3. That is, at a high-level, running experiments to directly record from visual neurons *in-vivo* during development, and analyze the population activity to look at representation. I'll leave it up to a future meandering graduate student with the energy and capacity, who luckily tumbles, like I did, into this line of work.

In the remainder of this chapter, I'll discuss the guiding principle of this thesis, the efficient coding hypothesis, and how it can be used to understand the problem of representation (Section 1.2). In Section 1.3, I'll then give a broad overview of the development of the visual system and the questions that concern Chapters 2 and 3.

1.2 The Efficient Coding Hypothesis

A central goal of systems neuroscience is to establish a precise quantitative description of how neurons learn to encode sensory stimuli. A principle that inspires the majority of the analyses in this thesis is the efficient coding hypothesis, which posits that the goal of visual perception is to efficiently represent incoming visual stimuli [10]. It was later hypothesized that sensory neurons do this by minimizing statistical redundancy of sensory inputs [11].

One useful hypothetical that helped me get an intuition for this principle is, given a set of sensory neurons, how would I pick their tunings to represent a yellow car? One way of doing this is to make each neuron tuned for both object and color, so that one neuron fires when it sees a yellow car, another neuron fires when it sees a red car, another neuron fires when it sees a yellow bus, etc. On one hand, whenever you see one of these represented objects, only one neuron is firing at a time, which is efficient as far as metabolic costs are concerned. However, this configuration is highly inflexible. The number of neurons in the visual cortex is finite, and there are an uncountable number of visual stimuli that could be drawn up that exceed the capacity of such a system. An alternative way to do this is to have each neuron represent one of any potentially observable car-related feature — one neuron represents wheels, one neuron represents windshields, etc. This system avoids the problem of the first, in that we can more flexible represent a greater number of objects, the number of which now depends on permutations of firing neurons, rather than the number of neurons itself. However, for any given stimulus, the a large number of neurons would have to fire, which is metabolically inefficient (assuming that the steady state of each neuron requires sufficiently low energy such that adding more neurons to the system doesn't significantly add to the overall metabolic cost). Metabolic cost aside, it may be difficult for downstream neurons to interpret a dense neural code, as opposed to a sparse one that prioritizes only the most salient features of the input.

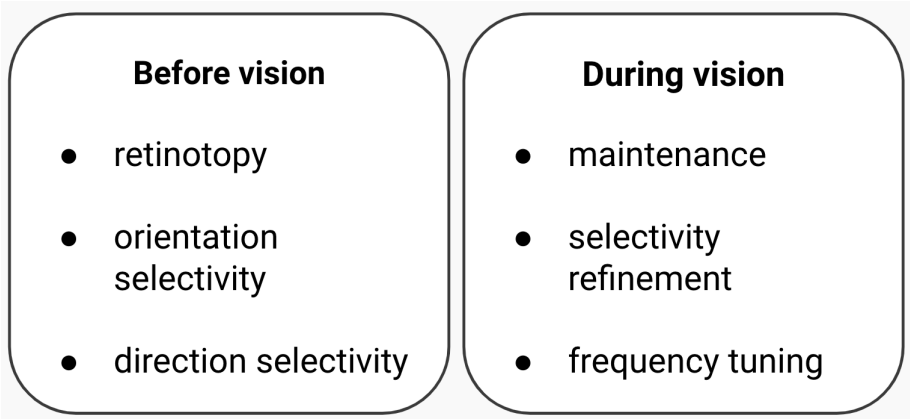


Figure 1.1: **Two stages of development.** *Left:* Many functions are established without visual experience. *Right:* The visual system is refined and maintained with visual experience.

Where on this spectrum of efficiency and flexibility does the visual cortex lie? The efficient coding hypothesis proposes that rather than have the system assign equal likelihood to every possible feature (as in system 2), the system encodes for the features of the environment that are most likely. It's not obvious, however, which features these are, and which configuration can balance the tasks the neurons (and more generally, the organism to which they belong) have to perform with metabolic costs.

One way of getting at this question is to directly examine the statistical properties of neural population activity in response to visual stimuli [12]. I mention in the previous section that this experiment can test all of the predictions that I make in this thesis based on computational results. An alternative approach is to derive a model of representation based on statistical properties of the external environment [9, 13]. While this thesis is a computational work, we draw inspiration from both approaches. In Chapter 2, we discuss directly how such a model of representation, the sparse coding model, prioritizes learning frequencies with more power in the data, which we show accords with experimental evidence. In Chapter 3, we discuss how features of population activity are a viable instructive signal for "training" the visual system to recognize objects under certain spatial transformations.

1.3 Development of the Visual System

Development is a huge area of research. For a comprehensive review of nearly a century's worth of research concerning visual development, I highly recommend [14]. For the purposes of this thesis, the key idea to understand is the distinction between experience-dependent and experience-independent development.

Interestingly, many key aspects of visual function are well-established before visual experience, such as topographic maps, orientation selectivity, and ocular dominance [14], suggesting external stimuli are not necessary for the initial development of the visual system, and axon targeting can largely be learned by internally generated signals such as spontaneous neural activity and molecular guidance cues [15, 16, 17, 18, 19].

Visual experience, on the other hand, doesn't appear to be necessary for the initial development of the visual system, but it does maintain responsiveness and selectivity of neural responses. For example, the V1 response to the ipsilateral eye becomes much stronger if animals are permitted visual experience, and responses to both eyes gradually deteriorate if the animals are binocularly deprived [20].

This roughly two-stage timecourse of development (Fig. 1.1) has interesting implications from a theoretical perspective. For one, given that visual experience is not necessary for the initial formation of the visual system, how do we quantify and describe the effects of visual stimuli on representation learning? This is the question we tackle in Chapter 2.

Moreover, the nature of biological development suggests that visual function can be learned without explicit external visual input in the form of natural image stimuli. This is contrary to how models of vision are typically trained: on millions of labeled natural images [6, 7, 8]. That is, while existing models may mimic visual responses, they do not address the question, how does the visual system learn the representations that lead to these responses in the first place? This is the question we tackle in Chapter 3.

Chapter 2

The Spectral Bias of Sparse Representations

2.1 Chapter Summary

This chapter deals with the order of emergence of learned representations. That is, given a set of data or stimuli, which features are learned first in forming a representation?

We are interested in this question on two levels. At a purely computational level, we want to characterize the behavior of a particular class of representation learning algorithms, sparse coding algorithms, which are a particular realization of the principle of efficient coding. At a biological level, we want to use sparse coding as a descriptive model for how neurons learn their sensory tunings during development. It's natural to assume there is indeed an order in which sparse representations are learned, both in a model and biological context. Any system, whether theoretical or biological, that has been optimized to perform a particular task, can only achieve high performance at the expense of performance on a different set of tasks, a notion that has been dubbed the "No Free Lunch Theorem" [21]. Neurons in particular have additional physical limitations — such as metabolic and wiring constraints — that impose restrictions on their computational capabilities [9]. Given these constraints, we hypothesize that there exists a task-dependent priority in representation learning that imposes an order in the rate at which features are learned.

Simple cells in the primary visual cortex (V1) have well-studied response properties [22, 23, 24, 25] and therefore offer a useful model system for understanding how these representations of the visual world are learned during development. In this chapter, we use computational models of neural encoding to understand how V1 simple cells learn to represent the world from a stream of visual input. While many response properties of V1 simple cells can emerge before eye-opening without the need for visual experience (e.g., orientation selectivity and ocular dominance), observations of changes in receptive field (RF) properties that depend on the nature of the visual environment suggest that plasticity

in V1 is experience-dependent [20, 14]. Experimental evidence also shows that early postnatal visual experience is necessary for natural scene representation and discriminability in V1 [26].

The process of learning to encode visual information in V1 has been modeled as an unsupervised learning problem in which neurons adapt their tuning properties in order to optimize some objective function based on the statistical structure of stimuli in the natural environment. One coding principle that has proven to be useful for understanding sensory representations is sparseness, which posits that the neural population should not only maximize fidelity to input stimuli, but also minimize the number of active units (L_0 population sparseness), or the amount of neural activity across the population (L_1 population sparseness) [9]. Sparseness is an appealing concept for biological systems, both in terms of conserving metabolic costs and efficiently representing natural scenes, which have sparse structure [27]. Indeed, sparse coding models trained on natural image data to jointly optimize both fidelity to the input and sparseness have been shown to learn basis functions whose response properties replicate simple cell receptive fields (RFs) of V1 neurons [13, 28, 29].

Experimental work demonstrates that over the course of development, the distribution of frequency tuning of V1 neurons shifts towards higher spatial frequencies, and this shift requires visual experience [30, 31]. However, the question remains whether this shift is due to high spatial frequency RFs emerging later after the early development of low spatial frequency RFs, or whether there is a global shift during development across all receptive fields towards higher spatial frequencies. We find that the Sparsenet model [13] predicts the former to be true: low spatial frequency basis functions tend to emerge earlier in training, and high spatial frequency basis functions tend to emerge later. In fact, we observe the same behavior for the SAILnet model [29] of sparse coding, which implements leaky integrate-and-fire neurons and synaptically local learning rules, suggesting both that this result is a general feature of sparse coding and that it is biologically plausible.

In the following section, we introduce the mathematics and intuition for sparse coding (Section 2.2). Next, we present results that demonstrate the spectral bias in traditional and biologically plausible implementations of sparse coding models (Section 2.3). We then propose a likely explanation for the spectral bias, namely that it arises from the statistics of the input data (Section 2.4). Finally, we discuss the implications of this prediction for the development of the visual system (Section 2.5).

2.2 Sparse Codes Are Efficient Representations of Data

Understanding the computations performed by visual neurons is a difficult task. One approach is to start from the statistical structure of the stimulus set, rather than the neuron, and use efficient coding strategies to gain insight into computations the brain may be implementing [32]. This is the motivation behind sparse coding, the goal of which is to find

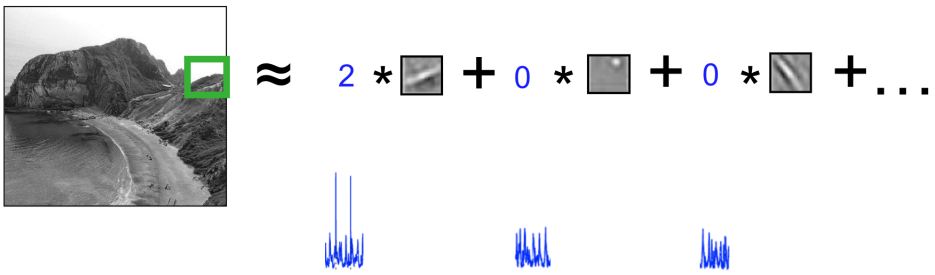


Figure 2.1: **Inferring a sparse representation of an image patch as a linear combination of basis functions.** The basis functions Φ (small patches multiplied by coefficients) are analogous to receptive fields, and the coefficients a (2, 0, 0) are analogous to neural activities (blue traces: the first one has two spikes, and the second and third have zero spikes, corresponding to the coefficient values) in response to the given input x (green square).

a model of the data x (Fig. 2.1). Sparse coding models the data as a linear combination of basis functions Φ weighted by sparse coefficients a:

$$x = \Phi a + \epsilon, \tag{2.1}$$

where ϵ is the error of the reconstruction.

Without the constraint of sparsity, there are many ways to find a suitable Φ (a matrix with each column corresponding to a basis function in the form of a vector) and a (a vector of coefficients, each corresponding to a column/basis function), principal components analysis being one of the most common. In fact, there is a link between PCA and neuroscience via linear Hebbian learning, a biologically plausible synaptic learning rule, which can be formulated to learn the principal components of the data [33, 32]. However, PCA is not a suitable model for understanding V1: reconstructions based on the basis functions with the highest variance (the top PCs) do not resemble the images they encode, and more importantly, the basis functions themselves don't resemble the receptive fields of V1 [32].

An alternate way of tackling this problem is to find the maximum likelihood,

$$p(x|\Phi) = \int p(x|a, \Phi)p(a)da, \tag{2.2}$$

where $p(a)$ is a sparse prior over the coefficients. The integral with respect to a signifies the mean of $p(x|a, \Phi)$ over the distribution of all possible coefficients, which is intractable [34]. Rather than directly calculate this integral, we can approximate the mean using the maximum a-posteriori estimate of a, given by

$$\hat{a} = \max_a \left[\log p(x|a, \Phi) + \log p(a)\right]. \tag{2.3}$$

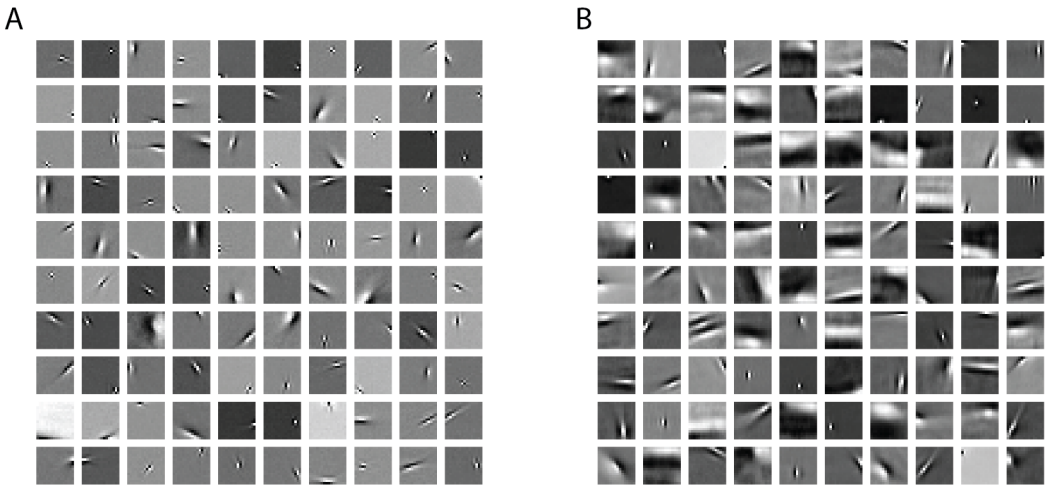


Figure 2.2: **Overcomplete dictionaries are composed of basis functions that resemble V1 simple cell receptive fields.** A) A sample of 100 basis functions from a 10 $\times$ overcomplete sparse coding model (Sparsenet). B) A sample of 100 basis functions from a 10 $\times$ overcomplete sparse coding model with biologically plausible learning rules (SAILnet).

The first term in the brackets in Eq. 2.3 is the log of the likelihood of the signal, which we model as Gaussian with mean Φa and variance σ_x. The second term is the log of the sparse prior over the coefficients, which we assume to be exponentially distributed. Taking the negative of Eq. 2.3, we can recast the optimization as a minimization problem of the sparse coding objective function E with respect to a:

$$\hat{a} = \min_a \frac{1}{2\sigma_x^2}\|x - \Phi a\|^2 + \lambda \sum_i |a_i| \tag{2.4}$$

$$\hat{a} = \min_a E(x, a, \Phi). \tag{2.5}$$

We infer a, as well as learn Φ, by gradient descent on E, which can be re-expressed in component-wise notation as

$$E = \frac{1}{2}\sum_n \left[x_n - \sum_j a_j \Phi_{n,j}\right]^2 + \lambda \sum_j |a_j|. \tag{2.6}$$

The resulting expression for E is the sparse coding objective. Alternately optimizing this objective subject to a and Φ leads to a sparse representation 2.1.

It's not obvious from the mathematical description that such a model would have biological relevance, but it turns out that the learned basis functions resemble V1 receptive fields: localized, oriented, bandpass filters that tile the space of natural images [13]. We show examples of these basis functions in Figure 2.2 from two sparse coding models: Sparsenet, the original version of sparse coding reported in [13], and SAILnet, a biologically plausible version of sparse coding reported in [29]. Importantly, unlike PCA, which only captures the lowest frequency components of natural images (since they have the highest variance), a small number of these basis functions can be used to reconstruct an input image.

2.3 Sparse Coding Predicts a Spectral Bias in the Order of Development of V1 Simple Cell Receptive Fields

The Spectral Bias of Sparse Coding

Because sparse representations are parsimonious reconstructions of input data, it's natural to assume that sparseness imposes a priority on certain basis functions being learned sooner than others. Indeed, it has been observed that only classical Gabor filters are learned in a complete or slightly overcomplete regime, while center-surround and high-frequency bases emerge only in highly overcomplete dictionaries [35]. A similar phenomenon has been observed in overparametrized neural networks, which learn lower frequency functions earlier in training [36, 37]. In this particular case, it was shown that given uniformly distributed data, lower degree spherical harmonics are learned more easily, and the learning rates for individual harmonics correspond to a direction determined by the eigenfunctions of the neural tangent kernel, a construct which allows neural networks to be analyzed as kernel methods [38, 39]. This suggests the driving cause of the spectral bias is the model specification (over-parametrized neural network) rather than arising from properties of the data.

In this section, we analyze sparse coding models during training to answer the following question: do some types of basis functions develop sooner than others, and if so, why? There are two preliminary steps to tackling this question: 1) Establish a notion of development, and 2) Establish a system to assign "type" to each learned basis function.

To address 1), we introduce a metric we call **similarity** that quantifies the development of a basis function. Given a basis function at a point t in training, denoted by BF_t, its development is measured by its degree of similarity to its final learned shape at the final training time step, denoted by BF_{final}. We quantify this using the cosine similarity, the cosine of the angle between two vectors, of BF_t and BF_{final}. This is expressed as

$$\textbf{similarity}(BF_t, BF_{final}) = \frac{BF_t \cdot BF_{final}}{\|BF_t\| \, \|BF_{final}\|}, \tag{2.7}$$

where $\|\cdot\|$ denotes the L_2 norm. By definition, the maximum similarity between any two basis functions is 1, which indicates that they are equal up to a re-scaling of the pixel intensities. Two orthogonal basis functions have a similarity of 0.

Regarding 2), prior work in sparse coding has typically described learned basis functions qualitatively by manual sorting each function into a "canonical" type (Gabor, edge detector, etc.) [35] or quantitatively by reporting parameter values obtained from fitting Gabor functions [29]. We found that in the highly overcomplete regime, neither approach is robust enough to account for the full diversity of the learned dictimonary. Qualitative inspection by eye is simply too time consuming, low-throughput, and likely unreliable. We found some success through fitting — simple gradient descent on different random initializations of a 2D Gabor function turns out to be robust for most basis functions [40]. However, classification on the learned function parameters still requires either user-set thresholds or labeling data to train a network classifier, both of which reintroduce the biases of manual inspection. Unsupervised clustering on the parameters via k-means or UMAP gave inconsistent results, but is perhaps worth revisiting.

We get around these issues by classifying basis functions according to their power spectra. Not only can we calculate the power spectrum for any basis function via discrete Fourier transform, regardless of its resemblance to a canonical type, it provides a direct way to classify a basis function into a discrete class without manually imposing thresholds, namely the frequency with maximum power f_{max} in its power spectrum $P(f)$

$$f_{max} = \text{argmax}\, P(f). \tag{2.8}$$

To ensure that we only categorize converged basis functions, rather than those that are still at or near their random white noise initialization, we only admit f_{max} categories with at minimum 100 basis functions.

Taking 1) and 2) together, we can track the convergence of each basis function to its final learned state and examine the mean convergence for each frequency bin, defined by f_{max} (Fig. 2.3). To obtain these results, we train a Sparsenet model via gradient descent (Euler's method) with a Laplace prior (L_1 penalty) over the coefficients [41]. The model is trained on 16 $\times$ 16 patches of whitened natural images that were obtained from David Field's original dataset for sparse coding [13, 41]. Prior to drawing patches, the full images are whitened with a cutoff frequency of $f_0 = 150$ (Section 2.4). To ensure sufficient diversity of the learned dictionary, we used an overcomplete model with 512 basis functions. The model is trained for 50000 iterations with a batch size of 100 patches, 300 iterations of inference per batch, sparsity penalty $\lambda = 0.3$, coefficient learning rate $\eta =$ 1e-3, and a dictionary learning rate $\alpha =$ 1e-3. The basis functions in the model are initialized with Gaussian-distributed white noise.

We find that on average, low frequency basis functions converge to their final learned shapes (reach similarity of 1) first, followed by the mid frequency basis functions, and the high frequency basis functions converging last (Fig. 2.3). We also find that at a dictionary-wide level, the learned representation has more power at lower frequencies earlier in train-

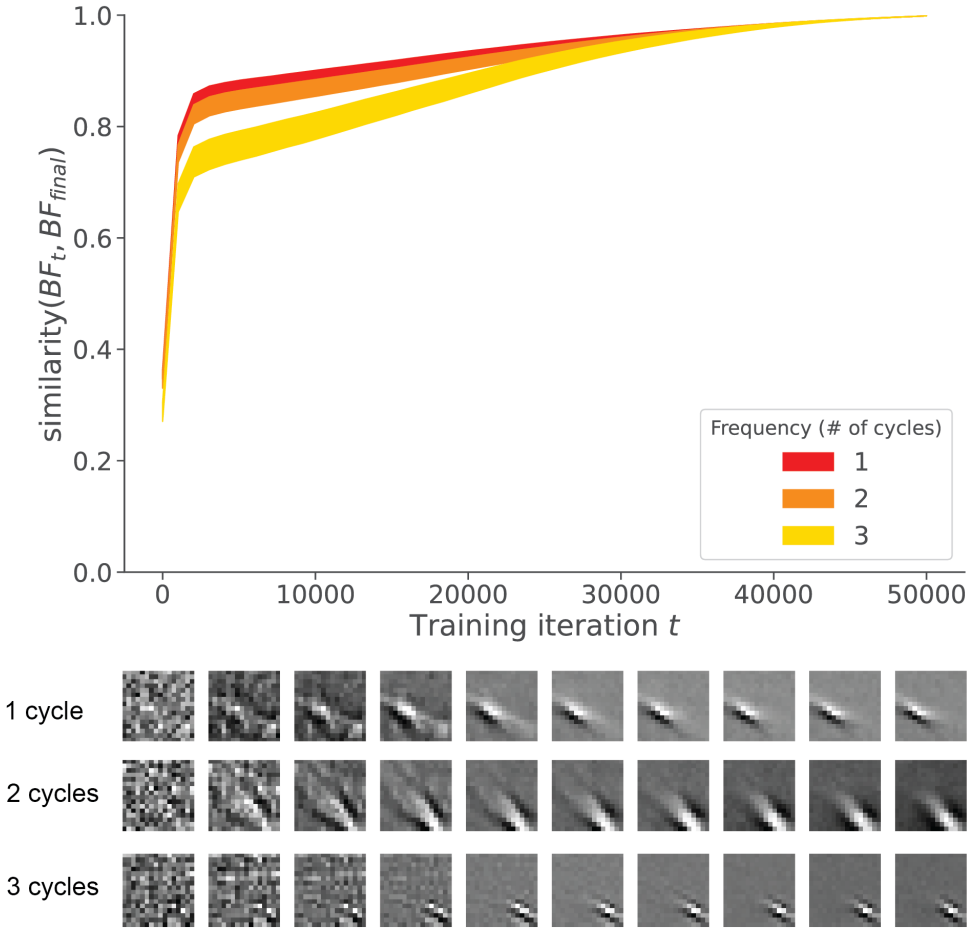


Figure 2.3: **Convergence of basis functions in Sparsenet grouped by f_{max}.** *Top:* A value of 1 on the y-axis denotes a basis function that has fully converged to its final learned shape BF_{final}, so the higher the curve, the faster the basis function has converged. The shaded regions denote standard deviation about the mean for each category. *Bottom:* Representative examples of one basis function developing (left to right, shown every 5000 training iterations) from each frequency category, starting from random initialization.

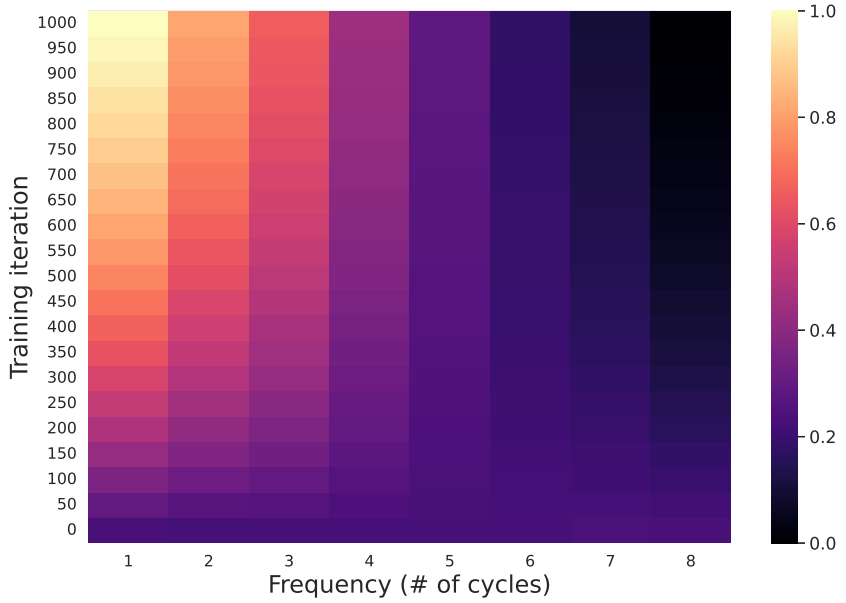


Figure 2.4: **Power spectrum of Sparsenet dictionary over training.** We characterize the power of the whole dictionary at each training iteration by taking the mean power spectrum of the basis functions. We normalize the power on a 0-1 scale by the maximum and minimum power at a given frequency and iteration. Notably, the spectral bias emerges quickly in the first 1000 iterations of training.

ing (Fig. 2.4). Taken together, these results demonstrate that the rate at which basis functions are learned is characterized by a spectral bias towards lower frequency features in the data. Moreover, this spectral bias occurs at both the local level of individual basis functions — the lower frequency basis functions emerge first — and at a global level across all basis functions — the dictionary as a whole is tuned for lower frequencies earlier in training.

The coincidence of local and global spectral biases is non-trivial, in that one does not necessitate the other. Consider a population of developing sensory neurons. In a scenario where local but not global spectral bias would occur during their development, a small handful of neurons with f_{max} values at low frequencies may emerge early on in development, while the rest of the neurons have initially uniform responses that gradually increase in power at higher frequencies throughout development. In this case, since the small population of low-frequency tuned neurons develops more quickly, a similarity-like metric would reveal a local spectral bias. However, because there are so few of them, at a population level, the whole population neuronal response to all frequencies may appear uniform throughout development, given the many small contributions to the population response

from neurons slightly tuned for higher frequencies. In a scenario where global but not local spectral bias would occur, a majority of neurons are tuned for low-frequencies, but develop at vastly different rates. Meanwhile, a handful of high-frequency tuned neurons develop quickly. However, because there are so few, they don't contribute much to the global response to high-frequency stimuli, whereas the many low-frequency neurons on average can encode low-frequency stimuli well early on in development. There are many other hypothetical edge cases where the notion of spectral bias as defined here would fail, for instance, if neurons change their optimal tuning frequency over development. In this setting, which we discuss in Section 2.3, local spectral bias would be impossible to define based on a convergence metric, though perhaps some other method could be used to track the average response of a neuron over time.

All this being said, the simultaneous occurrence of both local and global spectral biases indicates that 1) the relative sizes of each frequency category also follow a spectral bias that is consistent with the bias in their convergence rates, and 2) the convergence rates within the same frequency category defined by f_{max} are similar. We explore the reasons for the spectral bias in Section 2.4. But first, we examine whether these results hold in a biologically plausible implementation of sparse coding.

Biologically Plausible Sparse Coding

We perform the same analysis as in the previous section using the SAILnet model, a biologically plausible sparse coding model that uses leaky integrate-and-fire (LIF) neurons and local learning rules [29]. The model is distinct from traditional sparse coding in that the inferred coefficients are in the form of discrete spike counts, and the learning rules are synaptically local, neither of which is true in Sparsenet.

In SAILnet, each model neuron, indexed by i, is associated with a receptive field (basis function) Q_i. At each inference step t, the neuron's spike is recorded by the binary-valued activity variable y_i, and at the end of inference, its total activity over the inference period n_i is calculated as

$$n_i = \sum_t y_i^{(t)}, \tag{2.9}$$

a discrete value that serves as the inferred coefficient. Whether or not the neuron fires ($y_i = 1$) is determined by two factors: 1) excitatory feed-forward input from the stimulus image pixel X_k weighted by Q_{ik}, and 2) inhibitory input from the activity $n_{j \neq i}$ of other neurons in the model, weighted by inhibitory synapses W_{ij}. If the net input to a neuron exceeds its threshold value θ_i in response to a given image at inference step t, the neuron will fire:

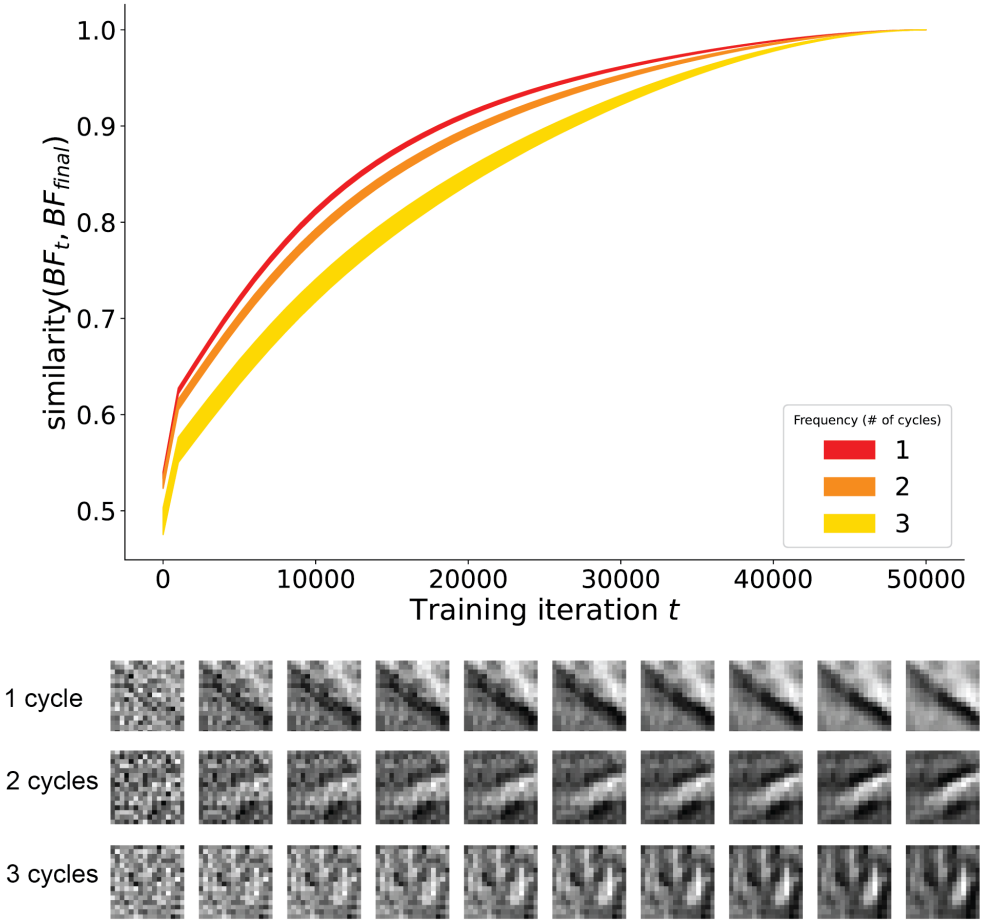


Figure 2.5: **Convergence of basis functions in SAILnet grouped by f_{max}.** Notably, additional considerations apply to training SAILnet that don't apply to Sparsenet. First, SAILnet requires a greater degree of whitening of the input data. Here, we impose a cutoff frequency of $f_0 = 256$ cycles per image, the Nyquist frequency of the Field images. Second, a learning rate schedule is set to ensure the stable convergence of the learned basis functions.

$$y_i = \begin{cases} 1 & \text{if } \sum_k Q_{ik} X_k - \sum_{j \neq i} W_{ij} n_j \geq \theta \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

The approximate pixel reconstruction is given as

$$\hat{X_k} = \sum_i n_i Q_{ik}. \tag{2.11}$$

After each inference period, SAILnet is trained on Hebbian and anti-Hebbian rules similar to those used in [42], with the additional constraint that learning is localized to each synapse without information from any other synapses in the network. The feed-forward weights Q and inhibitory synapses W are both trained by iterative local learning rules such that the update rule for an individual synaptic weight only depends on information available at that synapse during training. A third learning rule trains each neuron's firing threshold θ_i, which modulates how often that neuron fires for a given amount of input. This rule is also local in that it only trains each threshold based on the current firing rate of that neuron, without access to the firing rates of any other neurons in the network. We emphasize these features of the model to demonstrate that this model achieves sparse representations while incorporating biologically plausible learning mechanisms — as could occur in real neuronal networks such as the population of simple cells in V1. The three learning rules are summarized below:

$$\Delta Q_{ik} \propto n_i X_k - n_i^2 Q_{ik} \tag{2.12}$$

$$\Delta W_{im} \propto n_i n_m - p^2 \tag{2.13}$$

$$\Delta \theta_i \propto n_i - p, \tag{2.14}$$

where p is a globally set target firing rate. The two synaptic updates modulate the image reconstruction and pairwise decorrelation, respectively. The latter update encourages neurons to learn distinct receptive fields. The threshold update encourages lifetime sparseness for each individual neuron so that it doesn't fire too frequently. It can also be shown that these learning rules approximately minimize the Lagrangian,

$$\mathcal{L} = \sum_k (X_k - \sum_i n_i Q_{ik})^2 + \sum_i \lambda_i (n_i - p) + \sum_{i \neq m} \tau_{im} (n_i n_m - p^2), \tag{2.15}$$

where λ_i and τ_{im} are Lagrange multipliers corresponding to the sparseness and pairwise decorrelation parameters, respectively [29].

There are two additional considerations to training SAILnet that don't apply to training Sparsenet. The first is that while Sparsenet does not require whitened training data to

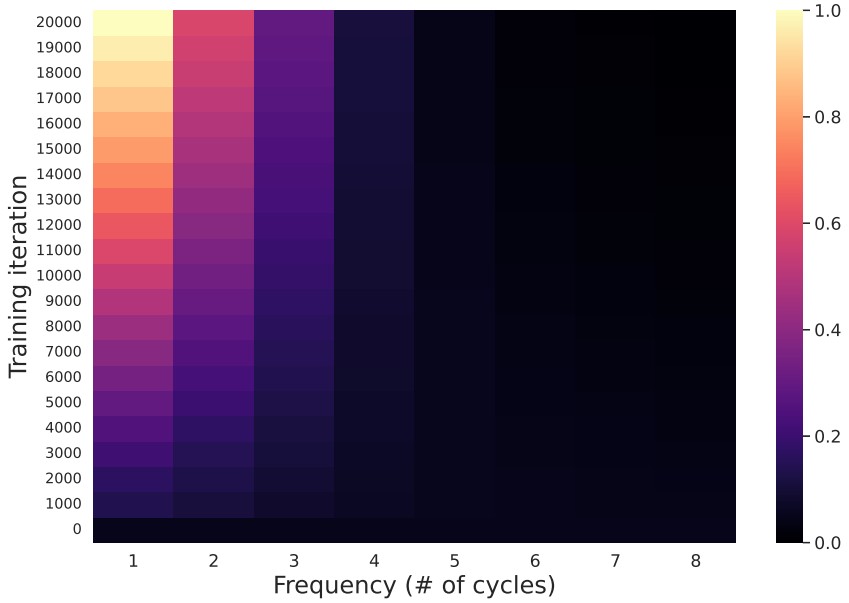


Figure 2.6: **Power spectrum of SAILnet dictionary over training.** Despite being highly overcomplete, the model hardly learns frequency content of 4 cycles or higher, if at all.

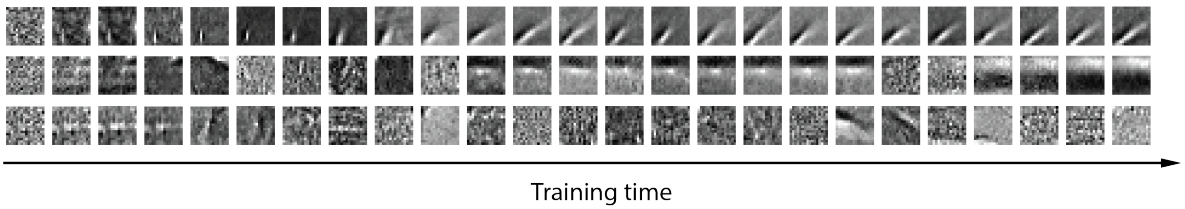


Figure 2.7: **SAILnet basis functions exhibit fluidity during training in certain hyperparameter regimes.** Each row depicts the time evolution of a basis function over training, each corresponding to a observed mode of fluidity. Panels are plotted every 1000 training iterations. *Top row:* In the first mode, a basis function will converge to a particular solution, then gradually diverge into a new, similar solution. *Middle row:* In the second mode, a basis function will converge to a solution for a fixed number of training iterations, then abruptly shift into a seemingly dissimilar solution. *Bottom row:* In the third mode, a basis function will rapidly fluctuate without any temporary or long term stability.

learn V1-like basis functions, SAILnet does. This observation suggests sparse coding with biologically plausible local learning rules requires decorrelated inputs [43]. We explore this idea further in Section 2.4. For now, we note a key difference in our SAILnet training procedure is that during preprocessing, we whiten the data with a cutoff frequency of $f_0 = 256$, the Nyquist frequency of the full Field images. The second consideration in training SAILnet is that in certain hyperparameter regimes, the basis functions might only learn low-frequency features or exhibit what we will refer to as "fluidity" and continue to change indefinitely rather than converge to a final shape (Fig. 2.7). The latter phenomenon was independently identified as "drift" in [44], where they find that the drifting receptive fields of individual neurons trained using Hebbian/anti-Hebbian learning rules can be characterized by a coordinated random walk. This work proposes that the objective has a degenerate solution space, and fluidity is the result of exploring this space via noisy synaptic updates.

To properly measure basis function convergence (Eq. 2.7), we need a set of hyperparameters that guarantees that each basis function stably reaches a particular solution. To do this, we impose a learning rate schedule: the initial learning rate in the original SAILnet for the first 10^3 iterations; the initial learning rates reduced by a factor of 10 from that point until 5×10^4 iterations; and tuned down by another factor of 10 from that point until the end of training. To ensure a range of learned basis functions, we used a highly overcomplete dictionary of 2048 elements, relative to the originally reported dictionary size of 1536 [29]. Accordingly, we set the lifetime sparseness parameter p, which modulates the target number of spikes per image, at $p = 0.025$, and θ_0, the initial firing thresholds, at $\theta_0 = 4.0$. Notably, these values impose twice as much sparseness as the original hyperparameters in [29].

Despite its differences from Sparsenet, SAILnet also learns basis functions in a hierarchical manner: lower frequency basis functions are learned early in training, and higher frequency basis functions are learned later in training, as measured by individual basis function convergence (Fig. 2.5). This effect also persists at the global level, although fewer frequencies are learned, which is surprising given the greater degree of whitening (Fig. 2.6). We explore this further in Section 2.4. Because Sparsenet and SAILnet represent two substantially different model architectures, we argue that these results are indicative of a general property of sparse coding, as opposed to being particular to a specific model architecture or optimization algorithm.

Highly Overcomplete Sparse Coding

In the previous section, we analyzed an $8\times$ overcomplete SAILnet model relative to the patch dimension, whereas we analyzed a $2\times$ overcomplete Sparsenet model. The reasons for this is that first order optimization methods used in training the original formulation of Sparsenet are computationally slow, particularly when scaling up the number of basis functions [13, 45, 46]. However, the overcomplete case is of particular interest due to the fact that higher frequency basis functions only emerge in overcomplete models. [35].

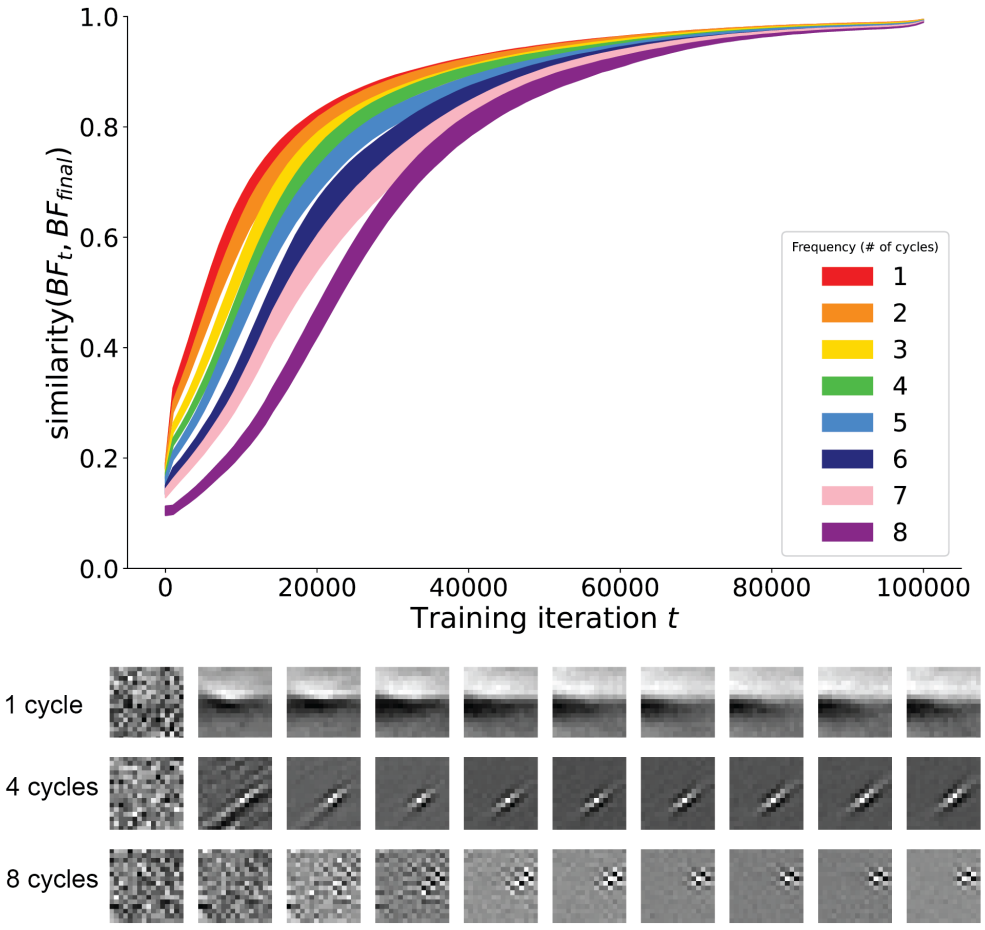


Figure 2.8: **Convergence of basis functions in a highly overcomplete Sparsenet model grouped by f_{max}.**

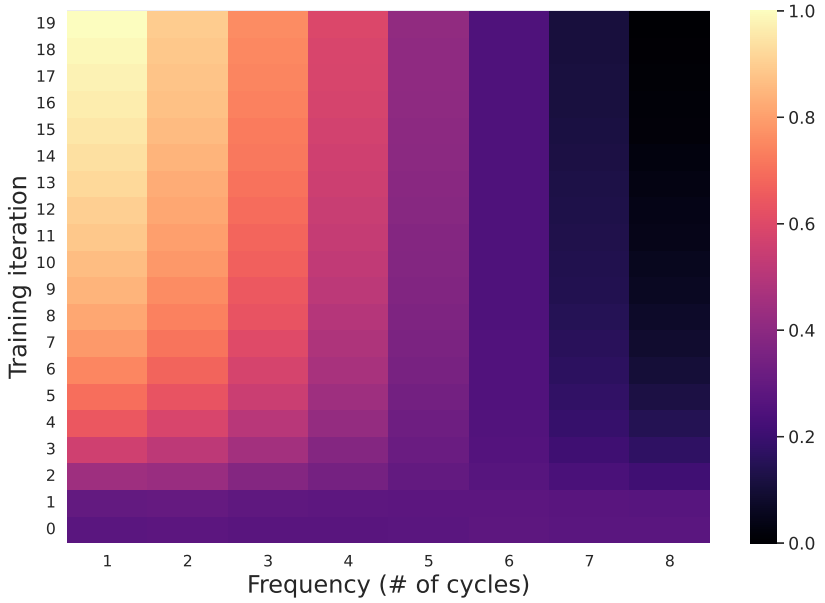


Figure 2.9: **Power spectrum of highly overcomplete Sparsenet model over training.** Notably, due to FISTA, the power spectrum is more visible at a much earlier training timepoint than in the traditional, Euler-optimized Sparsenet model.

This case provides us a way to test whether the degree of overcompleteness influences the degree of spectral bias, analogous to how overparametrization drives the spectral bias in neurla networks [39].

To get around the computational cost, we train a highly overcomplete Sparsenet model (2048 basis functions) using the Fast Iterative Shrinking-Thresholding Algorithm (FISTA) to solve the linear inverse problem of sparse coding, as described in [46]. We implement this algorithm in PyTorch based on code publicly distributed by Yubei Chen (`https://github.com/yubeic/Sparse-Coding/`). The model is trained for 10^5 iterations with a sparseness parameter of $\lambda = 0.8$. All learning rates are held constant throughout training and are specified in the source code. We find that, as in the $2\times$ overcomplete case, a highly overcomplete Sparsenet dictionary learns basis functions in a hierarchical manner, with low frequencies learned earlier in training (Figs. 2.8, 2.9). Notably, a complete range of frequencies (up to the basis function Nyquist frequency of 8) is learned.

In the following section, we show that the observed spectral bias in both Sparsenet and SAILnet is determined by the statistics of the input data.

2.4 The Spectral Bias Arises from Input Data Statistics

There are several possible causes for the spectral bias. One possibility is that it requires more spatial precision to specify higher frequency basis functions, and therefore they may take more time (i.e., more training data) to converge. Another is that the sparse coding objective itself rewards converging towards low frequencies. It's also possible that over-parametrization encourages learning simple patterns first, which can generalize to more complex patterns [36, 37]. While we don't disprove any of these possibilities, in this section we present evidence that one explanation can sufficiently explain the observed spectral bias in training, namely that the frequencies with more power in the training data are those that are learned first.

Whitening and the power spectrum of natural images

We initially discounted this explanation due to the whitening procedure used in pre-processing prior to training the models. Whitening effectively flattens the power spectrum of the data. It's suggested that this procedure simulates the function of the retina in visual processing: retinal ganglion cell spike trains are less correlated compared to the corresponding visual input stimuli, and the effect of this decorrelation is to enhance efficient coding [43, 47]. However, we find that the whitening procedure traditionally used in sparse coding doesn't completely eliminate the spectral characteristics of the input data.

Natural images, like the ones used as training data, have characteristic power spectra that roughly obey a $1/f^{2-\eta}$ power law with $0 < \eta < 0.3$ [48, 49]. Remarkably, any image you take of the real world (i.e., not computer generated) roughly conforms to this power law.

Given the falloff at high frequencies, whitening natural images effectively boosts high frequencies while attenuating low ones. We whiten in the frequency domain by multiplying with the following filter:

$$W(\vec{f}) = |\vec{f}|e^{-\left(\frac{|\vec{f}|}{f_0}\right)^n}, \tag{2.16}$$

where $\vec{f}$ denotes the two-dimensional spatial frequency. The steepness parameter n is set to 4 to produce a sharp cutoff without introducing ringing in the space domain [41]. Increasing the cutoff frequency f_0 produces a more whitened image. In our case, where the original images are 512×512 pixels, the maximum f_0 possible is the Nyquist frequency of 256 cycles per image.

We compute the mean power spectra across 10 images before and after whitening under different values of f_0 (Fig. 2.10). We perform a best line fit ($R^2 = 0.99$) to the average power spectrum of the raw, pre-whitened images and find that the average estimated power $\hat{P} = 10.1/f^{2.39}$, slightly steeper than the theoretical $1/f^{2-\eta}$ power law. The power spectra under varying f_0 do not make it obvious why sparse coding might have such a clear spectral bias in training, especially given the behavior at lower frequencies. However, recall that we're training on patches drawn from whitened images, rather than the

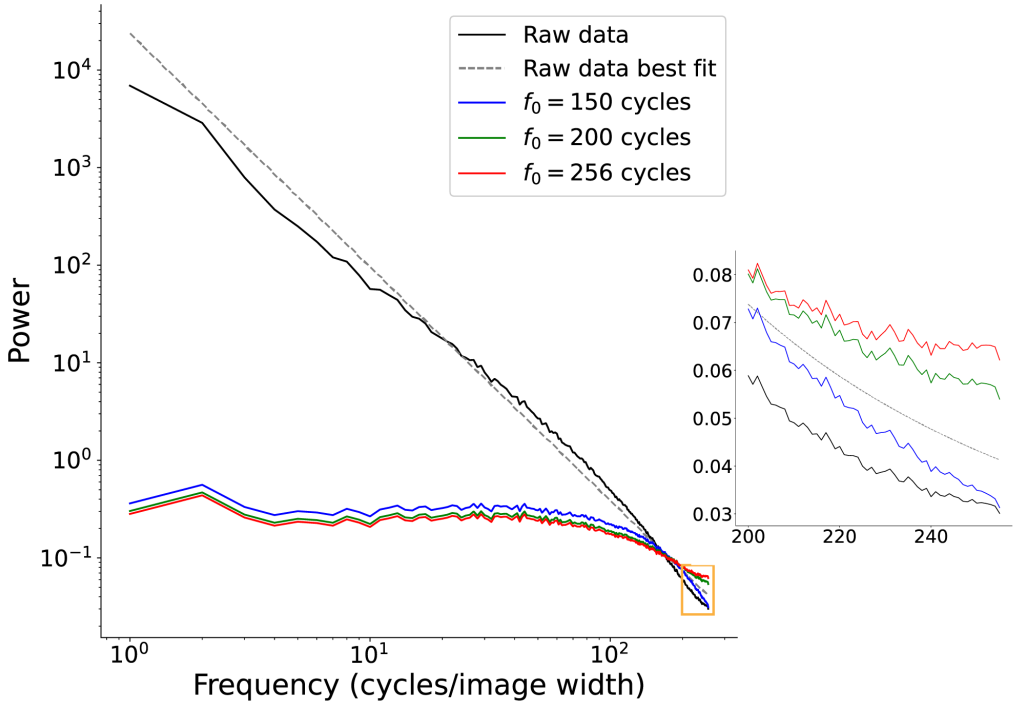


Figure 2.10: **Power spectrum of natural image data before and after whitening with various cutoff frequencies f_0.** We perform a best line fit ($R^2 = 0.99$) to the average power spectrum of the raw, pre-whitened images and find that the average estimated power $\hat{P} = 10.1/f^{2.39}$, slightly steeper than the theoretical $1/f^{2-\eta}$ power law. *Inset:* The power spectra at high frequencies. Note the order of the traces is reversed, with the higher cutoff frequency data having higher power in this range.

full images themselves. The average power spectra taken over a sample of 100 natural image patches drawn from the full whitened images does in fact have a clear power law (Fig. 2.11) that is consistent with the observed spectral bias in training.

That being said, this fact alone doesn't directly show that the power spectrum of the training data is sufficient to produce the effect. Without any other evidence, it's still possible, for instance, that lower frequencies are learned first, regardless of the frequency content of the data. To more precisely interrogate of how the power spectra of the data affect the spectral bias during training, we must devise a method to manipulate directly the power spectrum of the training data and observe the effect on basis function convergence.

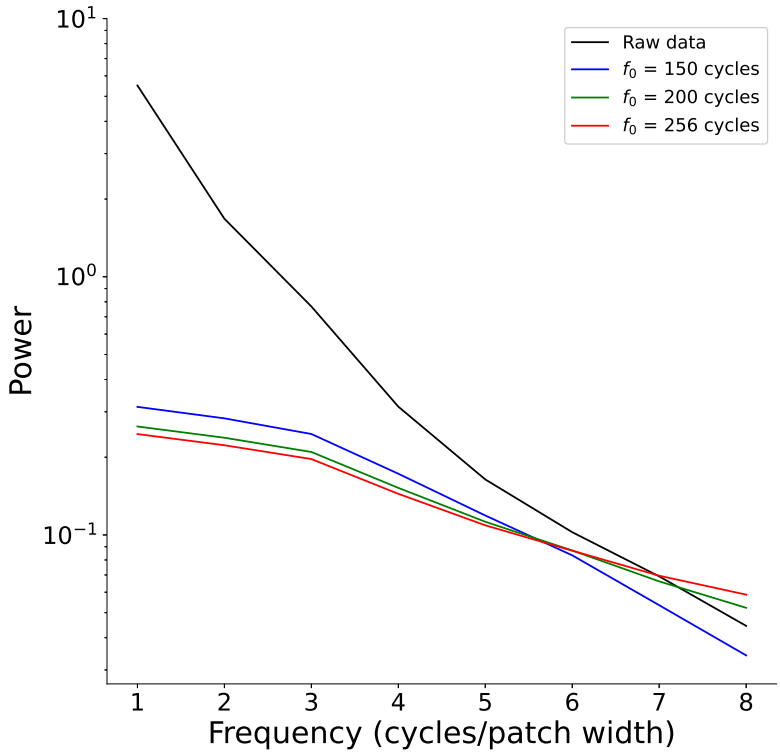


Figure 2.11: **Power spectrum of natural image patches.** Mean power spectra across 100 16×16 patches drawn from the full whitened 512×512 images obey a clear power law.

The input data power spectrum fully accounts for the spectral bias

To directly examine the effect of the frequency content in the data on basis function convergence, we generate synthetic 1-D training data with varying power spectra. The setup is similar to the one used in [37]: given frequencies $\mathbf{k} = (k_1, k_2, ...)$ with associated amplitudes $\mathbf{A} = (A_1, A_2, ...)$ and phases $\mathbf{\Phi} = (\Phi_1, \Phi_2, ...)$, we generate data vectors $\boldsymbol{\lambda}$ according to the function

$$\boldsymbol{\lambda}(z) = \sum_{i}^{N} A_i \sin\left(2\pi k_i z + \Phi_i\right), \tag{2.17}$$

where z is a positive, non-zero integer equal to the index of the corresponding vector element. To create a training set, we generate 200000 unique vectors $\boldsymbol{\lambda}$, each with $\mathbf{k} = (1, 2, ..., 14, 15)$, $N = 256$, and $\Phi_i \sim U(0, 2\pi)$. We generate three such training sets: 1) low frequency, where $\mathbf{A}$ falls off with increasing $\mathbf{k}$ according to a power law $P(k)$; 2) flat,

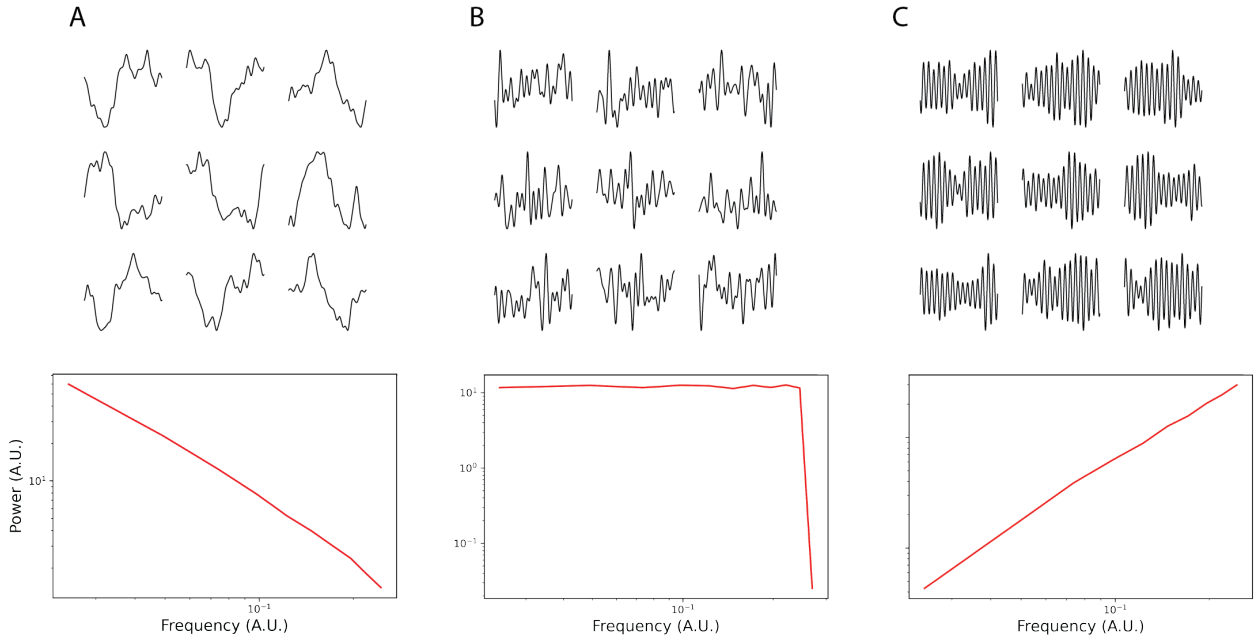


Figure 2.12: **Examples drawn from three synthetic 1-D datasets.** A) Low frequency B) Flat C) High frequency.

where $\mathbf{A}$ is equal at all $\mathbf{k}$; and 3) high frequency, where $\mathbf{A}$ increases with $\mathbf{k}$ according to a power law $P'(k)$, which is obtained by reversing the order in which power is assigned to frequency in $P(k)$ (Fig. 2.12). The high frequency data is generated this way so that the high and low frequency datasets have symmetric power laws. Each $\boldsymbol{\lambda}$, regardless of dataset, is normalized to have the same total power as every other $\boldsymbol{\lambda}$.

For Sparsenet, we use synthetic data generated according to $P(k) = 1/k^{1.3}$. We train a model with 512 1-D basis functions for 100 iterations with a batch size of 100 vectors, 300 iterations of inference per batch, sparsity penalty $\lambda = 0.3$, coefficient learning rate $\eta =$ 5e-4, and a dictionary learning rate $\alpha =$ 1e-3. The basis functions in the model are initialized with Gaussian-distributed white noise. If the spectral bias during training occurred independently of the power spectrum of the synthetic data, we would expect that 1) the low and high frequency convergence plots and heatmaps would be asymmetric, with low frequency basis functions converging sooner for the low frequency training set than high frequency basis functions for the high frequency training set; and 2) low frequencies converge sooner for the flat training set. However, we do not observe either. Rather, the low frequency and high frequency convergence plots and heatmaps are nearly perfectly symmetric, and all basis functions converge at a nearly equal rate for the flat dataset, with no clear frequency-dependent pattern emerging (Fig. 2.13). These results suggest that the spectral bias is entirely determined by the input data statistics.

For SAILnet, we use synthetic data generated according to $P'(k) = 0.1k$. Due to the

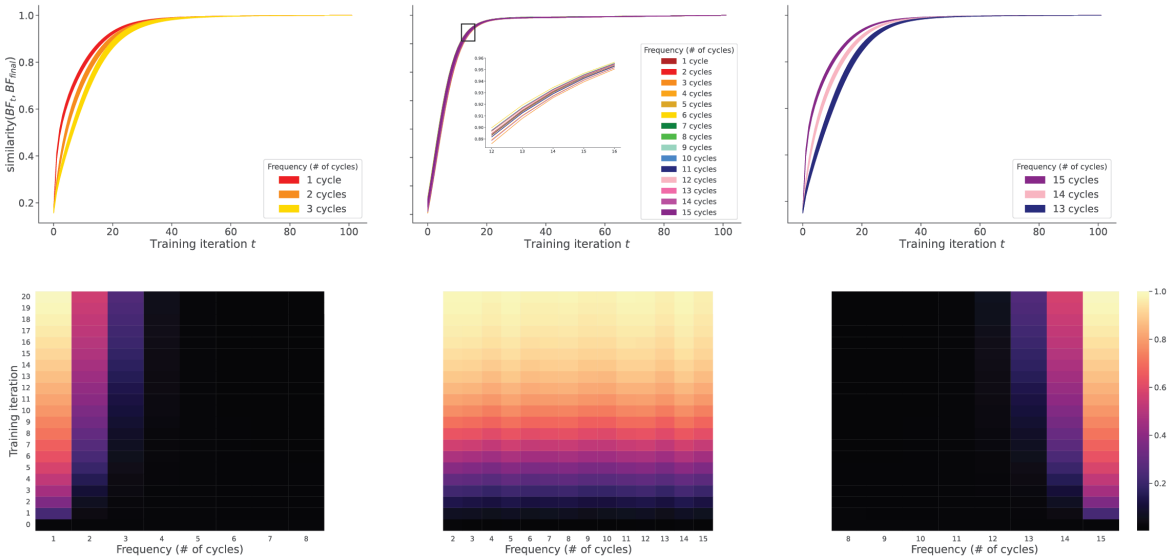


Figure 2.13: **Convergence of 1-D basis functions in Sparsenet.** *Top:* Convergence plots for the low frequency (left), flat (middle), and high frequency (right) training sets. Inset of mean similarities (plotted without standard deviations to clearly show the mean values by category) for the flat dataset shows no clear frequency-dependent effect on convergence, with all basis functions converging at the same rate. *Bottom:* Corresponding power heatmaps showing power spectra of learned dictionary over training.

well-documented poor performance of SAILnet on datasets with steep power laws [43], a gentle, linear power law was necessary to achieve learning while still being able to visualize the separation in convergence between basis functions. We train a model with 2048 1-D basis functions for 1000 iterations with a batch size of 100 vectors, inhibitory synaptic learning rate of 0.2, excitatory synaptic learning rate of 0.002, threshold learning rate of 0.02, lifetime sparseness parameter $p = 0.05$, and initial firing thresholds $\theta_0 = 2.0$. We do not employ a learning rate schedule, since we only train for a short period of time, and good convergence was achieved on the training data without decreasing the learning rates at later iterations. As in Sparsenet, we observe a symmetry in the low and high frequency regimes, as well as equal convergence for all basis functions on the flat dataset (Fig. 2.14). Our results in SAILnet suggest that even under synaptically local learning rules and nearly perfectly whitened data, the spectral bias is entirely determined by input data statistics.

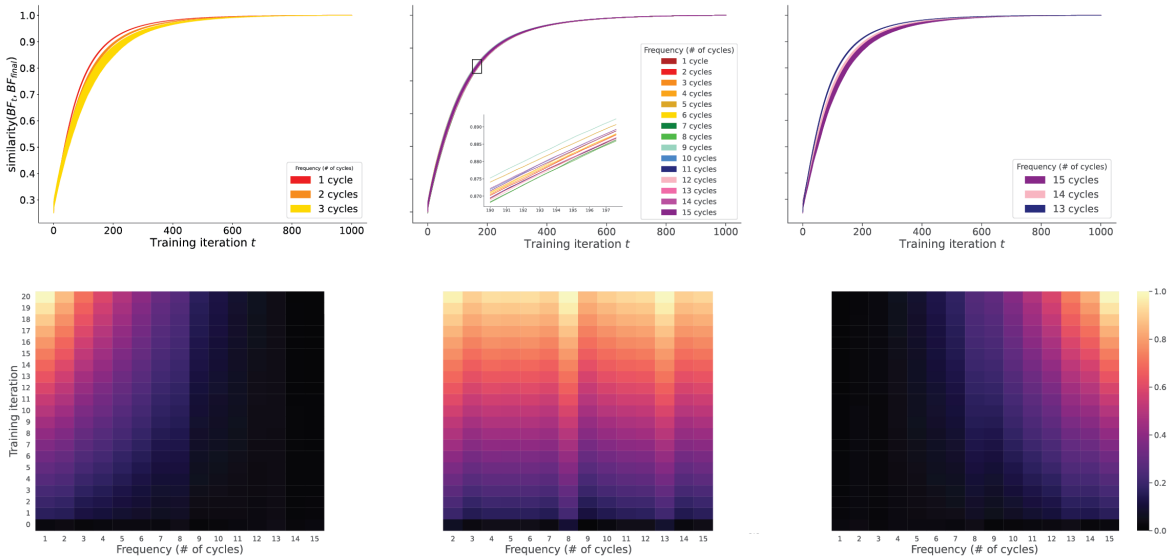


Figure 2.14: **Convergence of 1-D basis functions in SAILnet.** *Top:* Convergence plots for the low frequency (left), flat (middle), and high frequency (right) training sets. Inset of mean similarities (plotted without standard deviations to clearly show the mean values by category) for the flat dataset shows no clear frequency-dependent effect on convergence, with all basis functions converging at the same rate. *Bottom:* Corresponding power heatmaps showing power spectra of learned dictionary over training. Despite the gentle linear power law of the training set, the convergence of SAILnet basis functions by frequency is still highly sensitive to the input data statistics.

2.5 Implications for Experience-Dependent Development of the Visual System

In this chapter, we demonstrate that sparse coding models learn basis functions in a hierarchical manner: lower frequency basis functions are learned early in training, and higher frequency basis functions are learned later in training. Given the substantial algorithmic differences between Sparsenet and SAILnet, these results suggest a general property of sparse coding, rather than a property of a specific architecture or optimization procedure. In addition, because SAILnet is a biologically plausible model of sparse coding, it further suggests that this observed spectral bias may be a feature of RF development in V1.

Prior work has also used SAILnet as a model of V1 development using a similar approach. By tracking the changes in sparseness of the learned representations throughout model training, it has been shown that SAILnet recapitulates the experimentally observed decrease in sparsity of V1 neural encodings over development in ferrets [50]. Traditional sparse coding has also been used to understand the role of visual experience in devel-

opment. Previous work has shown that training sparse coding models with unnatural training images results in basis functions resembling the RFs that arise when animals are reared with abnormal visual input (such as only being able to view stimuli consisting of a single orientation), suggesting that sparse coding is a feature of experience-dependent development [51].

Our results make a prediction for the timecourse of development of V1 neurons that is consistent with experimental evidence. It has been shown that during development, the distribution of frequency tuning of V1 neurons shifts towards higher spatial frequencies, and this shift requires visual experience [30, 31]. However, the question remains whether this shift is due to high spatial frequency RFs emerging later after the early development of low spatial frequency RFs, or whether there is a global shift during development across all receptive fields towards higher spatial frequencies. Our results may provide additional insight into this phenomenon because we are able to directly observe the development of each individual basis function in the model, as opposed to just sampling from the distribution of tuning across the V1 neuronal population at different timepoints during training. In particular, our results suggest that this shift in the distribution is due to higher frequency receptive fields emerging later than the low frequency receptive fields. Future experimental work can help distinguish whether one or both of these explanations can account for the observations in [30] and [31]. It may be experimentally challenging to track individual neuronal receptive fields over the full course of development, which would be the ideal way to tackle this problem. Whether or not this can be done, it should be possible to sample from the population of receptive fields at various points in development and estimate the relative proportions of low, mid, and high frequency receptive fields at each time point. This could provide indirect evidence for one or the other of these possibilities, depending on the details of the distribution of RF shapes.

Here, we only consider the development of V1 simple cells, rather than complex cells or neurons in higher visual areas. Other sparse coding models have the capacity to learn complex cell receptive field properties and topography [52]. We also do not consider excitatory connections between cells, which are a feature of the sparse coding model described in [53]. Future work could analyze development of the basis functions in these extensions of sparse coding. We also do not account for the temporal properties of receptive fields, since the receptive fields we model here only have spatial dependence, not time dependence. Indeed, accounting for the full spatio-temporal receptive fields of V1 simple cells would give us a more complete picture of spectral bias during development, as it has been shown that a large proportion of V1 neurons are "two-peak" cells in that they shift their preferred tunings to higher spatial frequencies over the stimulus period; meanwhile, "one-peak" cells, which have peak responses that occur at the same time, prefer lower spatial frequencies [54]. However, results are still consistent with these experiments, which show that visual experience increases the relative proportion of two-peak cells, and therefore the relative proportion of cells that are tuned for higher spatial frequencies.

As discussed in Section 2.3, for certain hyperparameters, individual model neurons in SAILnet do not each converge to one final learned shape. Rather, individual neurons fluc-

tuate, morphing from one shape to another throughout training; these fluctuations do not terminate even after training for many iterations. We refer to this phenomenon as fluidity. This is in contrast to Sparsenet, for which we have only observed smooth convergence of basis functions to their respective final shapes after many training iterations. We note that when the basis functions are fluid during training, the metric of similarity over training time, and by extension, most simple metrics of convergence, are no longer meaningful, as there is no point at which every basis function has fully converged. Therefore, our main findings hold for a set of parameters and initial conditions that are sufficient to suppress fluidity. Whether fluidity is a biological phenomenon is an interesting open question that we hope will be the subject of future experimental work.

Finally, there are many factors potentially affecting spectral bias in biological development that we do not consider in our modeling. For example, increasing spatial acuity occurs during development [55], both in terms of the optical properties of the eyes [56] and changes in visual perception from infancy to adulthood [57], both of which are likely to influence the ability to distinguish between visual stimuli. Moreover, changes in the RFs of neurons upstream of V1, such as in the retinas or the lateral geniculate nucleus of the thalamus, are undoubtedly changing during development.

Our spectral bias prediction is derived in the context of experience-dependent development, during which neuronal tuning adapts to natural scene statistics. It is possible that this particular order of development may not hold for experience-independent development, such as occurs in V1 prior to eye opening. This question could be addressed by considering different input data to the model. One possible input could be internally generated spontaneous neural activity, such as retinal waves, which play a role in the wiring of circuitry in early visual areas. For example, Dähne and colleagues implement slow feature analysis to encode retinal wave signals and find that the learned features correspond to the shapes of V1 complex cells [58]. We explore the role of spontaneous neural activity in the development of visual functions in the next chapter.

Chapter 3

Efficient Representation Geometry Emerges from Structured Spontaneous Neural Activity

3.1 Chapter Summary

At the end of the previous section, we discussed how receptive fields can be learned via spontaneous neural activity, which suggest that innate mechanisms prior to visual experience prime V1 for feature extraction on external stimuli [59, 58]. We now ask whether the internal representations that support higher-level visual function in the cortex can also be learned via innate mechanisms.

One such high-level visual function is object recognition. The visual system has an extraordinary capacity for rapidly and accurately recognizing distinct objects in the face of identity-preserving transformations [60, 61, 62]. Evidence suggests that this is a result of efficient representation: neural recordings reveal a high degree of linear separability between neural responses to different stimuli [63, 61]. More precisely, sensory representations of distinct objects in the early visual system are tangled together and gradually untangle as they are transformed and re-mapped in a feedforward manner along the ventral stream [61]. However, the manner in which such representations are learned in the brain is still unknown.

Models trained to classify images can perform invariant object categorization at near human-level accuracy [64]. However, the supervised learning methods used to train these models are unlikely to explain how the brain learns object recognition, given that large amounts of labeled examples are not necessary for visual development [65, 66, 67, 68]. In this work, we explore the potential of innate neural activity as pre-training data for neural networks and ask whether the internal representations that enable object recognition can be learned without access to any external visual information.

The motivation for this work is grounded in developmental neurobiology. Many key

aspects of visual system organization are well-established before visual experience, such as topographic maps, orientation selectivity, and ocular dominance [14]. Notably, axon targeting can largely be learned by innately generated signals such as spontaneous neural activity and molecular guidance cues [69]. These findings suggest external stimuli are unnecessary for the initial development of the early visual system.

Here, we investigate whether a particular form of spontaneous activity known as retinal waves can instruct formation of the feed-forward connections that support object recognition. Retinal waves are a developmental phenomenon characterized by correlated patterns of propagating, network-level activity among groups of retinal ganglion cells (RGCs) prior to eye-opening [70]. Experimental and computational evidence suggests that retinal waves instruct the formation of retinotopic maps, enabling RGC axons to reach their targets in the superior colliculus and lateral geniculate nucleus before the onset of visual experience [15, 16, 17, 18, 19, 71]. Given that 1) higher-order feature extraction in the visual system presumably depends on the representations induced by these axonal projections and that 2) these projections are well formed prior to visual experience, in this chapter we explore whether retinal waves are sufficient for learning the mappings that enable object recognition.

Our core result in this chapter is that networks pre-trained on movies of retinal waves produce more linearly separable representations of natural images compared to randomly initialized networks, despite the fact that these representations were never trained on natural images. This task-independent phase is meant to simulate the experience-independent period of visual development prior to eye-opening. To quantify the efficiency and robustness of the learned representations, we turn to the framework of manifold geometry, which we present in the first section. Manifold geometry is a statistical framework for determining the separability of object representations in high-dimensional feature space [72, 73].

We characterize the geometry of the networks' internal feature representations in two networks: a simple network with one hidden layer pre-trained on retinal wave patterns via Hebbian learning [74, 75], and a DCNN pre-trained on retinal wave patterns via a contrastive learning objective [76]. In both cases, we find the efficiency of the learned representations increases network performance on a set of image classification tasks, particularly classifying noisy data (in the case of the Hebbian network) and classifying spatially translated data (in the case of the DCNN). Our results suggest that the spatiotemporal information in retinal waves is relevant for object recognition in natural scenes and point towards an instructive role for retinal waves during early synapse formation in visual circuits.

3.2 The framework of manifold geometry

The neural population response to different presentations of the same perceptual object under different transformations — such as orientation, pose, lighting and location — con-

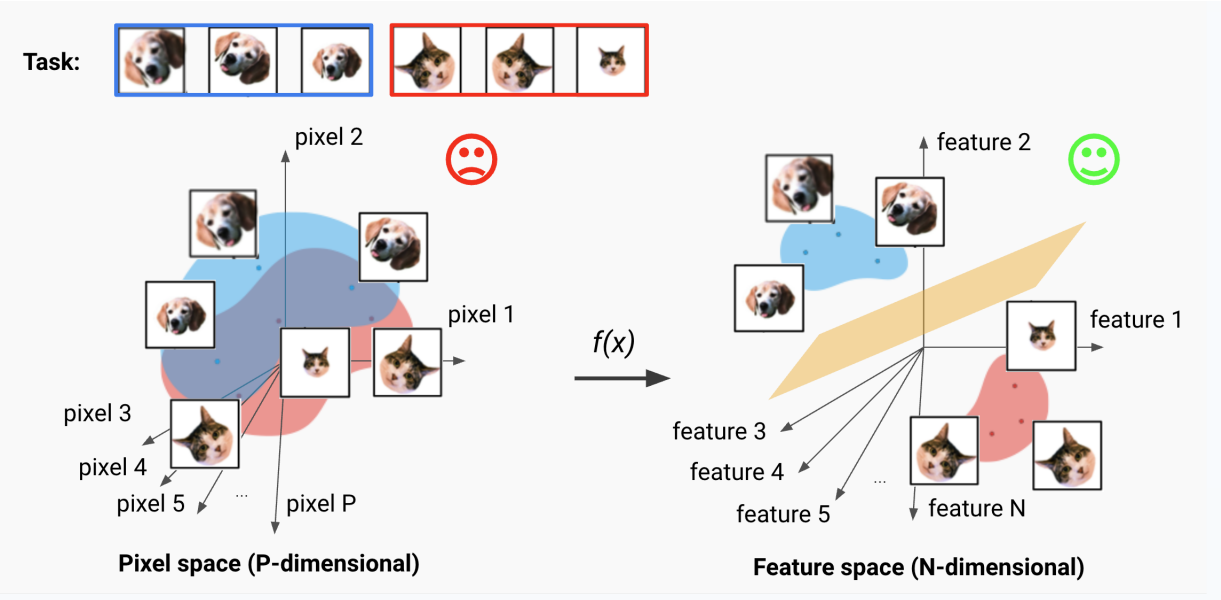


Figure 3.1: **Binary classification of 2 object manifolds.** The task of discriminating between the dog and cat image manifolds can be thought of as finding re-mapping $f(x)$ of the data x into a space where the manifolds are more easily separable. On the left, the two manifolds — each consisting of different presentations of the same object in varying scales and orientations — are highly tangled in pixel space, making them difficult to separate with a linear classifier. On the right, a transformation by a well-trained $f(x)$ compresses and pushes apart the manifolds in feature space, enabling classification with a linear hyperplane.

stitutes a neural object manifold (Fig. 3.1). Discriminating between different objects is therefore a problem of separating object manifolds. This is analogous to finding a separating hyperplane in the perceptron problem, only instead of the counting units for data being individual points, they are manifolds of different objects. The theory of manifold geometry provides a statistical framework to quantify the linear separability of these manifolds as a function of their geometry [73]. We examine three quantities of manifolds that determine their separability, namely the capacity α_c, the dimension D_M, and the radius R_M.

Activation extraction: For all theoretical manifold quantities, the outputs of the intermediate network layer activations are extracted to analyze the internal representations of the task or wave manifolds at each layer.

Capacity α_c: We consider a set of P object manifolds fully linearly separable if they can be classified into binary classes by a hyperplane in N-dimensional feature space. The theory of manifold geometry shows that the value of the manifold capacity α_c determines the ex-

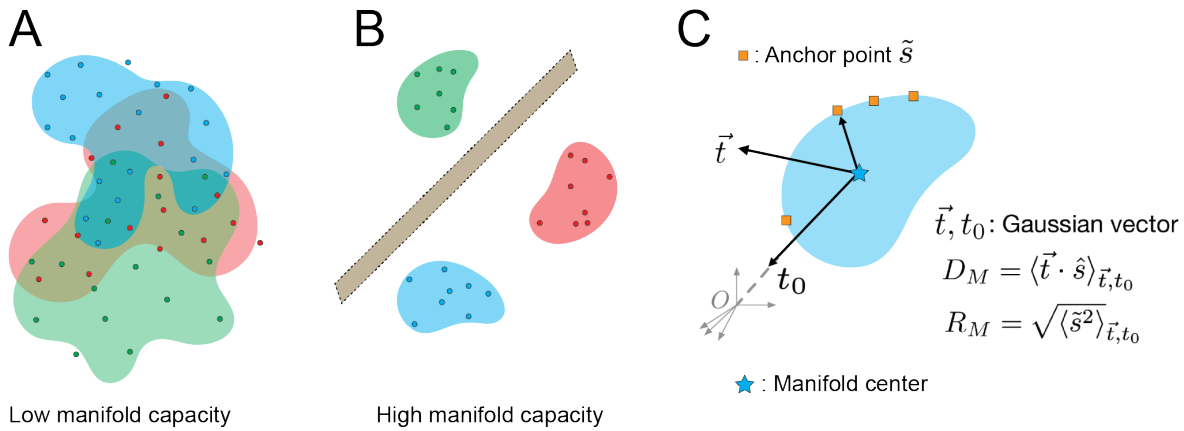


Figure 3.2: **Illustration of point cloud manifolds**. (A) Tangled manifolds exhibit low capacity. (B) Untangled manifolds exhibit high capacity and are separable by a hyperplane (C) Manifold dimension measures the spread of anchor points across the manifold axes by projection of a Gaussian vector onto an anchor point. Manifold radius measures the norm of an anchor point in the manifold subspace. These two geometrical quantities determine the manifold capacity.

tent of separability in the limit of large P and N: if $P/N < \alpha_c$, the manifolds are separable with high probability; if $P/N > \alpha_c$, the manifolds are inseparable with high probability. Therefore, the higher the value of α_c, the higher the probability of separability for a given set of manifolds (Figs. 3.2A,B). For point-cloud manifolds, in which each manifold consists of M data points each corresponding to an example of the given object, the capacity can be shown to be bounded as $\frac{2}{M} \leq \alpha_c \leq 2$ [72]. The theory of manifold geometry also shows that capacity is determined by two quantities which describe the geometry of the object manifolds in N-space: the dimension D_M and the radius R_M. These are statistical quantities defined for each manifold by considering spread of points in the manifold's convex hull, called anchor points, over variations in the manifold's labeling and location in N-space (Fig. 3.2C). For large N, α_c is inversely proportional to $\sqrt{D_M}$ and R_M [77]. All three quantities — α_c, D_M, and R_M — are estimated using algorithms based on statistical mechanical mean-field techniques described in [78].

Dimension D_M: Dimension is the spread of anchor points across the manifold axes and estimates the average embedding dimension of the manifold (Fig. 3.2C).

Radius R_M: Radius is the average distance between the manifold center and anchor points and reflects the scale of the manifold compared to the overall data distribution. (Fig. 3.2C).

Simulation capacity α_{sim}: We note that α_c is a theoretical estimate of linear separability that may deviate from the true capacity in the regime of finite manifolds P and feature dimensions N [72]. Simulation capacity provides a numerical approximation of the ground-truth manifold capacity. We calculate simulation capacity by first running linear

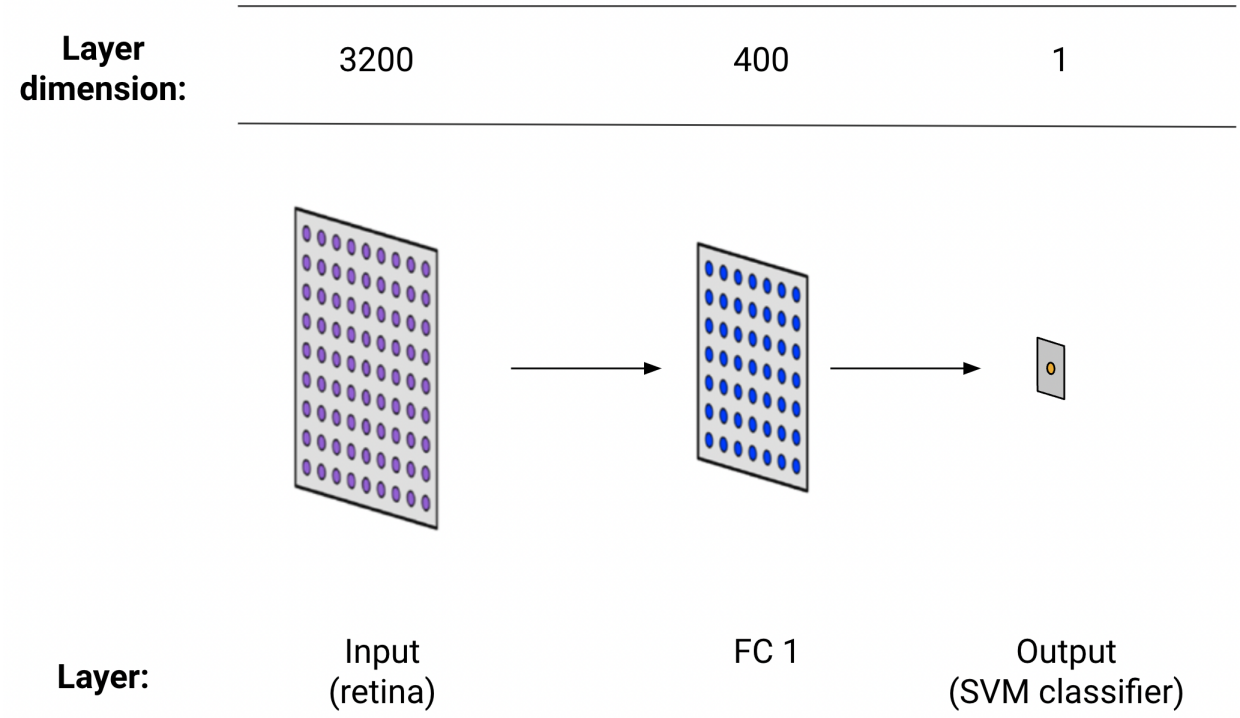


Figure 3.3: **3-layer feed-forward network with ReLU activations after the hidden (FC1) layer.**

classifications with fixed P and varying N until the probability of manifold separation converges to 0.5. The final value of $N = N_c$ is used to calculate the simulation capacity $\alpha_{sim} = P/N_c$.

3.3 Analysis of a simple linear network

Architecture

The first model architecture we consider is a feedforward network consisting of an input layer, a layer with pre-trained weights (the receptive fields) and ReLU activations, and a support vector machine (SVM) as the classifier/output layer (Fig. 3.3). The input layer consists of 3200 model retinal ganglion cells randomly placed on a square grid. The hidden linear layer consists of 400-units with ReLU activations. The output/classifier layer consists of 1 unit.

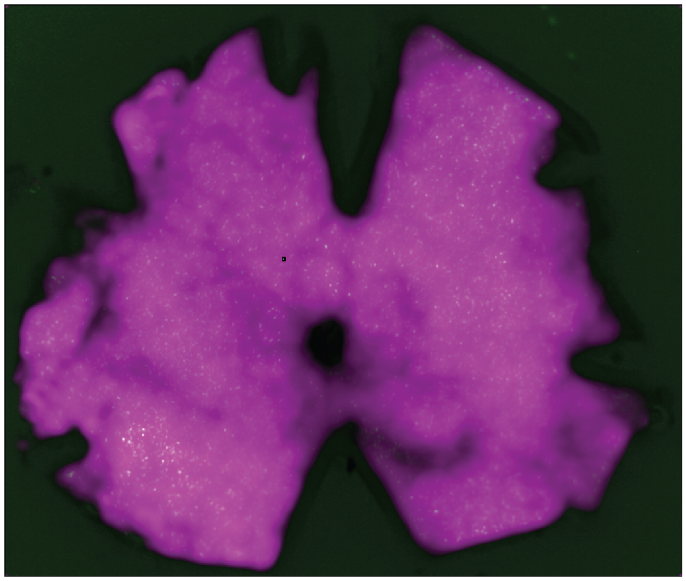
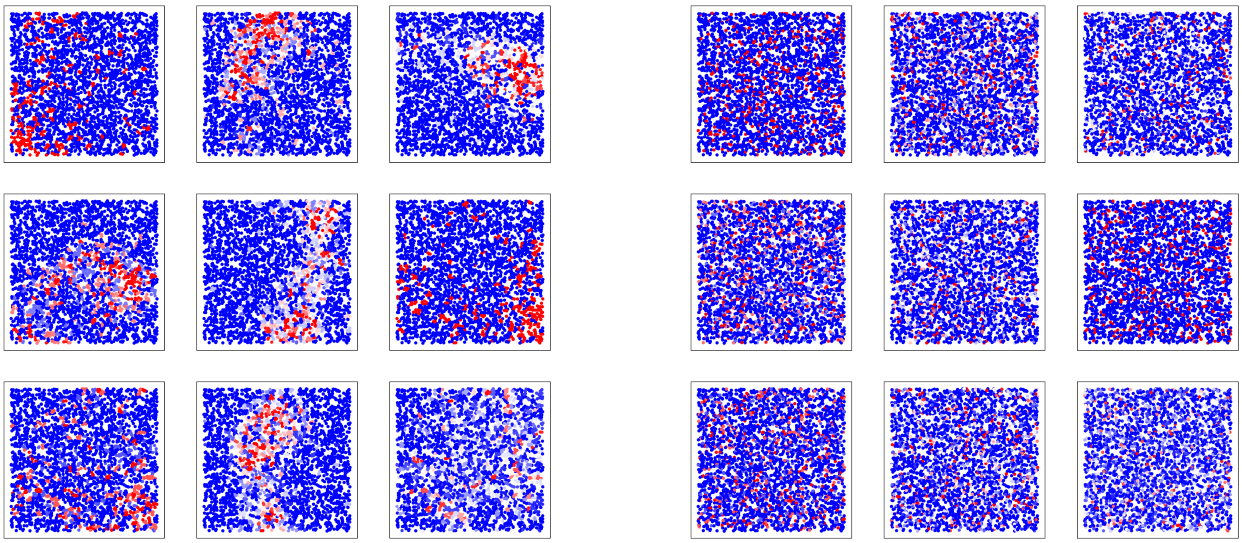


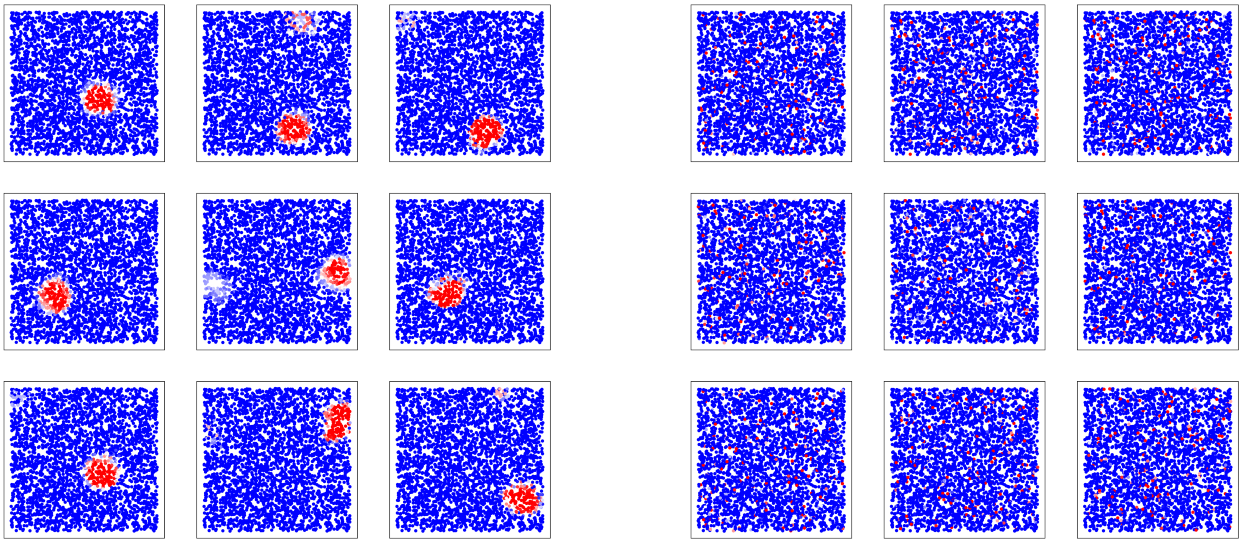
Figure 3.4: **Area of an isolated retina used to obtain real retinal wave data.** Retina (11 mm^2) shown in pink.



(a) Pre-trained (real)

(b) Scrambled (real)

Figure 3.5: **Receptive fields of hidden layer weights (real waves).** Left: Receptive fields for Pre-trained (real) network. Right: Receptive fields for Scrambled (real) network, which are obtained by shuffling the pixels of the receptive fields in the Pre-trained (real) network.



(a) Pre-trained (sim.) (b) Scrambled (real)

Figure 3.6: **Receptive fields of hidden layer weights (simulated waves).** Left: Receptive fields for Pre-trained (sim.) network. Right: Receptive fields for Scrambled (sim.) network, which are obtained by shuffling the pixels of the receptive fields in the Pre-trained (sim.) network.

Pre-training on retinal waves

To pre-train the network, retinal waves are propagated across the input layer to train the weights (receptive fields) to the first hidden layer units by a winner-take-all Hebbian learning rule as in [75]. Retinal waves are obtained from epifluorescent macroscope calcium imaging of mouse retinas (Fig. 3.4). For comparison, we also pre-train a network on simulated spontaneous activity according to the structured noise dynamics in [75]. We analyze five networks: "Pre-trained (sim.)", a network with receptive fields pre-trained on simulated retinal waves according to the dynamics in [75] (Fig. 3.6a), "Pre-trained (real)"; a network with receptive fields pre-trained on real retinal wave data (Fig. 3.5a); "Scrambled (sim.)" and "Scrambled (real)", networks in which each receptive field is the result of randomly permuting the pixels of its corresponding receptive field in "Pre-trained (sim.)" and "Pre-trained (real)", respectively (Figs. 3.6, 3.5); and a network whose hidden layer is a Gaussian random projection "Control (rand. proj.)", which reduces dimensionality while preserving geometrical properties of the input [79].

Task training on MNIST

After pre-training, we train the SVM to classify MNIST digits, which are pre-processed by binarizing the pixel values and projecting them onto the input layer units. The net-

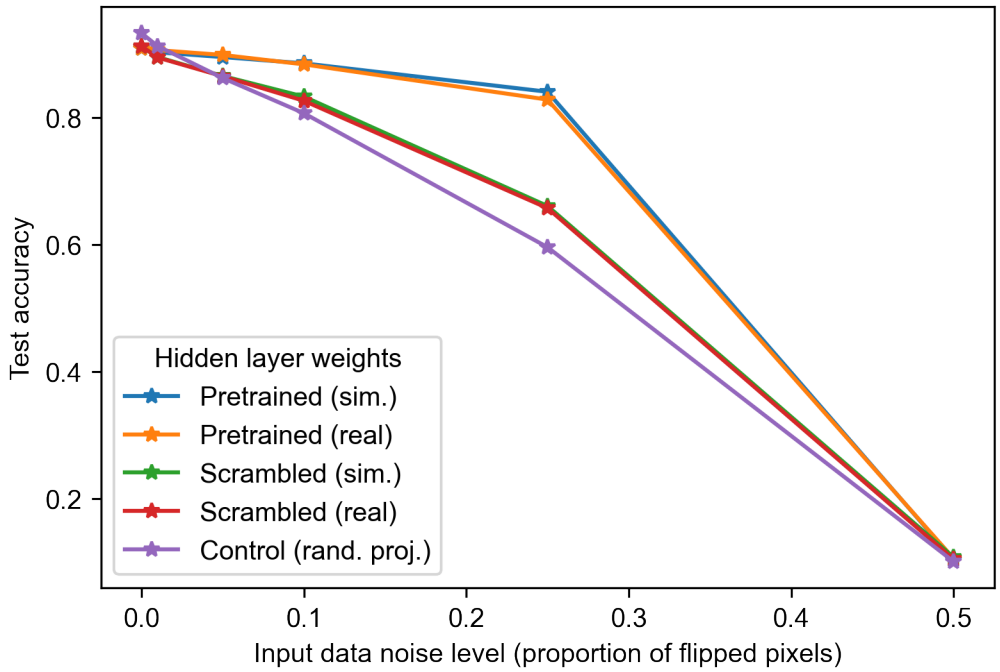


Figure 3.7: **MNIST classification accuracy with increasing noise.** While all networks have similar accuracy in the 0-noise regime, networks pre-trained on retinal waves have the highest robustness to noise perturbations as noise increases.

work is trained by freezing the input and hidden layer weights as is, doing a forward pass of the projected task data through the second layer, and training the SVM layer on the 400-dimensional hidden layer output representations of the data. Only the SVM layer is trained on the MNIST data, such that it learns to classify the network's internal representations of the digits. Once trained, models are then evaluated on test data. The training set consists of 60,000 samples and the test set consists of 10,000 samples.

All networks considered have the same width and depth. We find that classification accuracy is higher for pre-trained networks relative to their scrambled counterparts (Fig. 3.7. Pre-trained networks also maintain higher classification performance given noisy versions of MNIST data, which are generated by bit-flipping a randomly selected proportion of the pixels. Interestingly, in the 0-noise regime, the networks all have similar performance and in [75], it was found that pre-training on what we are calling the Pre-trained (sim.) data was slightly better than random (all networks were reported to have between 89 and 92 % test accuracy). It's possible that given the relatively simple dataset (MNIST), pre-training does not substantially increase performance — however, adding noise to the data increases the performance gap, as we show here. Similar noise robustness has been observed for sparse coding, which has been applied to de-noising tasks [35]. Nevertheless, we build on this finding by analyzing the manifold geometry of the object manifolds

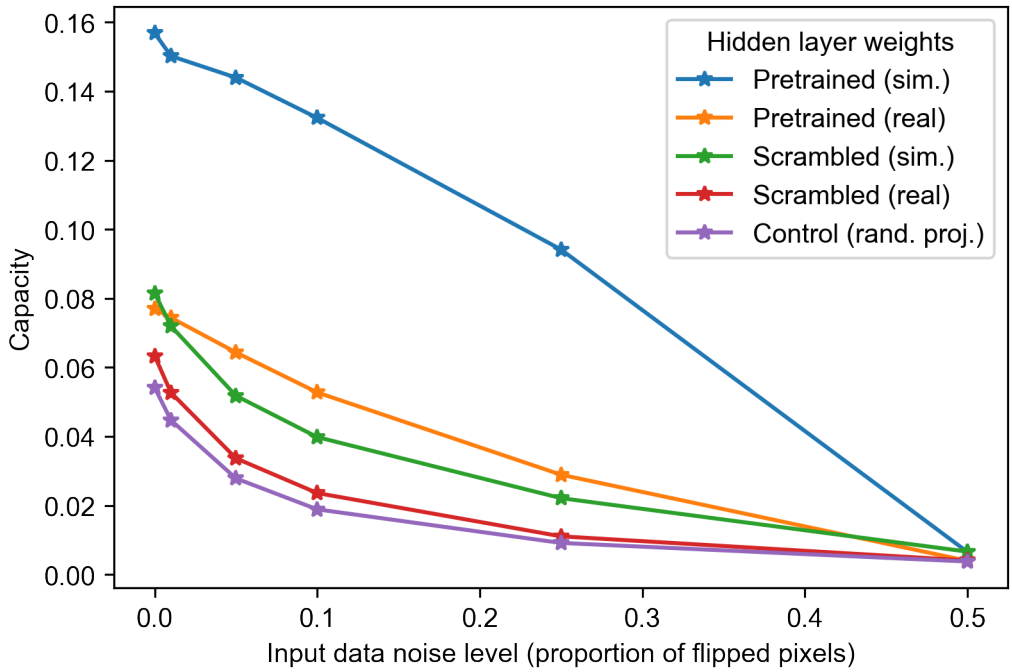


Figure 3.8: **Capacity of object manifolds with increasing noise.** The pre-trained networks have manifolds with higher capacity relative to their scrambled counterparts.

as represented by the network's hidden layer weights, which can give us a richer insight on the effects of pre-training on representation.

Manifold analysis

To examine how pre-training with retinal waves affects the geom- etry, and in turn the separability, of neural object manifolds for each task, we extract the network activations at the ReLU layer for P = 10 manifolds (each corresponding to an MNIST digit) consisting of M = 100 examples. Based on these digit manifolds, we analyze three manifold properties of the networks' internal representations of the MNIST data: capacity, defined as the maximum number of object manifolds that can be linearly separated using random binary labels divided by the dimension of the representation; manifold dimension, the spread of anchor points (which define the optimal separating hyperplane between two perceptual manifolds) along the manifold axes; and manifold radius, the variance of anchor points normalized by the average distance between manifold centers. For each of these quantities, we report the average across the MNIST digit classes/manifolds.

The lower the dimension and radius of the object manifolds, the more linearly separable they are, yielding higher manifold capacity (Fig. 3.2). We show that the internal representations of pre-trained networks generally exhibit higher capacity, lower dimen-

sion, and lower radius relative to those of their scrambled counterparts, particularly in the presence of moderate noise. Notably, the pre-trained networks exhibit consistently lower dimension than the control (random projection) network (Fig. 3.10), but lower radius only in the presence of considerable noise (Fig. 3.9). Moreover, the dimension of the networks pre-trained on simulated waves tends to be lower than that of networks pre-trained on real waves, while the real wave networks tend to exhibit lower radius.

Our results show that pre-training with both simulated and real retinal waves yields receptive fields with spatial structure favorable for separation and classification of object manifolds. The purpose of using simulated waves is a groundtruth to determine whether receptive fields are sufficiently in response to a simple pattern, as the real wave data contains noise that make it harder to determine whether learning has occurred based on inspection of the network weights. Because we are comparing to scrambled networks, whose receptive fields share the same pixel distribution as the pre-trained networks but lack their spatial structure, these results suggest that learning the spatial structure in retinal wave patterns, as opposed to just the overall distribution, is relevant for object recognition. We also find that networks trained on simulated waves exhibit higher accuracy and manifold capacity compared to networks trained on real retinal waves. We suspect this is because the real data contains considerably more noise and is less local in its spatial structure, which may lower performance on the relatively simple binarized MNIST dataset.

Interestingly, it appears simulated wave and real wave networks have different effects on the manifold geometry. Simulated wave networks have a greater effect on reducing the manifold dimension, while real wave networks have a greater effect on reducing the manifold radius. The disparity in dimension reduction may be due to the local structure of the simulated wave receptive fields, which could act as feature detectors for low dimensional structures induced by correlated nearby pixels. The reasons for disparity in radius reduction do not appear as straightforward, considering that the random projection network tends to exhibit lower radii. One factor could be that the high magnitude of the synaptic strengths induced by the slower, localized simulated waves increases the effective sizes of the data manifolds and thus their radii. The real retinal waves propagate less frequently and diffusely, so the synaptic strengthening during the Hebbian learning phase occurs at a lower rate. This explanation is consistent with the fact that the random projections have unit magnitude and generally lower radii, though it is not clear why their radii are higher in noisier regimes. Normalizing the receptive fields in the pre-trained networks to control for the effect of synaptic strength on radius may elucidate these questions

Finally, we highlight that the task (as well as the dynamics of the simulated retinal waves) were chosen as a simple benchmark to compare with previous similar work [75] and as a proof of concept, limiting the scope of our findings. A more ethologically relevant task, like classifying natural images (without binarization of the pre-training and training data, as is done here), would be a more direct examination of the role of retinal waves in biological development and for this reason may be more amenable for networks pre-trained on real data. Another direction for future work is to examine the effect of network architecture in this regime, in particular by introducing layers that more accurately mimic

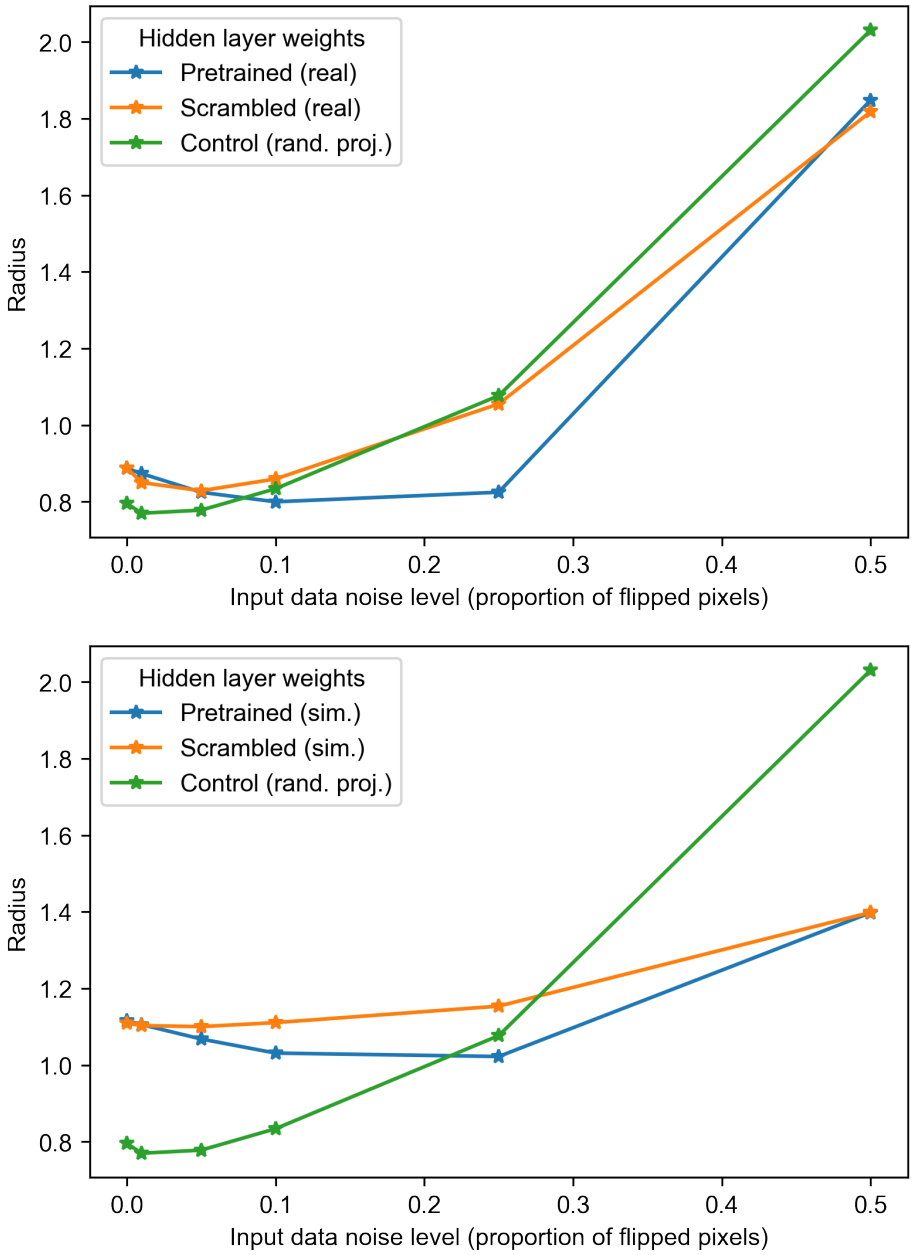


Figure 3.9: **Radius of hidden layer representations**. We report averages across digit classes for networks trained on real retinal waves (top) and simulated retinal waves (bottom).

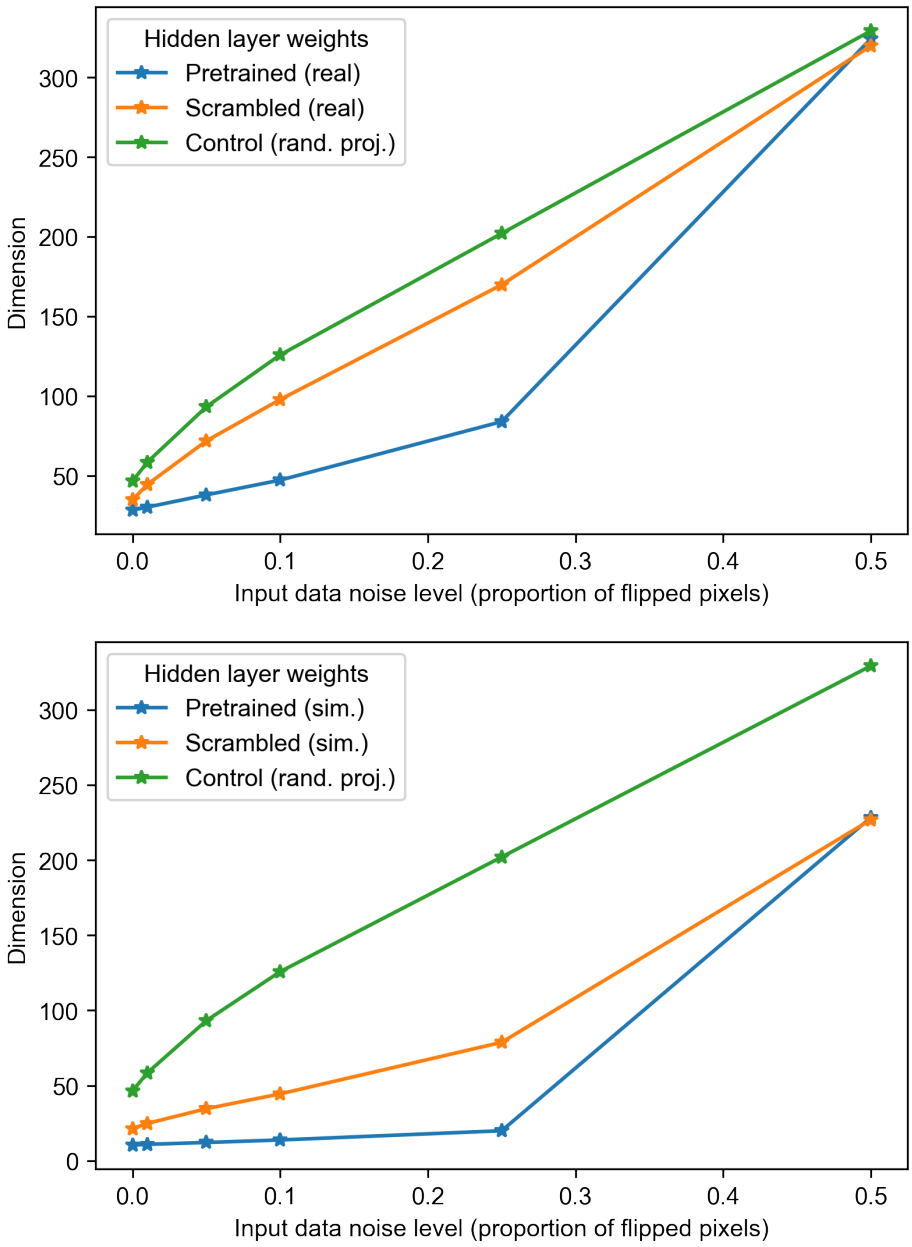


Figure 3.10: **Dimension of hidden layer representations**. We report averages across digit classes for networks trained on real retinal waves (top) and simulated retinal waves (bottom).

the structure and dynamics of higher visual areas like LGN and V1. We address this in the next section, where we perform a more in-depth analysis on a deep neural network.

3.4 Analysis of a deep network

Deep convolutional neural networks (DCNNs) can classify objects at near human-level accuracy [64] and have been shown to exhibit representations similar to neural activities in mammalian systems [80, 81]. Furthermore, DCNN layers have analagous properties to the visual hierarchy, whereby feature transformations at each layer induce linear separability in the object manifolds [82]. DCNNs therefore offer a useful testbed for modeling the visual system [83, 84, 85, 86]. These features of DCNNs allow us to explore how self-supervised pre-training on retinal waves affects representation for more complex data and tasks that require higher-order feature extraction.

Pre-training on retinal waves

To test whether spatiotemporal features of retinal waves learned during pre-training will improve performance on visual tasks, we follow the pipeline described in Fig. 3.11. Given a movie of a neurally active developing retina (Fig. 3.11A), we first train a ResNet-18 to compress temporally consecutive frames of the movie in output space, while pushing apart temporally distant frames (Fig. 3.11B) using the SimCLR training objective [76]. This is in accordance with the finding that temporally close activity bursts convey the most spatial information about relative RGC position [87, 88]. **Because retinal waves occur only before eye-opening in developing animals, this phase is meant to simulate the period of cortical development *prior* to visual experience.** We pre-train two kinds of networks: the first using macroscope movies of retinal waves obtained via calcium imaging of whole retinas dissected from postnatal mice, and the second using simulated movies of retinal waves from a parametrized, reaction-diffusion based model [89].

To filter out calcium transients, periods of inactivity, and random noise in the calcium imaging data, watershed image segmentation is used to identify periods of continuous retinal wave activity spanning a given number of frames, with each period denoted as a "wave event". We aggregate movies from four retinas, resulting in $\sim$60,000 total frames of real retinal wave pre-training data. The frames are downsized to 32×32 pixels.

We consider simulated retinal wave data generated using the model in [89] "out-of-the-box" (Fig. 3.12). The area parameter of the simulation is changed to match the average pixel-wise area of the four isolated real retinas, which was calculated using open source Fiji software and then converting to metric units based on macroscope resolution. The "strength" parameter α is modified to 0.5 to increase the wave frequency and eliminate long periods of inactivity. The model frame rate is matched to that of the macroscope data. The model is run to obtain a total of $\sim$237,000 frames of simulation data. Because the simulated data is far less noisy than the real data, wave events are simply taken as the sets

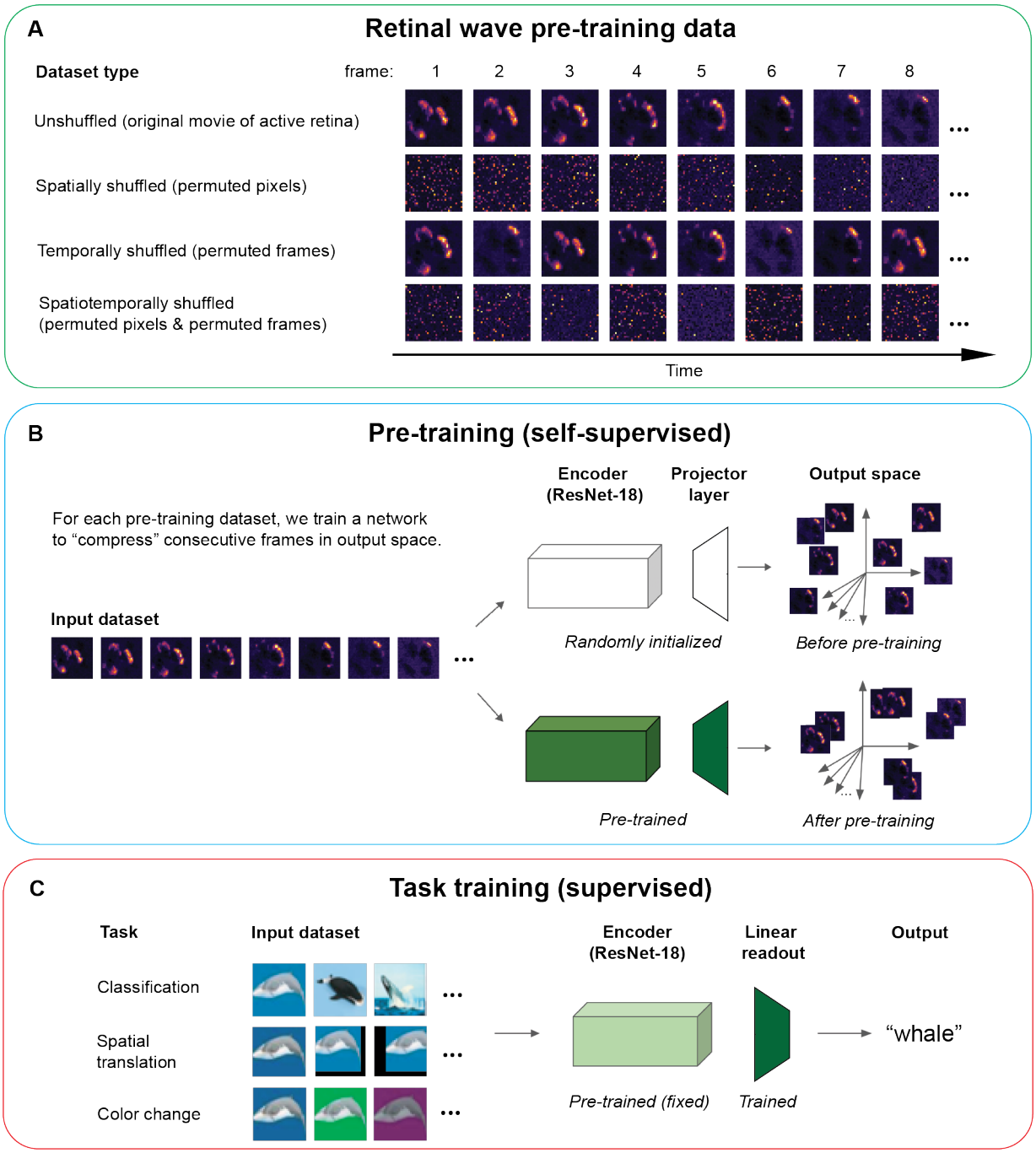


Figure 3.11: **Network training pipeline.** (A) Retinal wave movies and three permutations of the original movies are used as pre-training datasets. As an example, three permutations are shown on the same 8-frame excerpt taken from an original movie, which consists of consecutive frames of retinal wave activity. (B) Contrastive learning is used to train networks to learn temporally close spatial correlations in the movies. (C) Each's network's performance is evaluated on three labeling tasks.

of frames in between periods of cell inactivity, without the need for image segmentation. Both real and simulated retinal wave datasets are normalized to have a global mean and variance of 0 and 1, respectively.

To isolate the effects of the spatial and temporal characteristics of retinal waves, we pre-train networks on three additional types of datasets created by modifying the original movies, described below and depicted in Fig. 3.11A.

Spatially shuffled waves: Pixels of each frame are randomly permuted. Spatially shuffled waves contain information about how the overall distribution of RGC activities changes over time, but lack the continuously varying spatial structure present in the original movies. As such, pre-training on spatially shuffled waves controls for how much task information can be inferred only through temporally local changes in the population statistics of RGC activity.

Temporally shuffled waves: The sequence of frames is randomly permuted. This condition controls for the amount of task-relevant, temporally non-local information in retinal waves.

Spatiotemporally shuffled waves: Both the pixels of each frame and the sequence of frames are randomly permuted. If correlations between temporally distant frames are relevant for a given task, networks pre-trained on temporally shuffled waves should perform better than those trained on spatiotemporally shuffled waves.

We compare all pre-training conditions to a He random initialized control network that has not been pre-trained, for a total of nine conditions (four sets of pre-training data — unshuffled, spatially shuffled, temporally shuffled, and spatiotemporally shuffled — for both real retinal waves and simulated retinal waves, plus one randomly initialized control netowork).

Model architecture: For all pre-training, we use a ResNet-18 network [90] (without the fully connected classification layer) followed by a projector layer. The projector consists of three linear layers with 8192 output units. The first two linear layers in the projector are each followed by a batch normalization layer and ReLU activations. The ResNet-18 backbone without the classification layer is sometimes referred to in self-supervised learning as an "encoder", and the outputs of the projector layer are referred to as "embeddings" [91]. The SimCLR loss is computed on the embeddings during pre-training, and during task training, the projector is swapped out with a 512×10 linear readout layer. This procedure of swapping out the projector has been shown empirically to be beneficial in transfer learning, where there is a misalignment between the pre-training and training tasks [92].

Hyperparameters: Networks are pre-trained with a projector layer [91] of dimensions $8192 \times 8192 \times 8192$ for 100 epochs with a learning rate of 0.0001 and Adam optimization based on a grid hyperparameter search. Because wave events occur for varying lengths of time, batches are formed by randomly sampling whole wave events from the movie until the total number of sampled frames exceeds a threshold value of 3000. Positive examples

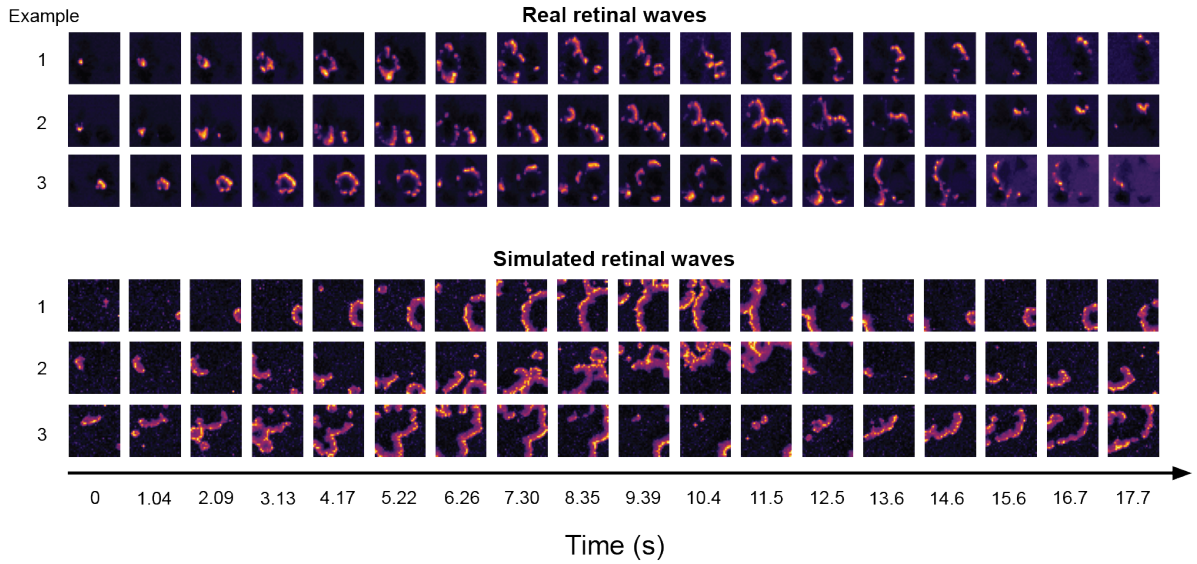


Figure 3.12: **Qualitative comparison of representative examples from real and simulated retinal wave movies.** While we do not perform a direct quantitative comparison between the real and simulated retinal waves in this work, we present 3 representative examples from each dataset taken over a time period of about 18 sec. For each example, every 12th frame is presented in order to visualize wave activity over longer a period of time. A key difference between the two datasets is that in the real retinal wave movies, the waves must terminate when they reach a boundary of the imaged retina (Fig. 3.4), but in the simulated retinal wave movies, the "retina" is a uniform surface that extends beyond the field of view. For this reason, in the simulated movies, the waves may continue past the frame. We partially adjust for this difference by setting the area parameter of the simulated retinal wave model as the average area of the calcium imaged retinas, though this adjustment does not account for any variations in wave characteristics induced by the retinal border.

are defined as consecutive frames within the same wave event, and negative examples are defined as all frames outside of that wave event.

Task-training on CIFAR images

To test the effects of pre-training on task performance, we add a linear readout layer to the pre-trained weights and train linear readout layer weights on labeled images while leaving the pre-trained hidden layer weights fixed (Fig. 3.11C). **This phase is meant to simulate a test of the functionality gained from retinal wave activity at the onset of visual experience.** We evaluate network performance on three labeling tasks to examine and bound the scope of functions that be learned from pre-training on retinal waves.

Classification task: The first task is standard image classification on CIFAR-10. The test

Figure 3.13: **Base images and labels for spatial translation and color change tasks.**

of function this task is meant to simulate is linking semantic and visual information.

Spatial translation task: For the second task, we train networks to classify spatially translated images drawn from CIFAR-100. The test of function this task is meant to simulate is recognizing an object in the face of affine spatial transformations. To generate the task data, we first choose 10 of 100 classes at random and draw a random image from each class, which we denote as a "base" image (Fig. 3.13). An image in the task dataset is then generated as a random affine transformation (up to 16 pixels in the x and y directions) of one of the 10 base images. Using this procedure, each base image is used to generate 5000 training images and 1000 test images, for a total of 50,000 training images and 10,000 test images. The networks are trained to classify a given training image with the label of its original base image.

Color change task: For the third task, we train networks to classify recolorations of the same 10 base images used in the spatial translation task. The test of function this task is meant to simulate is recognizing an object that has been recolored. The task data is generated by the same procedure, as the spatial translation task, only instead of random affine transformations, we apply random color transformations to the base image that range from 50 to 100% changes in saturation, brightness, contrast, and hue. The networks are trained to classify a given training image with the label of its original base image.

Model architecture: For task training, the projector from the pre-trained ResNet-18 is swapped out with a 512×10 linear readout layer. This procedure of swapping out the projector has been shown empirically to be beneficial in transfer learning, where there is a misalignment between the pre-training and training tasks [92].

Hyperparameters: In task training, the projector dimension used in pre-training is re-

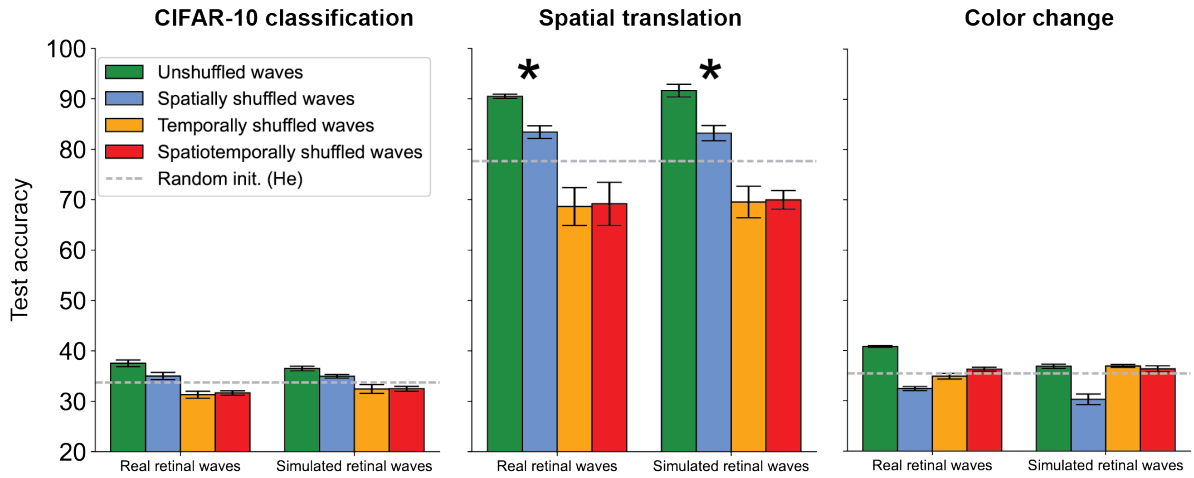


Figure 3.14: **Test accuracy for pre-trained networks in three labeling tasks.** Asterisks indicate that the performance increase from pre-training on both real and simulated retinal waves (relative to random baseline performance) is highest for the spatial translation task. Pre-training yields only a slight performance boost for the standard CIFAR-10 classification and color change tasks.

moved and replaced with a 512×10 linear readout layer [91]. The readout layer is trained for 100 epochs, batches of size 100, and learning rate of 0.0001 on 50,000 labeled training images. The performance is evaluated on 10,000 labeled test images.

Pre-training on retinal waves improves task performance

Our main result is that self-supervised pre-training of networks on movies of retinal waves improves object separability for labeled images. In particular, we find that pre-training on the original, unshuffled wave movies yields the highest performance increase in the spatial translation task (Fig. 3.14, middle). This suggests that retinal waves contain information that supports learning object invariance to spatial translation. Pre-training on spatially shuffled waves yields a moderate improvement above random initialization in this task, suggesting that learning temporally local changes in the overall distribution of activities is also relevant for this function. Destroying the temporal structure of the waves, however, yields performance below random initialization, as shown in the temporally and spatiotemporally shuffled pre-training conditions. This suggests that temporally local, rather than global correlations in retinal waves are most relevant for learning spatial invariance. This is consistent with the previous finding that little information is gained by considering RGC activity bursts more than 3 sec (around 35 frames) apart [87, 88]. These networks perhaps even learn non-local features that actually hinder task learning, as suggested by

their below-random-network performance. We further explore this idea in Sec. 3.4.

Classification is a far more complex task than spatial translation as it requires mapping visual information onto higher level semantic structures, information not present retinal waves. Accordingly, performance for this task is significantly lower for pre-trained networks overall than for spatial translation. However, networks trained on unshuffled waves still perform slightly better than the others (Fig. 3.14). A similar trend emerges for the color change task, for which we also did not expect pre-training to yield any advantage. A potential reason for the performance increases in both cases is the persistence of similar features across examples in the same class. Visual patterns like edges and curves are features that retinal waves may train the visual system to recognize [59]. We further explore reasons for these small performance boosts in Section 3.4.

While accuracy provides a proxy for the task-specific relevance of retinal waves, it does not give insight into how retinal waves influence learned feature representations. In the next section, we address this question by examining the geometry of task object manifolds across pre-training conditions.

Pre-training on retinal waves increases capacity for manifolds defined by invariance to spatial translation

To examine how pre-training with retinal waves affects the geometry, and in turn the separability, of neural object manifolds for each task, we extract the network activations at each ReLU layer for $P = 50$ manifolds consisting of $M = 20$ examples. For **standard classification**, each manifold corresponds to an image class in CIFAR-100. Examples for each manifold are drawn from the given class based on the ranked 20-highest softmax probability scores output by a well-trained classifier. For both **spatial translation** and **color change**, each manifold corresponds to one random base image drawn from CIFAR-100. Examples for each spatial translation manifold are generated by applying random affine shifts up to 3 pixels in both directions to the base image. Examples for each color change manifold are generated by applying random $50 - 150\%$ changes in saturation, brightness, hue, and contrast to the base image. For all theoretical manifold quantities (α_c, D_M, R_M, correlation, PR, EV) the outputs of the intermediate ReLU activations in the encoder (for a total of 9 activation layers [90]) are extracted to analyze the internal representations of the task or wave manifolds at each layer. Due to the high computational cost, we only calculate the simulation capacity at the last ReLU in the encoder.

Previous work shows that DCNNs trained to classify images increase the object manifold capacity from the input to output layers [82]. We only observe this behavior for the spatial translation manifold. Consistent with the accuracy results, networks trained on unshuffled waves and spatially shuffled waves yield increases in capacity relative to randomly initialized networks, while networks trained on temporally and spatiotemporally shuffled waves do not substantially change the capacity between the input and output layers (Fig. 3.15).

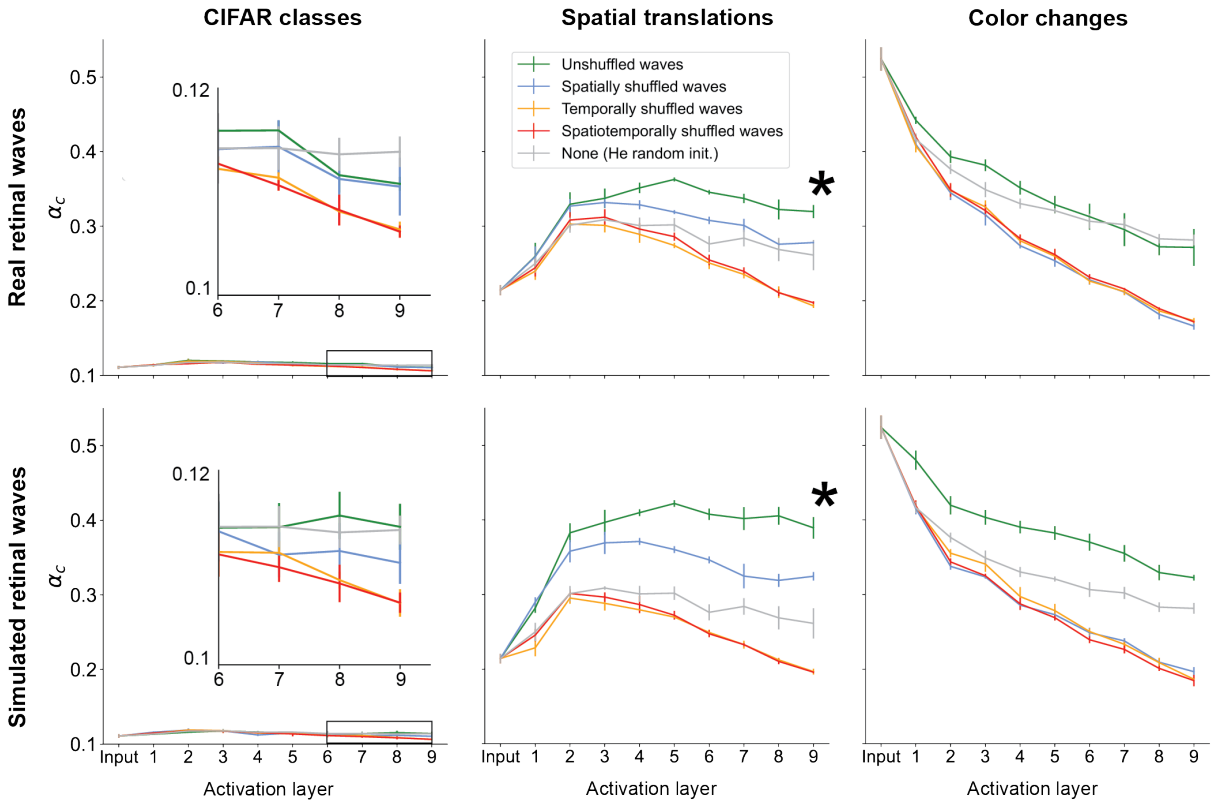


Figure 3.15: **Changes in classification capacity over network layers.** Asterisks indicate that the capacity of spatial translation manifolds increases the most along the hierarchy of the network pre-trained on unshuffled retinal waves. Insets (top left and bottom left plots) show that there is little difference in capacity across pre-trained and random networks for the CIFAR class manifolds. Unexpectedly, pre-training on simulated, but not real retinal waves yields a slight increase in capacity above random for the color change manifolds.

In all networks, the capacity of the CIFARs class manifold (see inset, Fig. 3.15) remains nearly constant around the theoretical lower bound of 0.1 (Sec. 3.2). All networks also yield a decrease in capacity for the color change manifold at each successive layer (Fig. 3.15). Although the network trained on simulated unshuffled waves appears to have a relatively high capacity for the color change manifold, this particular value actually overestimates the ground truth simulation capacity, which we show in Fig. 3.16. To determine why this overestimation is more pronounced for the network trained on simulated waves, a useful followup would be to more closely inspect the statistics of anchor points used in calculating the mean field theoretic capacity 3.2. The correction provided by calculating the simulation capacity, however, demonstrates that there is no substantial improvement in the color change task with regards to representation after pre-training on retinal waves.

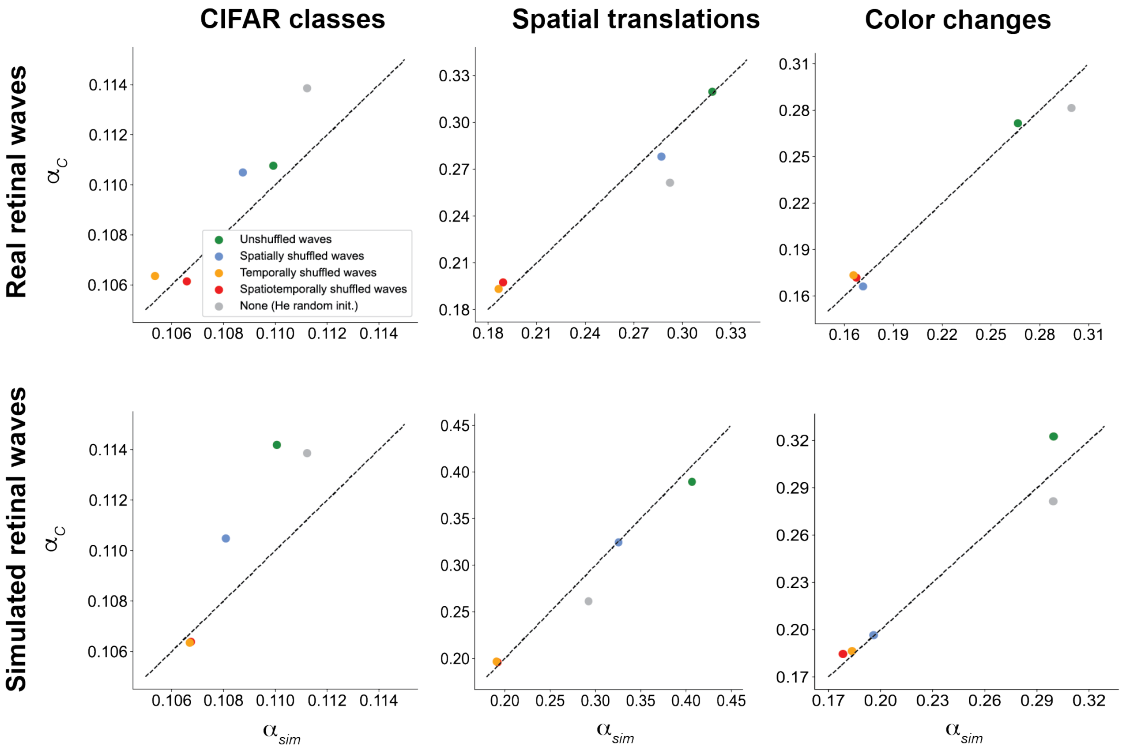


Figure 3.16: **Correspondence between theoretical and simulation capacity.** Each point represents mean over three random network initializations at the last activation layer in the encoder. Dotted gray line denotes exact match between α_c and α_{sim}. We note a high degree of correspondence between theoretical and simulation capacity, with the exception of the CIFAR and color change manifolds for networks pre-trained on simulated retinal waves (second row, first and third columns).

With exception of this overestimation, we generally observe a high degree of correspondence between theoretical and simulation capacity.

Pre-training on retinal waves decreases radius and dimension for manifolds defined by invariance to spatial translation

As expected, the spatial translation manifolds in networks pre-trained on unshuffled waves have lower dimension and radius compared to those in the other networks, while networks pre-trained on spatially shuffled waves only appear to decrease the radius (Fig. 3.17). These results suggest that pre-training on retinal waves has a direct influence on the geometry and separability of neural object manifolds for tasks that involve learning spatial invariance. Meanwhile, the dimensions and radii of the CIFAR class and color change manifolds do not show any consistent ordering that points to a clear advantage

of pre-training on retinal waves relative to the random baseline (Fig. 3.17). These results are consistent with the poor accuracy in the classification and color change tasks across all networks.

However, if pre-training does not substantially affect these object manifolds, what accounts for the slight boost in performance on these tasks for the networks pre-trained on unshuffled waves (Fig. 3.14)? To address this question, we explore two factors external to the geometry of individual manifolds, namely the inter-manifold correlation and the effective dimensionality of the feature space.

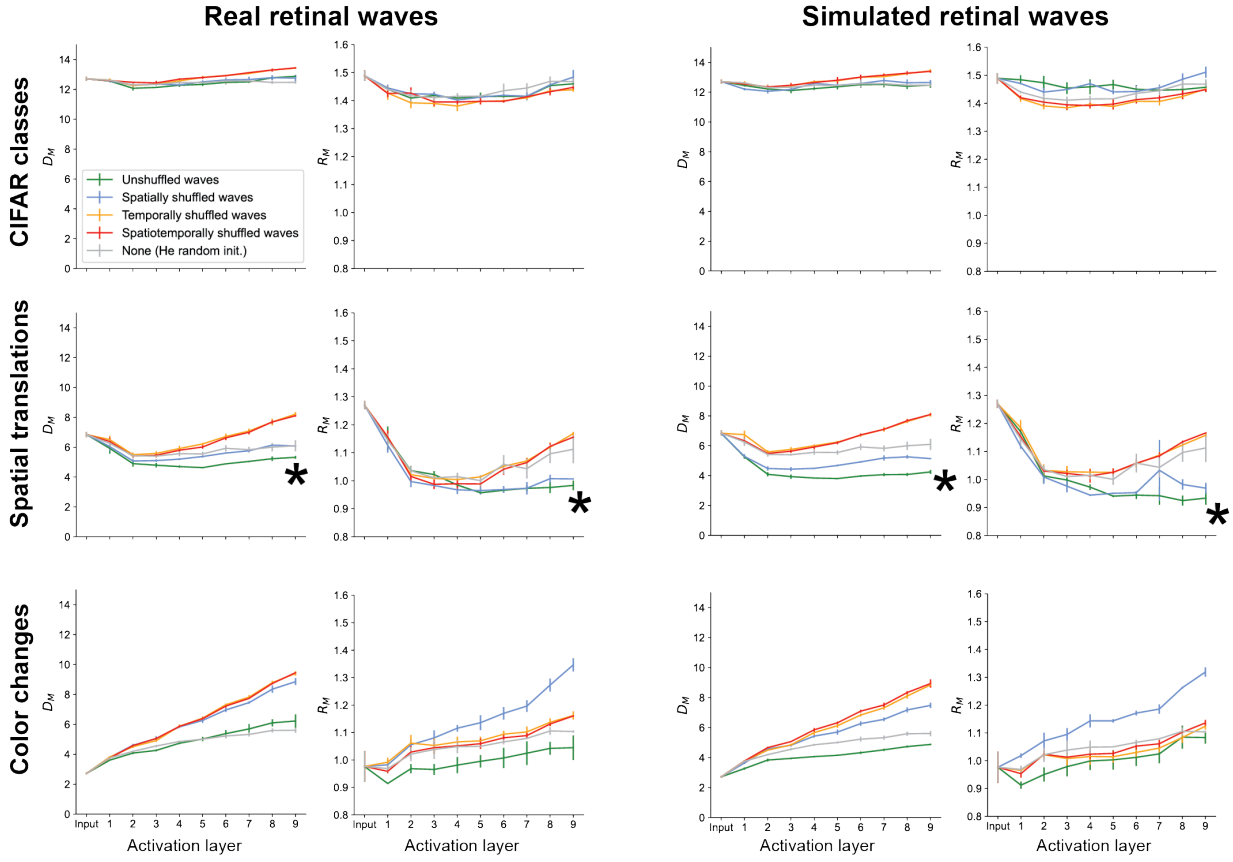


Figure 3.17: **Changes in manifold geometry over network layers.** Asterisks indicate that networks pre-trained on unshuffled retinal waves most effectively compress spatial translation manifolds, as indicated by the decreases in both dimension and radius in deeper layers.

Pre-training on retinal waves decreases inter-manifold correlations

A high degree of correlation between manifold centers may lead to clustering of object manifolds in feature space, making them more difficult to separate and decreasing the

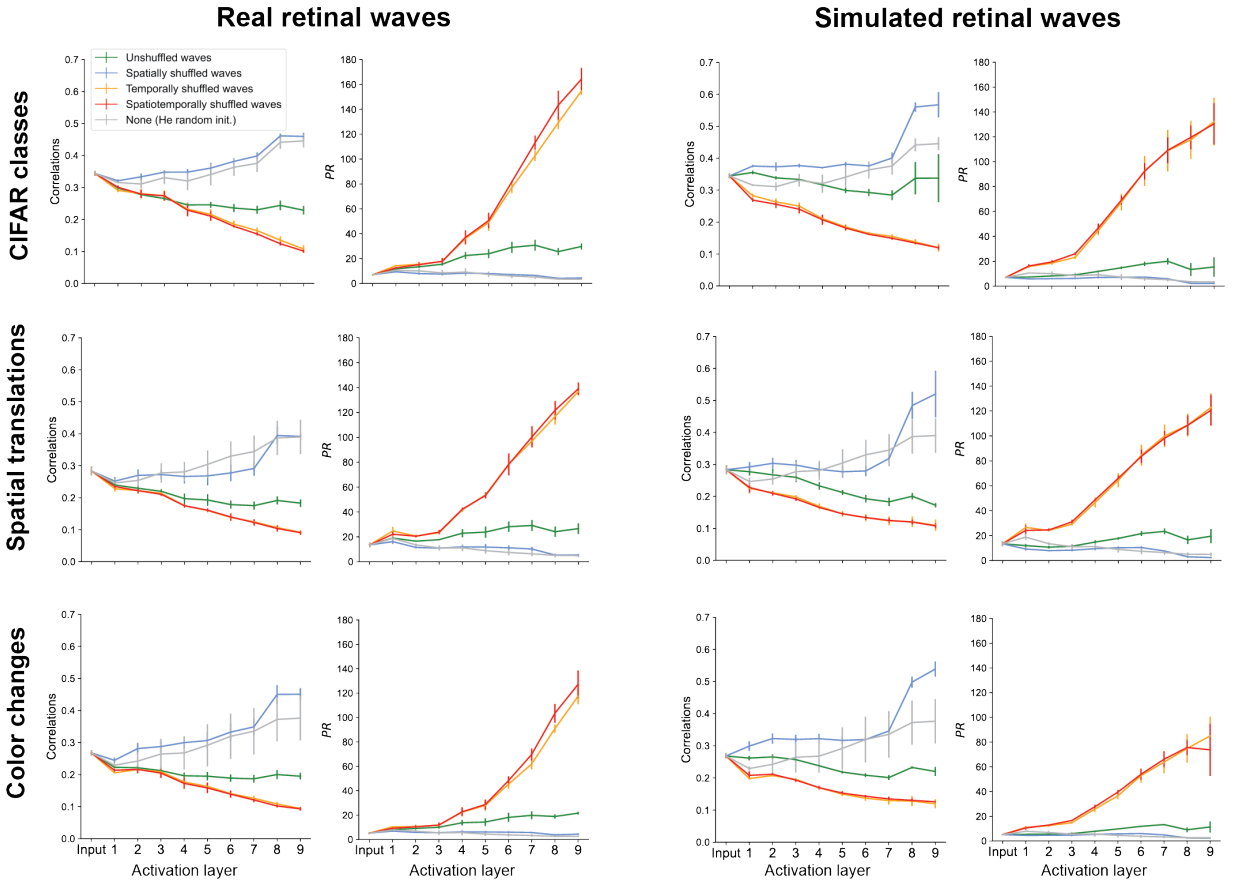


Figure 3.18: **Changes in inter-manifold correlation and participation ratio along network layers**. Only the network pre-trained on unshuffled waves consistently reduces correlation and avoids vanishing/exploding dimensionality.

effective capacity. Previous work demonstrates that training DCNNs leads to decorrelation of the manifold centers [82]. Here, we measure the pairwise correlation coefficient between manifold centers at each network layer and find that networks pre-trained on unshuffled retinal waves decrease center correlations relative to randomly initialized networks and networks pre-trained on spatially shuffled waves for all three tasks (Fig. 3.18). Unshuffled pre-training also leads to a generally consistent decrease in correlation along at each successive network layer. Interestingly, temporally and spatiotemporally shuffled pre-training also produce networks that exhibit this behavior, in addition to having lower correlations than in the unshuffled case. However, based on their poor task performance and low capacities of their feature representations, it is likely this is simply due to the explosion in dimensionality of their respective feature spaces, which we discuss next.

Pre-training on retinal waves maintains effective dimensionality of the data

Ideally, a well-trained classifier will extract the features that correspond to the highest sources of variance in the data, while separating out low-variance features that do not correspond to meaningful distinctions between samples. Participation ratio (PR) varies from 1 to N and measures how data variance is spread out across the feature dimensions: if $PR = 1$, the variance is concentrated entirely in one feature; if $PR = N$, the variance is spread out evenly across all features [93]. In general, a good classifier will maintain a $PR > 1$ in the feature dimensions so as to preserve the latent dimensionality in the data (which is in the vast majority of applications is higher than 1), while also keeping $PR < N$ so as to extract only the meaningful (high variance) features as the basis for classification. The layer-wise participation ratio suggests that networks pre-trained on unshuffled waves maintain this happy medium in all three tasks (Fig. 3.18). Networks pre-trained on spatially shuffled waves decrease participation ratio to near the lower bound, consistent with the idea that they broadly capture population-level statistics, but fail to learn many spatially local features that likely lie along other dimensions. The large increase in PR observed in networks trained on temporally and spatiotemporally shuffled waves suggests that they do in fact learn features that are not relevant for the task dataset, as proposed in Section 3.4. These extraneous features would account for the increase in PR above the values observed in other networks. Notably, correlation and PR are inversely related, suggesting that high effective dimensionality is a factor in separation of manifold centers.

The trends observed in PR are consistent with the trends in layer-wise explained variance, which measures how many feature dimensions account for a given percentage of variance in the data (Fig. 3.19). We measure explained variance as the number of dimensions in feature space that account for 90% of the variance in the examples considered for manifold analysis. The trend in EV in all tasks reflects that observed in center correlation and PR (Fig. 3.18), whereby the networks pre-trained on unshuffled waves maintain higher feature dimensionality and lower center correlation than the random networks, without producing a dimensionality explosion like the networks pre-trained on temporally shuffled waves.

The network learns to efficiently represent wave events

To analyze how the network learns to represent the retinal waves themselves over pre-training, we examine how the geometry of the retinal wave manifolds changes throughout the network layers. A wave manifold as described in Fig. 3.20 is defined from a set of 50 frames from a randomly chosen single wave event in the original, unshuffled wave movie. We consider 50 such manifolds for all such metrics computed in Fig. 3.20 across the five pre-training conditions. Explained variance is the number of dimensions in feature space that account for 90% of the variance in the frames considered for manifold analysis.

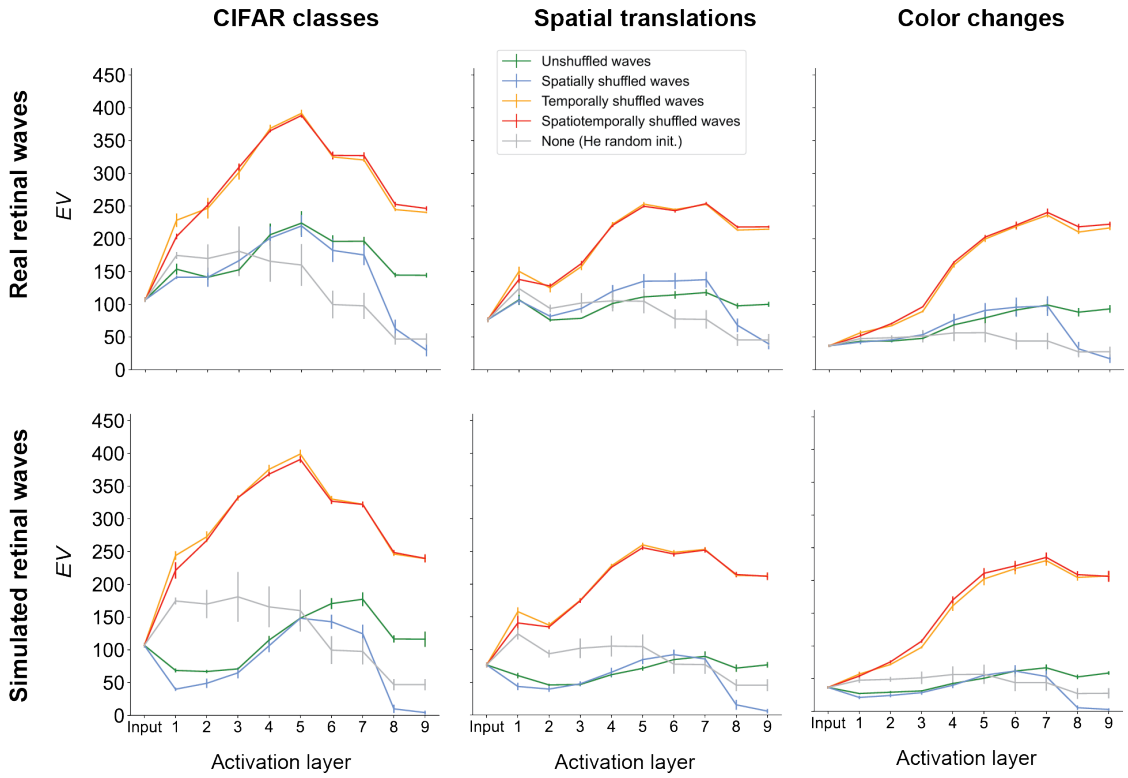


Figure 3.19: **Changes in inter-manifold correlation and participation ratio along network layers**. Only the network pre-trained on unshuffled waves consistently reduces correlation and avoids vanishing/exploding dimensionality.

As expected, networks pre-trained on unshuffled waves yield the highest capacity amongst all pre-training conditions for the wave manifolds. Interestingly, the manifolds for real retinal waves appear to have higher capacity at all layers compared to those for simulated waves. This may be due to the smaller size of the real retinal waves dataset, which could lead to less variability across frames than in the simulated dataset. This explanation is consistent with the fact that the PR and EV for simulated waves is higher than for real waves. The trend in correlation, PR, and EV for both real and simulated wave manifolds reflects that observed in the task manifolds (Fig. 6), whereby the networks pre-trained on unshuffled waves maintain higher feature dimensionality and lower center correlation than the random networks, without producing a dimensionality explosion like the networks pre-trained on temporally shuffled waves.

We also note that for retinal wave manifolds, calculation of α_{sim} is numerically unstable in the last activation layer for networks not trained on unshuffled waves (fourth column). This may occur when the manifold capacity is low relative to the feature dimension N, resulting in poor separability. For this reason, we instead report the values of α_c and α_{sim} for the wave manifolds in the projector layer.

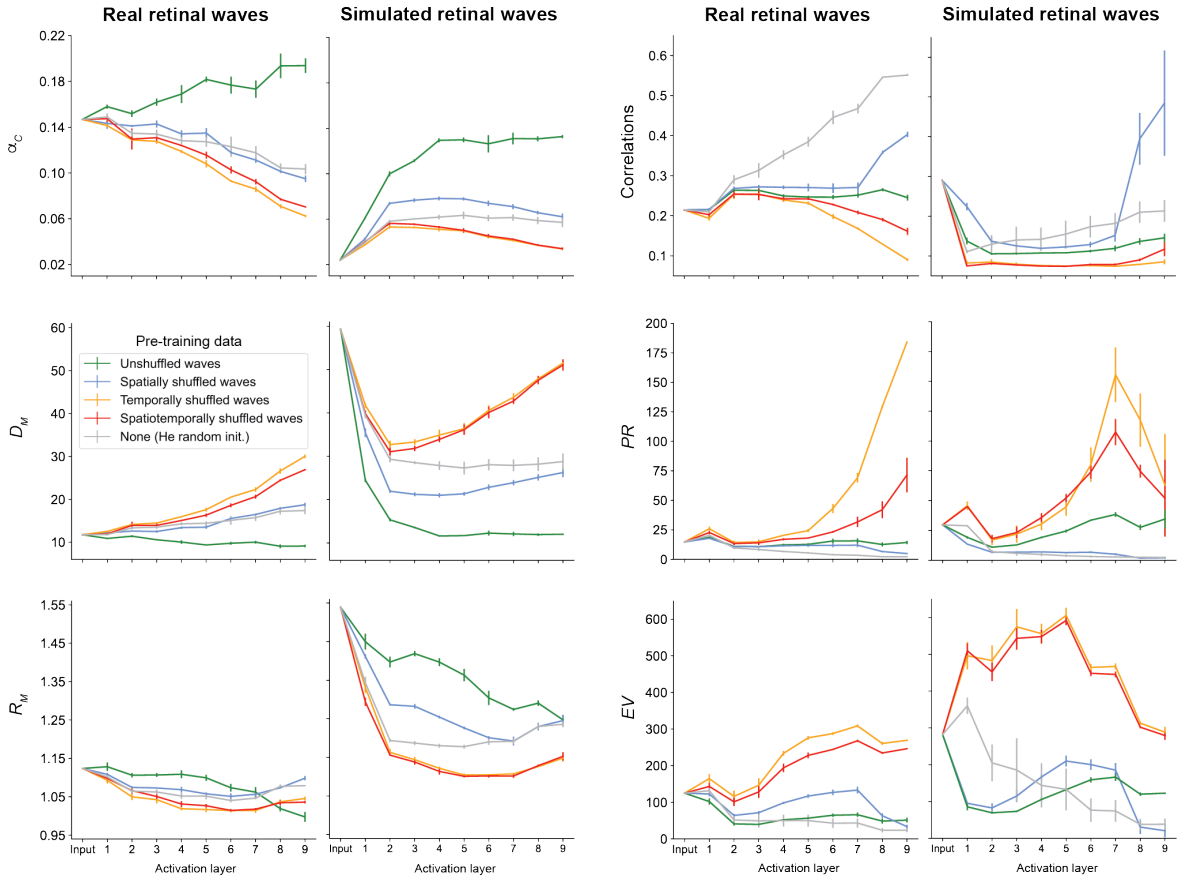


Figure 3.20: **Changes in (unshuffled) wave manifolds over network layers.**

3.5 Implications for Experience-Independent Development of the Visual System

To our knowledge, this is the first computational work that directly explores how real retinal waves can influence neural object representations, demonstrating a bioplausible means of learning spatial invariance without training on large datasets of labeled images. While DCNNs trained on labeled images achieve state-of-the-art performance and even predict neural responses [80, 7], these models are unlikely to explain how biological vision develops. Unsupervised and self-supervised learning mechanisms have therefore been proposed as biologically plausible means of learning object recognition [68]. However, standard implementations of these algorithms still require natural images or videos as training inputs, which effectively simulate a visual experience. Though visual experience certainly shapes cortical functional development [94, 95, 26, 96], models that wholly rely on image data do not account for the functionality, connectivity, and feature selectivity al-

ready observed in animals prior to the onset of vision [97, 98, 99, 100, 101]. Consistent with our results, previous work has demonstrated that self-supervised learning on structured noise can improve classification accuracy on unseen images [75, 74, 102]. Additionally, simulated retinal waves have been shown to yield V1-like receptive fields when used as inputs for sparse coding algorithms [59, 19, 103] and slow feature analysis [58].

We demonstrate that pre-training on retinal waves has two primary effects on learned representations that can account for increases in task performance. The first is an increase in the separability of individual object manifolds. This effect is pronounced in the spatial translation task, suggesting that the spatiotemporal characteristics of retinal waves train networks to learn spatial translation invariance. To show this, we analyze the geometry of the neural object manifolds defined by affine transformations of a single object (image) and find they are more linearly separable when represented in networks pre-trained on unshuffled retinal waves (Figs. 3.15, 3.17). Both the spatial and temporal characteristics of retinal waves are necessary for learning this task, as pre-training on spatially and/or temporally shuffled retinal waves leads to poor separability of spatial translation manifolds. Pre-training does not have a significant effect on the separability of the manifolds defined by CIFAR image classes or color changes of a single object (Figs. 3.15, 3.17), suggesting a qualitative bound on the scope of tasks for which retinal waves are useful training signals.

We also observe that pre-training on retinal waves reduces center correlations between neural object manifolds and increases the effective dimensionality of the feature space (Figs. 3.18). Both effects are directly correlated with linear separability and appear to be independent of the effect on individual manifold separability, as they are observed in all three tasks.

Together, these two effects of pre-training on retinal waves correspond to distinct *local* and *global* mechanisms of transforming object representations, both of which are important for separability. At the local level, pre-training increases the compressibility of individual neural object manifolds, as shown in the increase in capacity and the concurrent decreases in dimension and radius. At the global level, pre-training places neural object manifolds in higher dimensional feature space, as shown by the increase in participation ratio and concurrent decrease in center correlation. These two regimes point to distinct ways in which retinal waves may influence emerging sensory representations.

We do not observe a significant difference between pre-training on real versus simulated retinal waves from the model. The advantage of the model is that we can generate an arbitrarily large set of pre-training data, at the risk of introducing free parameters that may lead to deviations from real data. Though we do not perform a direct comparison between the simulated and real data in this work, no clear difference emerges between these two datasets in terms of model performance or the geometry of the object representations. This suggests that for the tasks considered, the common features of these datasets — such as spatiotemporal continuity between frames — are the primary drivers of the observed effects. In future work, the model may be a useful tool for examining the effect of changing the waves' spatiotemporal characteristics on representation learning.

We note that our findings are subject to our choice of network architecture (ResNet-

18), learning algorithm (SimCLR), and dataset (postnatal mouse retinal waves). Retinal waves occur during multiple stages of development [104] and drive formation of visual circuitry in numerous ways [70]. Retinal waves are also not the only form of spontaneous activity during development [105]. Along this line of work, future studies may consider the role of cortical feedback [106], introduce bioplausible, synaptically local learning rules [107], or investigate the role of spontaneous activity in other modalities like temporal prediction [108]. Additionally, laboratory experiments that test object recognition in mice [109] performed at the onset of vision could verify our model predictions and provide richer insight into the capacity of neural object manifolds during this early developmental period.

Bibliography

[1] Nobuhiko Watari and Ronald G. Larson. "The Hydrodynamics of a Run-and-Tumble Bacterium Propelled by Polymorphic Helical Flagella". In: *Biophysical Journal* 98.1 (Jan. 2010), pp. 12–17. DOI: 10.1016/j.bpj.2009.09.044. URL: https://doi.org/10.1016/j.bpj.2009.09.044.

[2] Peter Godfrey-Smith. *Other Minds*. New York, NY: Farrar, Straus & Giroux, Dec. 2016.

[3] Charles G. Frye et al. "Critical Point-Finding Methods Reveal Gradient-Flat Regions of Deep Network Losses". In: *Neural Computation* 33.6 (May 2021), pp. 1469–1497. DOI: 10.1162/neco_a_01388. URL: https://doi.org/10.1162/neco_a_01388.

[4] Eric Jonas and Konrad Paul Kording. "Could a Neuroscientist Understand a Microprocessor?" In: *PLOS Computational Biology* 13.1 (Jan. 2017). Ed. by Jörn Diedrichsen, e1005268. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005268. URL: http://dx.doi.org/10.1371/journal.pcbi.1005268.

[5] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, 2010. ISBN: 9780262289610. DOI: 10.7551/mitpress/9780262514620.001.0001. URL: http://dx.doi.org/10.7551/mitpress/9780262514620.001.0001.

[6] Mingyu Ding et al. *DaViT: Dual Attention Vision Transformers*. 2022. DOI: 10.48550/ARXIV.2204.03645. URL: https://arxiv.org/abs/2204.03645.

[7] Martin Schrimpf et al. "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" In: (Sept. 2018). DOI: 10.1101/407007. URL: https://doi.org/10.1101/407007.

[8] Martin Schrimpf et al. "Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence". In: *Neuron* (2020). URL: https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

[9] Eero P Simoncelli and Bruno A Olshausen. "Natural Image Statistics and Neural Representation". In: *Annual Review of Neuroscience* 24.1 (Mar. 2001), pp. 1193–1216. ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev.neuro.24.1.1193. (Visited on 03/11/2021).

[10] Fred Attneave. "Some informational aspects of visual perception." In: *Psychological review* 61 3 (1954), pp. 183–93. URL: https://api.semanticscholar.org/CorpusID:8453552.

[11] H. B. Barlow. "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Communication*. The MIT Press, Sept. 2012, pp. 216–234. DOI: 10.7551/mitpress/9780262518420.003.0013. URL: http://dx.doi.org/10.7551/mitpress/9780262518420.003.0013.

[12] William E. Vinje and Jack L. Gallant. "Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision". In: *Science* 287.5456 (Feb. 2000), pp. 1273–1276. ISSN: 1095-9203. DOI: 10.1126/science.287.5456.1273. URL: http://dx.doi.org/10.1126/science.287.5456.1273.

[13] Bruno A. Olshausen and David J. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (June 1996). Number: 6583 Publisher: Nature Publishing Group, pp. 607–609. ISSN: 1476-4687. DOI: 10.1038/381607a0. (Visited on 03/11/2021).

[14] J. Sebastian Espinosa and Michael P. Stryker. "Development and Plasticity of the Primary Visual Cortex". en. In: *Neuron* 75.2 (July 2012), pp. 230–249. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.06.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627312005697 (visited on 09/13/2021).

[15] Jianhua Cang et al. "Development of precise maps in visual cortex requires patterned spontaneous activity in the retina". eng. In: *Neuron* 48.5 (Dec. 2005), pp. 797–809. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2005.09.015.

[16] Anand R. Chandrasekaran et al. "Evidence for an instructive role of retinal activity in retinotopic map refinement in the superior colliculus of the mouse". eng. In: *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 25.29 (July 2005), pp. 6929–6938. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.1470-05.2005.

[17] Andrew D. Huberman, Colenso M. Speer, and Barbara Chapman. "Spontaneous Retinal Activity Mediates Development of Ocular Dominance Columns and Binocular Receptive Fields in V1". English. In: *Neuron* 52.2 (Oct. 2006). Publisher: Elsevier, pp. 247–254. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2006.07.028. URL: https://www.cell.com/neuron/abstract/S0896-6273(06)00625-8 (visited on 09/14/2022).

[18] Jeffrey Markowitz, Yongqiang Cao, and Stephen Grossberg. "From Retinal Waves to Activity-Dependent Retinogeniculate Map Development". en. In: *PLOS ONE* 7.2 (Feb. 2012). Publisher: Public Library of Science, e31553. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0031553. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031553 (visited on 09/16/2022).

[19] Jonathan J. Hunt, Michael Ibbotson, and Geoffrey J. Goodhill. "Sparse Coding on the Spot: Spontaneous Retinal Waves Suffice for Orientation Selectivity". In: *Neural Computation* 24.9 (Sept. 2012), pp. 2422–2433. ISSN: 0899-7667. DOI: 10.1162/NECO_a_00333. URL: https://doi.org/10.1162/NECO_a_00333 (visited on 04/25/2022).

[20] Michael C. Crair, Deda C. Gillespie, and Michael P. Stryker. "The Role of Visual Experience in the Development of Columns in Cat Visual Cortex". In: *Science* 279.5350 (Jan. 1998), pp. 566–570. DOI: 10.1126/science.279.5350.566. URL: https://doi.org/10.1126/science.279.5350.566.

[21] D H Wolpert and W G Macready. "No free lunch theorems for optimization". In: *IEEE Trans. Evol. Comput.* 1.1 (Apr. 1997), pp. 67–82.

[22] David H Hubel and Torsten N Wiesel. "Receptive fields of single neurones in the cat's striate cortex". In: *The Journal of physiology* 148.3 (1959), pp. 574–591.

[23] J. P. Jones and L. A. Palmer. "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex". In: *Journal of Neurophysiology* 58.6 (Dec. 1987), pp. 1233–1258. ISSN: 0022-3077, 1522-1598. DOI: 10.1152/jn.1987.58.6.1233. (Visited on 04/20/2021).

[24] Dario L. Ringach. "Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex". In: *Journal of Neurophysiology* 88.1 (July 2002), pp. 455–463. ISSN: 0022-3077. DOI: 10.1152/jn.2002.88.1.455.

[25] Dario L Ringach. "Mapping receptive fields in primary visual cortex". In: *The Journal of Physiology* 558.Pt 3 (Aug. 2004), pp. 717–728. ISSN: 0022-3751. DOI: 10.1113/jphysiol.2004.065771. (Visited on 08/10/2021).

[26] Nina N. Kowalewski et al. "Development of Natural Scene Representation in Primary Visual Cortex Requires Early Postnatal Experience". In: *Current Biology* 31.2 (Jan. 2021), 369–380.e5. DOI: 10.1016/j.cub.2020.10.046. URL: https://doi.org/10.1016/j.cub.2020.10.046.

[27] Daniel Graham and David Field. "Sparse Coding in the Neocortex". In: *Evolution of Nervous Systems* 3 (Dec. 2007). DOI: 10.1016/B0-12-370878-8/00064-1.

[28] Martin Rehn and Friedrich T. Sommer. "A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields". In: *Journal of Computational Neuroscience* 22.2 (Feb. 2007), pp. 135–146. ISSN: 0929-5313, 1573-6873. DOI: 10.1007/s10827-006-0003-9. (Visited on 04/08/2021).

[29] Joel Zylberberg, Jason Timothy Murphy, and Michael Robert DeWeese. "A Sparse Coding Model with Synaptically Local Plasticity and Spiking Neurons Can Account for the Diverse Shapes of V1 Simple Cell Receptive Fields". In: *PLoS Computational Biology* 7.10 (Oct. 2011). Ed. by Olaf Sporns, e1002250. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002250. (Visited on 03/11/2021).

[30] Yuzo M. Chino et al. "Postnatal Development of Binocular Disparity Sensitivity in Neurons of the Primate Visual Cortex". In: *Journal of Neuroscience* 17.1 (1997), pp. 296–307. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.17-01-00296.1997. eprint: https://www.jneurosci.org/content/17/1/296.full.pdf.

[31] Nana Nishio et al. "The role of early visual experience in the development of spatial-frequency preference in the primary visual cortex". In: *The Journal of Physiology* 599.17 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1113/JP281463, pp. 4131–4152. ISSN: 1469-7793. DOI: 10.1113/JP281463. (Visited on 09/13/2021).

[32] B A Olshausen and D J Field. "Natural image statistics and efficient coding". In: *Network* 7.2 (Jan. 1996), pp. 333–339.

[33] Bruno A Olshausen. "Linear Hebbian learning and PCA". In: (2012), p. 8.

[34] Michael Fang et al. "Learning and Inference in Sparse Coding Models with Langevin Dynamics". In: (Apr. 2022).

[35] Bruno A. Olshausen. "Highly overcomplete sparse coding". In: ed. by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder. Mar. 2013. DOI: 10.1117/12.2013504. (Visited on 03/24/2021).

[36] Ronen Basri et al. "The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies". In: *arXiv* 1906.00425 (2019). arXiv: 1906.00425. URL: http://arxiv.org/abs/1906.00425.

[37] Nasim Rahaman et al. "On the Spectral Bias of Neural Networks". In: *arXiv* (2019). arXiv: 1806.08734 [stat.ML].

[38] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: (2018). DOI: 10.48550/ARXIV.1806.07572. URL: https://arxiv.org/abs/1806.07572.

[39] Yuan Cao et al. *Towards Understanding the Spectral Bias of Deep Learning*. 2019. DOI: 10.48550/ARXIV.1912.01198. URL: https://arxiv.org/abs/1912.01198.

[40] Gerrit Ecke. *fit2dGabor*. Version 1.0.1.0. Aug. 1, 2017. URL: https://www.mathworks.com/matlabcentral/fileexchange/60700-fit2dgabor-data-options?s_tid=srchtitle.

[41] Bruno A. Olshausen and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research* 37.23 (Dec. 1997), pp. 3311–3325. ISSN: 0042-6989. DOI: 10.1016/s0042-6989(97)00169-7. URL: http://dx.doi.org/10.1016/S0042-6989(97)00169-7.

[42] P. Földiák. "Forming sparse representations by local anti-Hebbian learning". In: *Biological Cybernetics* 64.2 (1990), pp. 165–170. ISSN: 0340-1200. DOI: 10.1007/BF02331346.

[43] Eric McVoy Dodds, Jesse Alexander Livezey, and Michael Robert DeWeese. "Spatial whitening in the retina may be necessary for V1 to learn a sparse representation of natural scenes". In: (Sept. 2019). DOI: 10.1101/776799. URL: https://doi.org/10.1101/776799.

[44] Shanshan Qin et al. "Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning". In: *Nature Neuroscience* 26.2 (Jan. 2023), pp. 339–349. ISSN: 1546-1726. DOI: 10.1038/s41593-022-01225-z. URL: http://dx.doi.org/10.1038/s41593-022-01225-z.

[45] Ingrid Daubechies, Michel Defrise, and Christine De Mol. *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*. 2003. DOI: 10.48550/ARXIV.MATH/0307152. URL: https://arxiv.org/abs/math/0307152.

[46] Amir Beck and Marc Teboulle. "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems". In: *SIAM Journal on Imaging Sciences* 2.1 (Jan. 2009), pp. 183–202. ISSN: 1936-4954. DOI: 10.1137/080716542. (Visited on 04/27/2021).

[47] Xaq Pitkow and Markus Meister. "Decorrelation and efficient coding by retinal ganglion cells". In: *Nature Neuroscience* 15.4 (Mar. 2012), pp. 628–635. ISSN: 1546-1726. DOI: 10.1038/nn.3064. URL: http://dx.doi.org/10.1038/nn.3064.

[48] Daniel L Ruderman. "The statistics of natural images". In: *Network: Computation in Neural Systems* 5.4 (1994), pp. 517–548. DOI: 10.1088/0954-898X_5_4_006.

[49] D. J. Tolhurst, Y. Tadmor, and Tang Chao. "Amplitude spectra of natural images". In: *Ophthalmic and Physiological Optics* 12.2 (Dec. 2007), pp. 229–232. DOI: 10.1111/j.1475-1313.1992.tb00296.x. URL: https://doi.org/10.1111/j.1475-1313.1992.tb00296.x.

[50] Joel Zylberberg and Michael Robert DeWeese. "Sparse Coding Models Can Exhibit Decreasing Sparseness while Learning Sparse Codes for Natural Images". In: *PLOS Computational Biology* 9.8 (Aug. 2013). Publisher: Public Library of Science, e1003182. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003182. (Visited on 03/11/2021).

[51] Jonathan J. Hunt, Peter Dayan, and Geoffrey J. Goodhill. "Sparse Coding Can Predict Primary Visual Cortex Receptive Field Changes Induced by Abnormal Visual Input". In: *PLoS Computational Biology* 9.5 (2013). DOI: 10.1371/journal.pcbi.1003005.

[52] Aapo Hyvärinen and Patrik O Hoyer. "A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images". In: *Vision research* 41.18 (2001), pp. 2413–2423.

[53] Paul D King, Joel Zylberberg, and Michael R DeWeese. "Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1". In: *Journal of Neuroscience* 33.13 (2013), pp. 5475–5485.

[54] Rolf Skyberg et al. "Coarse-to-fine processing drives the efficient coding of natural scenes in mouse visual cortex". In: *Cell Reports* 38.13 (2022), p. 110606. DOI: 10.1016/j.celrep.2022.110606.

[55] Chand Parvez Danka Mohammed and Reem Khalil. "Postnatal Development of Visual Cortical Function in the Mammalian Brain". In: *Frontiers in Systems Neuroscience* 14 (2020). ISSN: 1662-5137. DOI: 10.3389/fnsys.2020.00029.

[56] Donald O. Mutti et al. "Ocular Component Development during Infancy and Early Childhood". In: *Optometry and Vision Science* 95.11 (Nov. 2018), pp. 976–985. ISSN: 1040-5488. DOI: 10.1097/opx.0000000000001296. URL: http://dx.doi.org/10.1097/OPX.0000000000001296.

[57] Oliver Braddick and Janette Atkinson. "Development of human visual function". In: *Vision Research* 51.13 (July 2011), pp. 1588–1609. ISSN: 0042-6989. DOI: 10.1016/j.visres.2011.02.018. URL: http://dx.doi.org/10.1016/j.visres.2011.02.018.

[58] Sven Dähne, Niko Wilbert, and Laurenz Wiskott. "Slow Feature Analysis on Retinal Waves Leads to V1 Complex Cells". en. In: *PLoS Computational Biology* 10.5 (May 2014). Ed. by Jeff Beck, e1003564. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003564. URL: https://dx.plos.org/10.1371/journal.pcbi.1003564 (visited on 08/25/2021).

[59] Mark V Albert, Adam Schnabel, and David J Field. "Innate Visual Learning through Spontaneous Activity Patterns". en. In: *PLoS Computational Biology* 4.8 (2008), p. 8.

[60] Simon Thorpe, Denis Fize, and Catherine Marlot. "Speed of processing in the human visual system". In: *Nature* 381.6582 (June 1996), pp. 520–522. DOI: 10.1038/381520a0. URL: https://doi.org/10.1038/381520a0.

[61] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. "How Does the Brain Solve Visual Object Recognition?" en. In: *Neuron* 73.3 (Feb. 2012), pp. 415–434. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2012.01.010. URL: https://www.sciencedirect.com/science/article/pii/S089662731200092X (visited on 09/15/2022).

[62] Rishi Rajalingham, Kailyn Schmidt, and James J. DiCarlo. "Comparison of Object Recognition Behavior in Human and Monkey". In: *The Journal of Neuroscience* 35.35 (Sept. 2015), pp. 12127–12136. DOI: 10.1523/jneurosci.0573-15.2015. URL: https://doi.org/10.1523/jneurosci.0573-15.2015.

[63] Chou P. Hung et al. "Fast Readout of Object Identity from Macaque Inferior Temporal Cortex". In: *Science* 310.5749 (Nov. 2005), pp. 863–866. DOI: 10.1126/science.1117593. URL: https://doi.org/10.1126/science.1117593.

[64] Charles F. Cadieu et al. "Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition". In: *PLoS Computational Biology* 10.12 (Dec. 2014). Ed. by Matthias Bethge, e1003963. DOI: 10.1371/journal.pcbi.1003963. URL: https://doi.org/10.1371/journal.pcbi.1003963.

[65] Elika Bergelson and Daniel Swingley. "At 6–9 months, human infants know the meanings of many common nouns". In: *Proceedings of the National Academy of Sciences* 109.9 (Feb. 2012), pp. 3253–3258. DOI: 10.1073/pnas.1113380109. URL: https://doi.org/10.1073/pnas.1113380109.

[66] Elika Bergelson and Richard N. Aslin. "Nature and origins of the lexicon in 6-mo-olds". In: *Proceedings of the National Academy of Sciences* 114.49 (Nov. 2017), pp. 12916–12921. DOI: 10.1073/pnas.1712966114. URL: https://doi.org/10.1073/pnas.1712966114.

[67] Michael C Frank et al. *Variability and consistency in early language learning: The Wordbank project*. MIT Press, 2021.

[68] Chengxu Zhuang et al. "Unsupervised neural network models of the ventral visual stream". In: *Proceedings of the National Academy of Sciences* 118.3 (2021), e2014196118. DOI: 10.1073/pnas.2014196118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2014196118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2014196118.

[69] M. B. Feller and D. Kerschensteiner. "Chapter 16 - Retinal waves and their role in visual system development". en. In: *Synapse Development and Maturation*. Ed. by John Rubenstein et al. Academic Press, Jan. 2020, pp. 367–382. ISBN: 978-0-12-823672-7. DOI: 10.1016/B978-0-12-823672-7.00016-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128236727000168 (visited on 04/25/2022).

[70] David A. Arroyo and Marla B. Feller. "Spatiotemporal Features of Retinal Waves Instruct the Wiring of the Visual Circuitry". In: *Frontiers in Neural Circuits* 10 (July 2016), p. 54. ISSN: 1662-5110. DOI: 10.3389/fncir.2016.00054. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4960261/ (visited on 04/25/2022).

[71] Ben Jiwon Choi, Yu-Chieh David Chen, and Claude Desplan. "Building a circuit through correlated spontaneous neuronal activity in the developing vertebrate and invertebrate visual systems". en. In: *Genes & Development* 35.9-10 (May 2021), pp. 677–691. ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.348241.121. URL: http://genesdev.cshlp.org/lookup/doi/10.1101/gad.348241.121 (visited on 04/25/2022).

[72] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. "Classification and Geometry of General Perceptual Manifolds". en. In: *Physical Review X* 8.3 (July 2018), p. 031003. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.8.031003. URL: https://link.aps.org/doi/10.1103/PhysRevX.8.031003 (visited on 04/12/2022).

[73] SueYeon Chung and L. F. Abbott. "Neural population geometry: An approach for understanding biological and artificial neural networks". In: *Current Opinion in Neurobiology* 70 (Oct. 2021). arXiv: 2104.07059, pp. 137–144. ISSN: 09594388. DOI:

10.1016/j.conb.2021.10.010. URL: http://arxiv.org/abs/2104.07059 (visited on 04/11/2022).

[74] Guruprasad Raghavan, Cong Lin, and Matt Thomson. *Self-organization of multi-layer spiking neural networks*. 2020. DOI: 10.48550/ARXIV.2006.06902. URL: https://arxiv.org/abs/2006.06902.

[75] Guruprasad Raghavan and Matt Thomson. "Neural networks grown and self-organized by noise". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/1e6e0a04d20f50967c64dac2d639a577-Paper.pdf.

[76] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 1597–1607.

[77] SueYeon Chung. *Statistical Mechanics of Neural Processing of Object Manifolds*. Tech. rep. arXiv:2106.00790. arXiv:2106.00790 [cond-mat, q-bio] type: article. arXiv, June 2021. URL: http://arxiv.org/abs/2106.00790 (visited on 05/13/2022).

[78] Cory Stephenson et al. "Untangling in Invariant Speech Recognition". In: *Neural Information Processing Systems*. 2020.

[79] Ella Bingham and Heikki Mannila. "Random projection in dimensionality reduction: applications to image and text data". In: *KDD '01*. 2001.

[80] Daniel L. K. Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences* 111.23 (May 2014), pp. 8619–8624. DOI: 10.1073/pnas.1403112111. URL: https://doi.org/10.1073/pnas.1403112111.

[81] H Wen et al. *Deep residual network predicts cortical representation and organization of visual features for rapid categorization. Sci. Rep. 8 (1), 3752 (2018)*. 2018.

[82] Uri Cohen et al. "Separability and geometry of object manifolds in deep neural networks". en. In: *Nature Communications* 11.1 (Dec. 2020), p. 746. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14578-5. URL: http://www.nature.com/articles/s41467-020-14578-5 (visited on 05/13/2022).

[83] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation". In: *PLoS Computational Biology* 10.11 (Nov. 2014). Ed. by Jörn Diedrichsen, e1003915. DOI: 10.1371/journal.pcbi.1003915. URL: https://doi.org/10.1371/journal.pcbi.1003915.

[84] Saeed Reza Kheradpisheh et al. "Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition". In: *Scientific Reports* 6.1 (Sept. 2016). DOI: 10.1038/srep32672. URL: https://doi.org/10.1038/srep32672.

[85] Daniel L K Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". In: *Nature Neuroscience* 19.3 (Feb. 2016), pp. 356–365. DOI: 10.1038/nn.4244. URL: https://doi.org/10.1038/nn.4244.

[86] Haiguang Wen et al. "Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization". In: *Scientific Reports* 8.1 (Feb. 2018). DOI: 10.1038/s41598-018-22160-9. URL: https://doi.org/10.1038/s41598-018-22160-9.

[87] Daniel A. Butts and Daniel S. Rokhsar. "The Information Content of Spontaneous Retinal Waves". In: *The Journal of Neuroscience* 21.3 (Feb. 2001), pp. 961–973. DOI: 10.1523/jneurosci.21-03-00961.2001. URL: https://doi.org/10.1523/jneurosci.21-03-00961.2001.

[88] Daniel A. Butts. "Retinal Waves: Implications for Synaptic Learning Rules during Development". In: *The Neuroscientist* 8.3 (June 2002), pp. 243–253. DOI: 10.1177/1073858402008003010. URL: https://doi.org/10.1177/1073858402008003010.

[89] Benjamin Lansdell, Kevin Ford, and J. Nathan Kutz. "A Reaction-Diffusion Model of Cholinergic Retinal Waves". In: *PLoS Computational Biology* 10.12 (Dec. 2014). Ed. by Olaf Sporns, e1003953. DOI: 10.1371/journal.pcbi.1003953. URL: https://doi.org/10.1371/journal.pcbi.1003953.

[90] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[91] Jure Zbontar et al. "Barlow twins: Self-supervised learning via redundancy reduction". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12310–12320.

[92] Randall Balestriero et al. *A Cookbook of Self-Supervised Learning*. 2023. arXiv: 2304.12210 [cs.LG].

[93] Peiran Gao et al. "A theory of multineuronal dimensionality, dynamics and measurement". In: (Nov. 2017). DOI: 10.1101/214262. URL: https://doi.org/10.1101/214262.

[94] Michael Pecka et al. "Experience-Dependent Specialization of Receptive Field Surround for Selective Coding of Natural Scenes". In: *Neuron* 84.2 (Oct. 2014), pp. 457–469. DOI: 10.1016/j.neuron.2014.09.010. URL: https://doi.org/10.1016/j.neuron.2014.09.010.

[95] Giulio Matteucci and Davide Zoccolan. "Unsupervised experience with temporal continuity of the visual environment is causally involved in the development of V1 complex cells". In: *Science Advances* 6.22 (May 2020). DOI: 10.1126/sciadv.aba3742. URL: https://doi.org/10.1126/sciadv.aba3742.

[96] Nana Nishio et al. "The role of early visual experience in the development of spatial-frequency preference in the primary visual cortex". In: *The Journal of Physiology* 599.17 (Aug. 2021), pp. 4131–4152. DOI: 10.1113/jp281463. URL: https://doi.org/10.1113/jp281463.

[97] H. Sherk and M. P. Stryker. "Quantitative study of cortical orientation selectivity in visually inexperienced kitten". In: *Journal of Neurophysiology* 39.1 (Jan. 1976), pp. 63–70. DOI: 10.1152/jn.1976.39.1.63. URL: https://doi.org/10.1152/jn.1976.39.1.63.

[98] H. Ko, T. D. Mrsic-Flogel, and S. B. Hofer. "Emergence of Feature-Specific Connectivity in Cortical Microcircuits in the Absence of Visual Experience". In: *Journal of Neuroscience* 34.29 (July 2014), pp. 9812–9816. DOI: 10.1523/jneurosci.0875-14.2014. URL: https://doi.org/10.1523/jneurosci.0875-14.2014.

[99] James B Ackman and Michael C Crair. "Role of emergent neural activity in visual map development". en. In: *Current Opinion in Neurobiology* 24 (Feb. 2014), pp. 166–175. ISSN: 09594388. DOI: 10.1016/j.conb.2013.11.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959438813002225 (visited on 04/25/2022).

[100] Hong-Ping Xu et al. "Retinal Wave Patterns Are Governed by Mutual Excitation among Starburst Amacrine Cells and Drive the Refinement and Maintenance of Visual Circuits". In: *The Journal of Neuroscience* 36.13 (Mar. 2016), pp. 3871–3886. DOI: 10.1523/jneurosci.3549-15.2016.

[101] Alexandre Tiriac, Karina Bistrong, and Marla B. Feller. *Retinal waves but not visual experience are required for development of retinal direction selectivity maps*. en. Tech. rep. Section: New Results Type: article. bioRxiv, Mar. 2021, p. 2021.03.25.437067. DOI: 10.1101/2021.03.25.437067. URL: https://www.biorxiv.org/content/10.1101/2021.03.25.437067v1 (visited on 04/25/2022).

[102] Manel Baradad et al. "Learning to See by Looking at Noise". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=RQUl8gZnN7O.

[103] Sahar Behpour, David J. Field, and Mark V. Albert. "On the Role of LGN/V1 Spontaneous Activity as an Innate Learning Pattern for Visual Development". In: *Frontiers in Physiology* 12 (2021). ISSN: 1664-042X. URL: https://www.frontiersin.org/article/10.3389/fphys.2021.695431 (visited on 05/17/2022).

[104] Christiane Voufo et al. "Circuit mechanisms underlying embryonic retinal waves". In: *eLife* 12 (Feb. 2023). DOI: 10.7554/elife.81983. URL: https://doi.org/10.7554/elife.81983.

[105] I. L. Hanganu, Y. Ben-Ari, and R. Khazipov. "Retinal Waves Trigger Spindle Bursts in the Neonatal Rat Visual Cortex". In: *Journal of Neuroscience* 26.25 (June 2006), pp. 6728–6736. DOI: 10.1523/jneurosci.0752-06.2006. URL: https://doi.org/10.1523/jneurosci.0752-06.2006.

[106] Yasunobu Murata and Matthew T Colonnese. "An excitatory cortical feedback loop gates retinal wave transmission in rodent thalamus". In: *eLife* 5 (Oct. 2016). DOI: 10.7554/elife.18816. URL: https://doi.org/10.7554/elife.18816.

[107] Bernd Illing et al. "Local plasticity rules can learn deep representations using self-supervised contrastive predictions". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: https://openreview.net/forum?id=Yu8Q6341U7W.

[108] Artur Luczak, Bruce L. McNaughton, and Yoshimasa Kubo. "Neurons learn by predicting future activity". In: *Nature Machine Intelligence* 4.1 (Jan. 2022), pp. 62–72. DOI: 10.1038/s42256-021-00430-y. URL: https://doi.org/10.1038/s42256-021-00430-y.

[109] Davide Zoccolan et al. "A rodent model for the study of invariant visual object recognition". In: *Proceedings of the National Academy of Sciences* 106.21 (May 2009), pp. 8748–8753. DOI: 10.1073/pnas.0811583106. URL: https://doi.org/10.1073/pnas.0811583106.