Of Machines and Men: Searching for the What, When, and Where of Perception

By

David A. Tovar

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Neuroscience

September 30, 2021

Nashville, Tennessee

Approved:

Mark Wallace, Ph.D.

Randolph Blake, Ph.D.

Alexander Maier, Ph.D.

Ramnarayan Ramachandran, Ph.D.

Frank Tong, Ph.D.

Dedication

To my mentors past and present who taught me invaluable skills and to my students who became my teachers along the way.

Acknowledgment

Pursuing a PhD is a journey full of ups and downs, where the end seems tantalizing close at times and then never ending at others. To some extent, that this is the conclusion of it, seems somewhat surreal. However, I can say that I've enjoyed each step and I feel fortunate to have had the people in my life that allowed me to have these experiences. Reflecting on the journey from start to finish, I ironically feel like my PhD experience breaks down into three parts, much like the dissertation itself.

The first part of my PhD was one of uninhibited exploration—any idea that I found interesting I would pursue. It does not make for the most efficient PhD, but it does make for a rewarding one. For providing me both the resources and the freedom to really explore my interests, I want to start off by thanking my advisor Mark Wallace. Not only did you provide me with financial resources, but you also provided guidance on the many projects I threw your way. Additionally, you allowed me to assemble a great team of undergraduates over the course of my PhD —Andrew, Garrett, Nitya, James, Courtney, Jonathan, Emma, Chidinma, Cokie. Together, we brought my ideas as well as their unique ideas to fruition. And truthfully, I feel that I learned as much from them as much as they learned from me. I am proud of what we accomplished together, both the projects that we published, as well as the ones where we simply learned new skills.

The second part of my journey was where I really honed in on the skillsets needed to complete this dissertation by visiting other labs for research visits. Here, I have to start off by thanking Tom Carlson, who by now is probably tired of my messages of gratitude. Tom was my undergraduate research advisor at the University of Maryland, and then became an essential part of my PhD as I visited his lab over at the University of Sydney. I enjoyed every minute I was in Sydney, and want to especially thank Tijl, Amanda, and Lina for the daily coffee talks that inspired a number of projects, and for your friendships. Shortly after, I also had the good fortune of visiting Micah Murray's lab in Switzerland where I com-

bined some of the decoding techniques with multisensory research paradigms. Micah, I am extremely thankful for your warm welcome to your lab, and our continued collaborations with several members in your lab—Paul, Ruxy, Nora, Anna, Solange.

The third part of the PhD is where I was finally ready to put everything together. And this is where my dissertation committee—Alex, Randolph, Ram, and Frank were invaluable. Starting off with Alex, your guidance and tutelage in how to best present results in figures as well as how to frame scientific stories was priceless. You made projects fun—both with your obvious scientific curiosity and your desire to learn more about the decoding analyses so we could apply them in novel ways to monkey laminar recordings. Randolph, it has been an honor working with you. I was reading about your work when I was a young undergraduate, and then to be able to sit down and work through projects together with you was a privilege. I truly appreciate how patient you were in working through my ideas and our chats in your office. Ram, you were a beacon of practical advice and were incredibly generous with your time. I would come to your office for one question, and I would spend hours talking through anything and everything. Coming from a vision science lab, my chats with you were really my introduction into the auditory world. Your blunt honesty was refreshing and I think I avoided a number of traps because of it. Thank you. Lastly, Frank as the machine learning expert, you became an essential arbiter on how I was applying different analyses to a variety of questions. While it was admittedly disheartening at times to hear that I needed to really reconsider a particular result, it was essential and critical to growing as a scientist. Also, thank you for inviting me to your neural network journal club where I was able to really expand on my ideas.

Last but not least, I want to thank my many friends, my lab mates past and present, and my family. I especially want to thank my current lab mates-Collins, Adriana, and Sarah, who have been critical in helping me with lab matters this past year as I conducted the majority of my research remotely. While my family knows how much they mean to me, I do want to really acknowledge how much of a constant pillar of support they've been for

me throughout my life. Without you, I would have been able to accomplish what I have done.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

General Introduction

"Every act of perception is to some degree an act of creation and every act of memory is to some degree an act of imagination"

-Oliver Sacks

When I was a child, I would often lay in the grass and take in the world around me. I would be in awe at how my senses were filled with information as I felt a caterpillar crawl up my leg or heard a dog bark in the distance. In these moments, I was always left wondering: how did I distinguish the caterpillar from the blades of grass? How did I instantly know it was a dog and not another animal? While I have added some layers of refinement, these are fundamentally the same questions I am asking in my studies today. How does the brain transform an exuberant number of incoming noisy signals at imperfect noisy receptors, and convert them into meaningful percepts? Ostensibly, this process must involve a number of different neural computations. Thus, understanding the process of perception requires a deeper dive into the underlying neural computations. Specifically, what neural computations are being performed by the brain and what features are extracted from incoming signals as a result? When do these neural computations occur and do they contain any temporal patterns? And finally, where in the brain do they occur and do these computations form repeating motifs that are present throughout several brain areas?

Studying the what, when, and where of neural computations benefits from analyzing the brain at different levels of detail in order to provide varying constraints to the questions. In this dissertation, I have adopted Marr's computational, algorithmic, and implementation levels of analysis (Marr, 1982) as initial guides of different levels of description to approach perceptual processes. While there are a number of criticisms regarding Marr's information processing framework, such as the independence between the levels (Bechtel & Shagrir, 2015; McClamrock, 1991), overemphasis of the computational level (Love, 2015),

omission of learning and development (Poggio, 2012), Marr's levels nevertheless provide a helpful framework for studying perceptual phenomena with varying constraints.

This dissertation is thus divided into three main parts (Figure 1.1) beginning with the implementation level and ending at the computational level. I will investigate: 1) information processing embedded within neural spikes, local field potentials, and current source density within a localized microcircuit 2) broaden out to whole field EEG recordings to study algorithms the brain uses to optimize object recognition, and 3) finally compare object recognition between humans and artificial neural networks to find how these two model systems of vision converge (or diverge). By focusing individually at each of these levels, I will sacrifice varying degrees of specificity in answering the what, when, and where of neural computations. For example, focusing on the implementation level by studying a specific neural circuit in primary visual cortex (V1) will necessarily constrain the neural computations I can uncover in other brain areas in response to a visual stimulus. However, I will gain improved resolution into when and where within V1 these neural computations are actualized. Similarly, I will lose spatial and temporal resolution when I study the algorithm and computational levels but gain insights into interactions between brain areas.

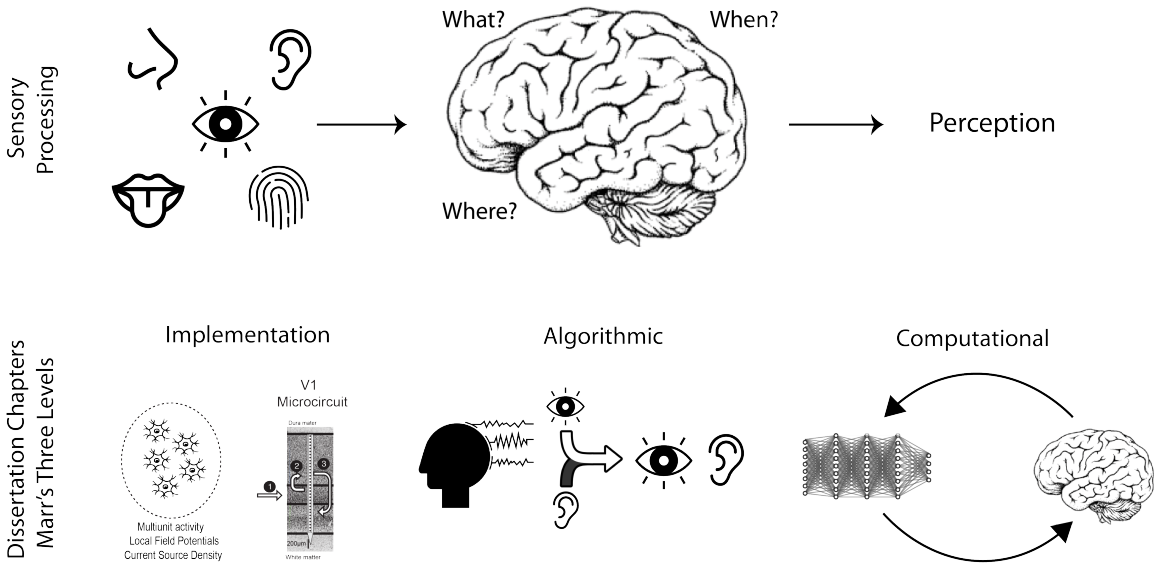

Figure 1.1: Dissertation organization and project summary.

A common thread that will pervade all of the chapters will be the use of machine learning decoding methods to characterize the information present across the respective neural spikes, current source density, local field potential, and layers activations within artificial neural networks. While there are a number of different methods including reverse correlations, choice probability, information theory, amongst others that could characterize stimulus features as well, I have chosen decoding methods because they provide a rich framework to generalize information states in space (Cichy, Pantazis, & Oliva, 2014), time (Carlson, Hoogendoorn, Kanai, Mesik, & Turrett, 2011; King & Dehaene, 2014), frequency, and between model systems (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, Mur, Ruff, & Kiani, 2008a; Kriegeskorte, Mur, & Bandettini, 2008). In effect, I will search for the what when and where of perception using "machines" as both a tool and as a comparative model to "men".

1.1 Implementation Level: Information embedded in neural spikes, local field potentials, and current source density

Looking at the physical properties that constitute the brain can be a daunting task. There are several different scales at which one can investigate the physical embodiment of perception: from the large structural white matter tracks that connect the two hemispheres together to the structure of specific proteins present in signaling cascades. Beyond scale, the physical properties of the brain can also be divided into different broad categories: electrical, chemical, structural, and vasculature. These categories are tightly interwoven and interdependent. For example, it is the chemical gradient between intercellular and extracellular spaces that leads to inward and outward electrical currents. While nonspecific, these categories serve as conceptual scaffolds and are closely linked to the tools we use in measuring brain function. In this section, I will limit my discussion to the measurement of the brain's electrical properties while ignoring other physical properties, with one notable exception when discussing how structure/cytoarchitecture directly affects electrical

measurements.

1.1.1 Action Potentials

Electrical activity in the brain can be measured using action potentials (neural spikes), local field potentials, and current source density. Each of these signals have slightly different biophysical origins. An action potential is observed when a neuron reaches its threshold potential leading to an abrupt depolarization due to changes in membrane permeability from the opening of voltage gated channels (Bean, 2007). Action potentials have historically been recorded either as single-unit activity or as multi-unit activity. Single unit activity is achieved by isolating a single neuron either extracellularly by being sufficiently close to a neuron or by piercing the neuron's membrane altogether in patch clamp or voltage clamp techniques (Perkins, 2006). Isolating individual action potentials from neurons without piercing the membrane, or spike sorting, can be difficult and subjective (Neymotin, Lytton, Olypher, & Fenton, 2011). While newer algorithms, including ones making use of convolutional neural networks, have shown promise in performing automated spike sorting (Buccino et al., 2020; Tolooshams, Song, Temereanca, & Ba, 2019), an alternative is to use multiunit activity if there is less emphasis on isolating individual units. Multiunit activity does not attempt to isolate individual neuron contribution but rather isolates the action potential signal more generally through the use of filters (Zeitler, Fries, & Gielen, 2006).

1.1.2 Local Field Potentials

In contrast to the action potential, which reflects the all-or-none depolarization of a neuron, the local field potential (LFP) is more complex and reflects a non-specific conglomeration of neuronal signals, including contributions from action potentials (Buzsáki, Anastassiou, & Koch, 2012), sodium currents (Ray, Crone, Niebur, Franaszczuk, & Hsiao, 2008), calcium currents (Schiller, Major, Koester, & Schiller, 2000), and gap junctions (Traub & Bibbig, 2000). Obtained by filtering out the low frequency signals from recordings, LFPs also capture the graded potentials, the input signals that do not depolarize a

cell enough to reach the threshold potential (Bijanzadeh, Nurminen, Merlin, Clark, & Angelucci, 2018).

A number of studies have found the LFP to be of particular behavioral and mechanistic relevance (for review see Fries, 2015), in some instances explaining behavior better than action potentials (Pesaran, Pezaris, Sahani, Mitra, & Andersen, 2002). In addition, different frequency bands within the LFP signal contain different types of information, with some associated more with feedforward processes, while other frequency bands associated with feedback processing (Bastos et al., 2015; Belitski et al., 2008; Van Kerkoerle et al., 2014). These oscillatory components have been particularly relevant in multisensory research, which I will expand upon in section 1.2.

However, there are a number of issues of interpretability with LFPs, in particular linking the relative contribution of different components to different frequency bands. For example, it has long been believed that high gamma is tightly linked to action potentials (Mukamel et al., 2005; Nir et al., 2007). However, a study which manipulated how effective the stimuli was in driving responses, found that MUA and high gamma disassociate in both primary auditory and visual cortex, especially during the sustained response for V1 and to a lesser degree A1 (Leszczyński et al., 2020). Thus, the relationship between different components, such as action potentials, to different frequency bands of the LFP is not well known. These are compounded by issues in localizing the origin of LFP signals due to passive volume conduction.

1.1.3 The role of volume conduction in local field potentials

Measuring the amount of spatial spread of the local field potential is difficult for a number of reasons. When a local neuron is activated, the signal may get propagated by exciting other neurons in the network, leading a traveling wave of activity (Sato, Nauhaus, & Carandini, 2012; Zanos, Mineault, Nasiotis, Guitton, & Pack, 2015), but it can also simply volume conduct. Therefore, it is difficult to know whether the stimuli evoked activity is

leading to LFP spread through passive or active ways (Dubey & Ray, 2016). Additionally, modeling work (Lindén et al., 2011) has shown that cell morphology and signal correlations between populations of nearby cells is a considerable factor in determining the extent of passive spread. For example, with uncorrelated synaptic activity, the spatial reach can be a few micrometers (Xing, Yeh, & Shapley, 2009) but for correlated signals in pyramidal cells the spread can be several millimeters (Kajikawa & Schroeder, 2011). Pyramidal cells, due to their asymmetry are amongst the largest contributors to the LFP signal, with symmetrical cells contributing considerably less to the LFP signal. In addition horizontal connections in the superficial layers which can lead to higher coherence have also been shown to be contributing factors in LFP spread (Dubey & Ray, 2016).

1.1.4 Examples of volume conduction potentially leading to incorrect localization inferences

There are a number of studies that have initially localized processes to various brain locations that have been later shown to reflect neural activity that was volume conducted from other distant brain structures. For example, theta oscillations in the dorsal lateral striatum were initially proposed to mediate behaviorally relevant interactions between striatum and cortex (Tort et al., 2008; von Nicolai et al., 2014). However, a recent study applied a bipolar derivation to the LFP to reduce volume conduction, and found that theta oscillations disappeared altogether from the striatum (Lalla, Rueda Orozco, Jurado-Parras, Brovelli, & Robbe, 2017). Similarly, local field potentials that were once thought to originate in the lateral habenula were found to be in fact volume conducted from theta rhythms originating in the hippocampus (Bertone-Cueto et al., 2020). In this study, theta waves persisted in the lateral habenula, despite pharmacological inactivation to the lateral habenula using a sodium channel blocker. Furthermore, independent component analysis showed that the lateral habenula was not the generator of the theta signal.

Similar localization problems have emerged in multisensory studies, where areas of

sensory convergence have been labeled as sites of cross-modal modulation when they are in fact volume conducted spread from two separate sensory areas (Galindo-Leon et al., 2019; Kajikawa, Smiley, & Schroeder, 2017). For example, one study used laminar probes in inferotemporal (IT) cortex and auditory cortex while macaques either viewed faces of other macaques or heard vocalizations, or both and found that the LFP signals in auditory cortex from visual stimulus presentations were no longer found in the CSD signal (Kajikawa et al., 2017). This study and others demonstrate the need to quantify and characterize local and distant signals, as volume conduction can particularly become an issue in localization when the distant signal is stronger than the local signal.

1.1.5 Current Source Density

One way to reduce the effects of volume conduction is to use the current source density (CSD), which is the second spatial derivative of the LFP (Mitzdorf, 1985). The CSD signal is composed of sinks which are the active inward currents, and sources which are the accompanying passive and equal passive outward currents, dissipated in time. The main contributors to the CSD signal are dendritic excitatory post synaptic potentials, with relatively less inhibitory post synaptic potentials coming from the soma (Mitzdorf, 1985). While the CSD signal might contain some passive diffusive and displacement current outside of synaptic transmembrane currents, these are typically only present at frequencies below 4hz (Gratiy et al., 2017). Interestingly, the CSD has shown to generally be a more complex signal than LFP and MUA, requiring more principle components to explain signal variance (Einevoll et al., 2007; Schaefer, Kössl, & Hechavarría, 2017). Thus, it captures the temporal and spatial summation found in the LFP signal, preserving the complexity in that signal, but also eliminates much of the volume conduction. It is for this reason, that a number of studies have adopted the CSD and CSD derived signals in order to quantify the amount of volume conduction present in a given area (Kajikawa & Schroeder, 2011, 2015; Kajikawa, Smiley, & Schroeder, 2017)

1.1.6 Constraints at the implementation level

The primary constraints of the implementation level is that the meaning of the findings largely depend on the algorithms and functions used at the two other levels. The biophysical components of the action potential, LFP, and CSD become meaningful in the context of the information these signals convey. It was for this reason that David Marr observed, "Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done" (Marr, 1982). Nevertheless, without proper understanding at the implementation level, incorrect inferences can be made regarding function and the algorithms underlying those functions as discussed earlier in this introduction. Therefore, the implementation level is best studied when the "what" of the neural computation has been previously explored. Thus, in chapter 2 and 3 and of this dissertation, I will constrain the what to specific features, namely eye of origin, orientation, and stimulus history, while studying where and when these features are extracted within the V1 microcircuit. This type of study can then be expanded upon with further studies with added complexity, such as exploring how features are combined when they are present in disparate sensory modalities, which I explore in the Appendix chapter. The combination of the senses will be the focus of the algorithmic level in the next section.

1.2 Algorithmic Level: Combining the senses

The integration of multisensory cues depends on the stimulus properties of incoming stimuli (Ernst & Banks, 2002; Parise & Ernst, 2016). For the brain to integrate and weigh the relative reliability of incoming sensory cues (Ernst & Banks, 2002; Morgan, DeAngelis, & Angelaki, 2008), they must be sufficiently close in space (Meredith & Stein, 1986) and time (Meredith, Nemitz, & Stein, 1987). In addition, the semantic congruence of sensory cues influences how well multisensory signals are combined (Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004). These multisensory integration principles, as written above, have traditionally been formulated from a stimuli-centric perspective. However, from a

brain-centric perspective these principles can be reduced back to the original three questions presented in this dissertation: where, when, and what types of sensory information must be present in the brain for integration to occur. Said in another way, these become questions of what stimulus features are extracted from each sensory modality and where and when in the brain might they converge.

Thus, understanding the neural underpinnings of how the brain combines the senses can be improved through a broader look into its functional organization. The neocortex has been commonly segmented into areas dedicated to processing incoming information from our five senses (Felleman & Van Essen, 1991). However, this compartmentalization has been questioned by many studies (Kayser, Petkov, Augath, & Logothetis, 2005; Kayser, Petkov, & Logothetis, 2008; Martuzzi et al., 2007; Murray et al., 2005), leading some to the other extreme—is the entirety of neocortex multisensory (Ghazanfar & Schroeder, 2006)? A number of fMRI studies in blind individuals have shown that in the absence of vision, visual cortex activation commonly associated with visual objects is utilized to encode sound objects (Amedi, Raz, Pianka, Malach, & Zohary, 2003; van den Hurk, Van Baelen, & Op de Beeck, 2017; Vetter, Smith, & Muckli, 2014). Similar recruitment of auditory areas and reweighting of visual cues has been found in deaf individuals and cochlear implant users (Benetti et al., 2017; Bola et al., 2017; Butera et al., 2018). Overall, these studies demonstrate the brain's capacity for marked cross-modal plasticity, in which areas normally associated with one sensory modality can be influenced (and even taken over) by other sensory modalities. Further, they speak to a general ability of the brain to use information across senses to optimize encoding even at early areas.

1.2.1 Evidence of feedforward cross-modal modulation

Conceptually, there are two broad ways by which disparate senses might modulate each other—during the feedforward pass of sensory processing or through convergence in association cortices following the initial feedforward sweep and subsequent feedback (Brand-

man, Avancini, Leticevscaia, & Peelen, 2020). Crossmodal activation early along the sensory hierarchy, suggesting potential feedforward modulation, has been found in a number of studies. For instance, an fMRI optogenetic study in rats found that excitation of infragranular excitatory pyramidal neurons in V1 enhanced auditory brainstem BOLD responses in the inferior colliculus (Leong et al., 2018). At the level of the cerebral cortex, an fMRI study found that noise bursts activated primary visual cortex and checkerboards activated primary auditory cortex, and when presented together these stimuli shortened the latency of the hemodynamic BOLD response in each area, suggesting multisensory facilitation (Martuzzi et al., 2007). In another fMRI study, investigators showed movies consisting of video, audio, and audiovisual components to awake and anesthetized macaques. Here, they found that core and belt auditory cortical areas were activated by just the visual components of the movie, and demonstrated audiovisual convergence in the caudal portion of primary auditory cortex, as well as in belt and parabelt areas (Kayser et al., 2008). Similarly, touch has been shown to modulate activity in early auditory areas with integration of touch and sound in the auditory caudal belt (Kayser et al., 2005). Using EEG, combined somatosensory and auditory stimulation has been found to elicit multisensory responses greater than the summed responses of either sound or touch alone as early as 50ms post-stimulus onset (Murray et al., 2005).

A number of different mechanisms may underlie the modulation of feedforward auditory processes. One potential mechanism is through oscillatory phase resets across the different sensory modalities (Fries, 2015). Links between phase reset and perception were found in an electrocorticography (ECoG) study in which epilepsy patients performed a speeded reaction time test in which they were asked to identify the presence of visual, auditory, and audiovisual stimuli. In the audiovisual condition, it was found that visual stimulation modulated auditory activity via phase reset in delta and theta bands. Furthermore, stronger synchrony between regions led to faster reaction times (Mercier et al., 2015). Similar phase resets have also been noted in a number of other studies (Romei, Gross, &

Thut, 2012; Simon & Wallace, 2017). However, it is important to note that oscillations can also play a role through attentional mechanisms with phase resets coming through feedback from supramodal areas (Lakatos et al., 2009). Further mechanisms by which other sensory modalities might influence auditory processes include nonspecific increases in membrane potential. They may come from increased arousal or other mechanisms such as stochastic resonance—the phenomenon where inserting noise into a non-linear system such as the human brain paradoxically increases perceptual awareness (Fujioka, Ross, Kakigi, Pantev, & Trainor, 2006; Lugo, Doti, & Faubert, 2008). Interestingly, these mechanisms do not rely on the stimulus being semantically congruent in order to enhance sensory processing.

1.2.2 Evidence of feedback cross-modal modulation

In contrast, top-down enhancement from feedback processes rely on higher level semantic properties to help with causal inference, helping bind sensory stimuli that are coming from a common source (Körding et al., 2007). Speech in particular relies on binding the semantic components found in the visual and auditory stream. In a study using EEG and fMRI, subjects listened/viewed auditory and visual syllables alone, congruent audiovisual syllables, and incongruent syllables. It was found that the reliability of the visual component influenced connectivity between visual and auditory cortices, but the congruence of the audiovisual stimulus determined the connectivity between superior temporal sulcus (STS) and primary visual and auditory areas (Arnal, Morillon, Kell, & Giraud, 2009). Further MEG and EEG studies found that there was a shift in oscillations from delta oscillations (3-4 Hz) in congruent speech to beta high-gamma coupling (15 Hz, 60-80 Hz) in incongruent and noisy speech in STS (Arnal, Wyart, & Giraud, 2011; Schepers, Schneider, Hipp, Engel, & Senkowski, 2013). A recent EEG study has further found that delta oscillations (1-4 Hz) specifically tracks speech comprehension, whereas theta (4-8 Hz) tracks speech clarity (Etard & Reichenbach, 2019). To further investigate the role of vision on speech comprehension, one study manipulated the timing between visual and auditory stimuli. In

this study, it was found that perception was better when audio lagged behind video, and resulted in reduced activity in STG, presumably due to inhibition of phonemes that would not be compatible with the video (Karas et al., 2019). These results complement another study which manipulated subjects' expectations of upcoming words, showing priming effects in STG at about 100 ms latency (Wang, Zhang, Zou, Luo, & Ding, 2019). Together, these results point to the importance of the STG in speech perception.

While top-down modulation occurs in association cortices, such as the STG, top-down influences can extend as far back as primary sensory cortices. A recent MEG study showed that visual lip reading can create a coarse auditory speech representation in early auditory cortices, independent of initial auditory input (Bourguignon, Baart, Kapnoula, & Molinaro, 2020). Complementing this finding, a study found frequency specific neural patterns from auditory predictions that activated auditory cortex in a tonotopic fashion (Demarchi, Sanchez, & Weisz, 2019). The interplay between feedforward and feedback were delineated further in a 7T fMRI study where subject viewed visual, auditory and audiovisual stimuli with varying levels of attention. Remarkably, they found that audiovisual interactions were found most prominently in infragranular layers of primary auditory cortex and attentional influence present in supragranulars layer, suggesting distinct circuits for these processes (Gau, Bazin, Trampel, Turner, & Noppeney, 2020).

1.2.3 The role of causal structure in cross-modal modulation

The interplay between feedforward and feedback activity has led to further exploration of the role of causal inference in multisensory integration. Recent EEG studies have suggested that multisensory integration occurs in a hierarchical manner beginning with an initial segregation of information at the level of the early sensory cortices, followed by information fusion according to stimulus reliability in intermediate areas, and finally by causal inference in decision-making areas which ultimately determines whether the stimuli should remain fused or segregated (Cao, Summerfield, Park, Giordano, & Kayser, 2019; Rohe &

Noppeney, 2018). However, what defines an early area, intermediate area, and area needed for decision making? And is this gradient fixed or can it change depending on how relevant the multisensory information is to behavior? These are important questions as even within typical integration sites such STG, demarcations have been found between anterior and posterior STG with decisional activity localizing to more posterior regions (Ozker, Schepers, Magnotti, Yoshor, & Beauchamp, 2017). The demarcation is corroborated by studies that show anterior STG responds more vigorously to clear auditory components while posterior STG responds more vigorously when speech has lower signal to noise, suggesting that posterior STG is more sensitive to the reliability of the incoming visual and auditory signals and thus more suited to perform multisensory integration (Ozker et al., 2017).

1.2.4 Towards characterizing the stimulus feature contained in sensory streams

The majority of the multisensory literature reviewed thus far have relied on univariate analyses, using response magnitudes as a way to gauge multisensory integration in both EEG and fMRI. However, a larger BOLD or EEG response does not necessarily equate to more information present at a given location (Harrison & Tong, 2009; Jehee, Brady, & Tong, 2011; Kok, Jehee, & de Lange, 2012; Laurienti, Perrault, Stanford, Wallace, & Stein, 2005). Furthermore, it is becoming increasingly evident that an understanding of how the brain codes for stimulus properties and their respective reliability will requires studying neuronal populations (Ma, Beck, Latham, & Pouget, 2006). In particular, multivariate pattern analysis (decoding) has been useful in abstracting the increased information present in a multisensory signal when compared to unisensory signals (Jung, Larsen, & Walther, 2018) rather than activation. Additionally, a decoding framework makes it possible to fuse the information gained from EEG and fMRI and place them into a common computational space with the use of representational similarity analysis (Cichy & Pantazis, 2017; Radoslaw Martin Cichy, Pantazis, Oliva, 2014, 2016; Kriegeskorte, Mur, & Bandettini, 2008b).

1.2.5 Representational Similarity Analysis (RSA) as a framework for testing algorithms

The basic premise behind RSA (Kriegeskorte et al., 2008b) is to use similarity as a metric between two "entities" to construct a representational map or geometry between all of the possible "entities" of interest. I purposely use the vague word "entity" here, as what is compared can be completely arbitrary and all depends on the research questions being asked. For example, the "entity" can be the neurophysiological responses to visual objects, or it can be of the relative number of vertical lines contained within visual objects. The similarity metric that is used is also somewhat arbitrary as similarity can be determined by the correlation between the entities, or the Euclidean distance, or determined through cross-validated decoding performance in distinguishing between two entities. The comparison of neural responses shares many common threads to earlier perceptual frameworks (see 'similarity rule' in Teller, 1984), but what has made RSA an especially powerful tool in recent years is the use of similarity across several stimuli comparisons to construct representational dissimilarity matrices (RDMs). The RDMs can be thought of as representational geometries, mapping out the unique distance/difference each entity has in relation to every other entity measured. The key advance in this process is that all of the entities are now within representational space and are no longer limited by the original measurements used to measure that entity. In other words, the RDMs make it possible to connect millions of voxels to EEG recordings containing 128 channels to laminar recordings with 24 channels to behavior. As a result RSA has become a framework for hypothesis testing of different algorithms (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cecere, Gross, Willis, & Thut, 2017; Cichy et al., 2014; Giordano, McAdams, Zatorre, Kriegeskorte, & Belin, 2013; Kriegeskorte et al., 2008a), as well as for comparing between model systems (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte et al., 2008a; Tovar, Murray, & Wallace, 2020; Xu & Vaziri-Pashkam, 2021). In this regard, RSA also aligns with the 'analogy rule' in Teller's perceptual framework in that psychophysical and physiological data can be

plotted on meaningfully similar axes.

1.2.6 Animacy as an organizing principle in the brain

Using RSA and other analyses, it has been found that one of the guiding principles or algorithms the brain utilizes is whether an object is living (animate) or non-living (inanimate). This organizational principle will be the focus of Chapter 4 and was first noticed in patients with brain damage that exhibited category-specific deficits in naming animate objects (Capitani, Laiacona, Mahon, & Caramazza, 2003; Kolinsky et al., 2002; Warrington & Mccarthy, 1987). Since then, a body of literature in both audition and vision have shown a distinct behavioral, fMRI, and M/EEG divide between animate and inanimate objects (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; De Lucia, Tzovara, Bernasconi, Spierer, & Murray, 2012; Grootswagers, Ritchie, Wardle, Heathcote, & Carlson, 2017; Huth, Nishimoto, Vu, & Gallant, 2012; Kriegeskorte, Mur, Ruff, & Kiani, 2008; Murray, 2006; Ritchie, Tovar, & Carlson, 2015). The division in the brain for animate and inanimate objects is thought to have arisen due to evolutionary forces (Mahon, Anzellotti, Schwarzbach, Zampini, & Caramazza, 2009; New, Cosmides, & Tooby, 2007). Supporting this theory, a study by Kriegeskorte et al. 2008 found a common categorical animacy distinction in monkey inferotemporal (IT) cortex and human IT cortex. The animacy distinction was more prominent in IT than primary visual cortex, and several lower level models of vision were not able to account for the category clustering observed in IT cortex. Despite these shared similarities between species, and evidence of animate and inanimate categories in infants (Simion, Regolin, & Bulf, 2008), it is an open question how much experience affects the animate/inanimate divide with a number of studies showing considerable category effects from stimulus exposure and subject expertise for different categories (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999; Livingstone et al., 2017).

Nevertheless, whether through innate brain development or through experience, the overall categorical nature of animacy has been reinforced with the use of carefully con-

trolled animate/inanimate stimuli that account for shape (Bracci, Ritchie, & Op de Beeck 2017; Ritchie & Op De Beeck 2018). Further, the animate/inanimate category boundary has been used to show that representational spaces are perceptually relevant, linking spatial and temporal properties of representational space with behavior (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; Ritchie, Tovar, & Carlson, 2015). Using the relative distances of objects from the animate inanimate category boundary, studies (Carlson et al., 2014; Ritchie et al., 2015) have been able to make predictions on categorization reaction times. Namely, objects closer to the animate inanimate category boundary are more difficult to distinguish as animate or inanimate and as such will have longer reaction times. Conversely, objects far apart from the category boundary are easier to distinguish as either animate or inanimate and as such have shorter reaction times. These predications fall within the framework of perceptual decision-models which state that evidence close to a decision boundary is more ambiguous, resulting in more decision time, while evidence far from a boundary is less ambiguous, resulting in more rapid decisions (Ashby & Maddox, 1994; Dunovan, Tremel, & Wheeler, 2014; Pike, 1973).

Similarly, auditory studies have also shown animacy to be an abstract category distinction, accounting for lower level features (Giordano et al., 2013; M. M. Murray, 2006). Furthermore, the visual cortex of blind individuals mirror the neural organization of visual objects in sighted individuals, supporting a possible shared semantic animacy distinction between sensory modalities (Bedny, Pascual-Leone, Dodell-Feder, Fedorenko, & Saxe, 2011; Mahon et al., 2009; van den Hurk et al., 2017). Beyond sharing a category distinction for animacy, a common perceptual advantage for animate objects over inanimate has been observed in both modalities. In vision, animate objects are categorized faster than inanimate objects, are consciously perceived more in the attentional blink and are found faster in visual search tasks (Carlson et al., 2014; Jackson & Calvillo, 2013; Lindh, Sligte, Assecondi, Shapiro, & Charest, 2019; New et al., 2007; Ritchie et al., 2015). Auditory studies have similarly found faster categorization times for animate objects (Vogler & Titchener,

2011; Yuval-Greenberg & Deouell, 2009). These behavioral differences suggest that these category classes may have neural encoding differences with more effective processing for animate objects. This difference may be from evolutionary origins with survival depending on recognition and further processing of living stimuli (Laws, 2000). Neurally, the number of specialized areas that have been identified for animate subcategories such as faces in the fusiform face area (FFA) and bodies in the extrastriate body area (EBA) further support an encoding difference between animate and inanimate objects (Downing, Jiang, Shuman, & Kanwisher, 2001; Kanwisher, McDermott, & Chun, 1997).

Furthermore, visual degradation of stimuli selectively contracts the representational space of animate objects without affecting inanimate objects (Grootswagers et al., 2017). The asymmetric compression for degraded animate objects suggests that the neural representational space is malleable and furthermore the initial encoding of neural representations influences how stimulus perturbations warp the representational space. Thus, there is ample opportunity to test how combining visual and auditory information might influence object encoding at the category level.

1.2.7 Constraints at the algorithm level

It is important to note that whole brain recording methods are far removed from the underlying neural spikes discussed in the implementation level section. The fMRI signal in particular has been shown to diverge from neural spiking under conditions of perceptual suppression (Maier et al., 2008; Self, van Kerkoerle, Goebel, & Roelfsema, 2017). In these circumstances, it more closely resembles the low frequency local field potential. Local field potentials in turn form the basis of EEG and MEG studies, which provide an estimation of the synaptic inputs in a given location. However, LFP signals carry potential problems as I have noted previously when investigating multisensory integration, as they are known to volume conduct across electrodes (Kajikawa & Schroeder, 2011). Thus, applying the multisensory concepts of superadditivity (Wallace, Meredith, & Stein, 1998) at a

given electrode becomes difficult as it can simply reflect the activity of a neighboring brain structure without necessarily signifying that there is integration (Laurienti et al., 2005).

1.3 Computational Level: Convolutional neural networks as models of the brain

One of the underlying assumptions when using model systems of the human brain is that models capture a function that the brain performs. For instance, Hubel and Wiesel's finding of orientation columns in cats would lose much of its impact if a cat's visual experience was significantly poorer than humans (Hubel & Wiesel, 1962). One of the reasons non-human primate research is incredibly valuable is the assumed perceptual similarity between the species and ability to therefore generalize results. This assumption is of course bolstered by histological and structural similarities that bare evidence to the shared evolutionary history between the species. While animal models have provided and will continue to provide a wealth of information regarding perceptual experience, they come with their inherent limitations in their ability to probe causal manipulations. Despite the advances in optogenetics, gene editing with CRISPR, DREADDs (designer receptors exclusively activated by designer drugs), as well as electrical stimulation and ablation studies of years past, the ability to flexibly manipulate neural architecture and connectivity is still quite laborious. This type of flexibility however can be gained in artificial models of the brain.

1.3.1 Computer vision and models of vision converge

Early models of the visual system assumed that function would follow form. Models such as HMAX (Poggio & Riesenhuber, 1999) aimed to recreate the architecture of the ventral visual stream, hierarchically building upon orientation tuned filters and then subsequently performing max pooling over receptor fields in downstream areas. In parallel, computer vision was using a number of different approaches, from light detection and ranging sensors (LiDAR) (Huang & Barth, 2009; Li & Olson, 2011) to feature segmentation algorithms. The feature segmentation algorithms required fine-tuning and expert selection of the features that would be extracted from images (Lecun, Bengio, & Hinton, 2015). In

the past decade, these two fields converged in the form of convolutional neural networks. In the 2012 ImageNet computer vision challenge, an 8-layer convolutional neural network (CNNs), AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), outperformed other computer vision models by an order of magnitude. Unlike previous models, CNNs do not require explicit feature selection. Instead, they use labeled examples coupled with gradient descent (LeCun, Haffner, Bottou, & Bengio, 1999) as an optimization algorithm to fine tune the connections and weights of all the layers sequentially in a process termed backpropagation. Early on, it was found that the early layers in these models remarkably resembled the orientation tuning in V1, despite these features never being explicitly programmed into the model (Lecun et al., 2015). When CNNs were used as models of human vision, several groups with different analytical approaches (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016), found that CNNs better explained vision than previous models of vision, including those that were directly biologically inspired such as HMAX (Kubilius, Bracci, & Op de Beeck, 2016). Thus, it has become apparent that form has followed function for vision models.

1.3.2 The rapid progress and sophistication of CNNs has unintentionally made them more brain like

Since the development of AlexNet, modifications to CNNs have continued to push the boundaries of computer performance on visual object recognition tasks. These modifications can be grouped into two broad categories, architecture and training data. A non-exhaustive list of changes to architecture include: increase depth with added layers, multiple sequential convolutional layers (inception layers) (Szegedy et al., 2014), skip connections (He, Zhang, Ren, & Sun, 2015), different types of pooling layers (max, average, pyramidal) (Kleppmann et al., 2018), and recurrence (bidirectional flow of information between layers) (Sherstinsky, 2018). A recent study shows the architectures that most resemble the brain also tend to be the architectures that have the best performance on object

recognition benchmarks, such as the ImageNet contest (Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Schmidt, et al., 2018a).

In terms of training data, a vast majority of visual CNNs are trained using ImageNet, a compilation of images grouped by 1000 object categories that amongst other things includes an inordinate amount of dog breeds. Recent studies have begun to incorporate training images based on ecological categories that are more representative of the objects people mention and are exposed to in everyday life (Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021). These studies have found that CNNs trained on ecological categories are more brain like, primarily assessed using CNN RDMs and neural RDMs. While no benchmarks have been set to compare the performance of these networks with networks trained on ImageNet, there is reason to believe that training set that make the CNNs more brain like will improve network performance. A number of studies have manipulated the ecological nature of the training images to be more similar to how humans learn to recognize objects with encouraging results in terms of network task performance. For example, one study manipulated a network so it was trained using images that have been obtained from video cameras that were mounted on infants and compared it to images obtained from cameras mounted on adults. The network that used the images from cameras mounted on infants had better performance and ability to generalize to new categories (Bambach, Crandall, Smith, & Yu, 2018). Another study emulated the gradual sharpening of an infant's spatial acuity during development, by training a network with images that were initially spatially low pass filtered but progressively sharpened during training. Here, too, the networks that were trained in this coarse to fine manner had the best object recognition performance and were able to generalize the most to image perturbations (Vogelsang et al., 2018; Avbersek, Zeman, & Op de Beeck, 2021).

1.3.3 Perturbing the networks: Adversarial examples to vision and CNNs

While CNNs have been able to model the ventral visual stream remarkably well (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Schmidt, et al., 2018b; Yamins & DiCarlo, 2016; Zhuang et al., 2021), small perturbations, changes to images that would not seemingly affect human visual recognition, can completely change the predictions made by a CNN. These examples are known as adversarial examples. One explanation for why they are thought to arise is because the neural computations in a CNN are more linear than the brain and thus extrapolate to points within its latent (i.e. computational) space that do not exist in real data sets (Goodfellow, Shlens, & Szegedy, 2015). However, other studies have pointed out that comparing human and CNN performance with adversarial examples is not always a reasonable comparison as machines are provided access to all properties of an image as the images are fed digitally, while humans must process the image through an imperfect sensor that is not privy to all of the information that is added to the image under an adversarial attack (Zhou & Firestone, 2019). Additionally, if humans are asked to predict how a CNN might label an adversarial example, they are often quite proficient at doing so, demonstrating that there is some common thread between CNNs and humans (Zhou & Firestone, 2019). Furthermore, if an adversarial image is shown sufficiently quick ($\tilde{6}$0ms as to prevent recurrent processing), they will also fool humans (Elsayed et al., 2018). However, these explanations are not entirely sufficient to explain recent studies that demonstrated that CNNs, including recurrent networks, fail to capture the variance in the fMRI signal if subjects are shown degraded and artificial objects (Xu & Vaziri-Pashkam, 2021). In chapter 4 of this dissertation, I will further explore the role of image perturbations on the correspondence between visual CNNs and human vision.

1.3.4 Visualizing the "black box"

One of the most frequent critiques of CNNs are that they are over parametrized, making them impossible to understand, replacing one black box with another (Goodfellow, Shlens, & Szegedy, 2015b; Kietzmann, Mcclure, & Kriegeskorte, 2019; Ribeiro, Singh, & Guestrin, 2016). While it is true that the millions of computations cannot be individually understood in the way one could understand simpler biologically inspired models, there are number of algorithms that make it easier to unravel some of the inner workings within CNNs. These approaches include DeconvNet (Zeiler & Fergus, 2014), GradCam (Selvaraju et al., 2017) and Google's DeepDream. For this introduction, I will briefly summarize the concept behind DeconvNet (Zeiler & Fergus, 2014). In this this procedure, an image is passed forward through the network up until the layer and particular neuron being visualized. From there, all other activations aside from the chosen neuron are zeroed out within that layer. Then as the name implies, there are a series of deconvolution and reverse pooling steps, using the forward pass of the image as a guide of where to reverse pool the activations and transpose the original convolutions of the image. In effect, these steps can be used to build salience maps of the image properties that were most important for classification. While there is some controversy regarding how rectification (i.e. ReLU units) are incorporated in the deconvolution/backpropagation procedure used to create the saliency maps, it lies beyond the scope of this brief overview (but see Adebayo et al., 2018 for further discussion). Together, these visualization techniques provide rich tools that can be used to manipulate and visualize CNNs.

1.3.5 Expanding beyond visual objection recognition: Using neural networks to model other sensory systems and cognitive processes

While much of the excitement in neuroscience for CNNs was initially found in their ability to model object recognition in the ventral visual stream, they are expanding to serve as models for other brain processes. For example, a recent study found that CNNs that

were trained on a visual categorization task also inadvertently coded for the memorability of the object. The overall activation magnitude of different layers, especially those in the last couple of fully connected layers to a given objects, predicted which objects were memorable and which ones were not. The magnitude code in the CNNs matched the ones in IT cortex (Jaegle et al., 2019). For auditory processes, it was found that early layers ostensibly capturing low level auditory features (timbre, loudness, etc.) of a CNN trained to classify between music genere showed more correspondence with fMRI voxels in anterior STG, while later layers tuned more towards the classification of music genres shared more correspondence with posterior STG (Güçlü, Thielen, Hanke, & Van Gerven, 2016). In addition to modeling fMRI responses, auditory CNN models are now showing human level performance for word and music genre classification tasks, while at the same time exceeding previous standard spectrotemporal models in terms of explained variance of auditory cortex voxel activations (Kell et al., 2018). Further studies have shown that increasing the phonetic similarity of languages used in the training sets of CNNs improves how well they model auditory brain responses (Millet & King, 2021). Much like the advances that improved object recognition in visual CNNs, manipulations to training data, architectures, and preprocessing are leading to improvements in auditory CNNs.

1.3.6 Constraints at the computational level and room for improvement in indexing a "brain-like" network

When comparing two model systems, one of the key challenges is to determine what metrics are used to assess correspondence. For CNNs and brains, this poses a particular challenge as several different analysis decisions are taken when measuring correspondence. The first choice is to choose which brain measure to use as a model. As discussed in the implementation level of this introduction, this is not a trivial choice as the dynamics and information captured varies considerably by brain measure. Thus far, fMRI studies and neural spikes have predominated the types of brain signals analyzed, with relatively fewer M/EEG

studies and no LFP or CSD studies of which I am aware. Another issue is choosing which network layers of the CNN to use to test for correspondence with brain activity. Different studies have taken different approaches, with some using convolutional and fully connected layers (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Mehrer et al., 2021) while others choose pooling layers (O'Connell & Chun, 2018; Xu & Vaziri-Pashkam, 2021). Beyond which layer selection, the analytical framework used to measure brain and network correspondence also vary between studies with some using RSA and RDMs (Cichy, Khosla, et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Mehrer et al., 2021; Xu & Vaziri-Pashkam, 2021) while others use linear transformations, training and cross validation to link fMRI voxels or electrodes to CNN layers (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Geiger, et al., 2018; Yamins & DiCarlo, 2016; Güçlü & van Gerven, 2015). Within RSA approaches, the distance measurement used to build the RDMs from neural network layer activations varies with some using Euclidean distance measurements (Xu & Vaziri-Pashkam, 2021) while other use correlation distance measurements (Cichy, Khosla, et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Mehrer et al., 2021). While these are not all of the possible analysis choices, this non-exhaustive list demonstrates the need to systematically assess the effects of these analysis choices and possibly offer broad recommendations if some analytical steps consistently show better correspondence.

1.4 References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):9505–9515.

Amedi, A., Raz, N., Pianka, P., Malach, R., and Zohary, E. (2003). Early 'visual' cortex

activation correlates with superior verbal memory performance in the blind. *Nature Neuroscience*, 6(7):758–766.

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, 29(43):13445–13453.

Arnal, L. H., Wyart, V., and Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, 14(6):797–801.

Ashby, F. G. and Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, 38(4):423–466.

Bambach, S., Crandall, D. J., Smith, L. B., and Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):1201–1210.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., DeWeerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401.

Bean, B. P. (2007). The action potential in mammalian central neurons. *Nature Reviews Neuroscience*, 8(6):451–465.

Bechtel, W. and Shagrir, O. (2015). The non-redundant contributions of marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science*, 7(2):312–322.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., and Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11):4429–4434.

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2008). Low-frequency local field potentials and spikes in primary visual

cortex convey independent visual information. *Journal of Neuroscience*, 28(22):5696–5709.

Benetti, S., Van Ackeren, M. J., Rabini, G., Zonca, J., Foa, V., Baruffaldi, F., Rezk, M., Pavani, F., Rossion, B., and Collignon, O. (2017). Functional selectivity for face processing in the temporal voice area of early deaf individuals. *Proceedings of the National Academy of Sciences of the United States of America*, 114(31):E6437–E6446.

Bertone-Cueto, N. I., Makarova, J., Mosqueira, A., García-Violini, D., Sánchez-Peña, R., Herreras, O., Belluscio, M., and Piriz, J. (2020). Volume-conducted origin of the field potential at the lateral habenula. *Frontiers in Systems Neuroscience*, 13(January).

Bijanzadeh, M., Nurminen, L., Merlin, S., Clark, A. M., and Angelucci, A. (2018). Distinct laminar processing of local and global context in primate primary visual cortex. *Neuron*, 100(1):259–274.e4.

Bola, , Zimmermann, M., Mostowski, P., Jednoróg, K., Marchewka, A., Rutkowski, P., and Szwed, M. (2017). Task-specific reorganization of the auditory cortex in deaf humans. *Proceedings of the National Academy of Sciences of the United States of America*, 114(4):E600–E609.

Bourguignon, M., Baart, M., Kapnoula, E. C., and Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, 40(5):1053–1065.

Bracci, S., Ritchie, J. B., and de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105(June):153–164.

Brandman, T., Avancini, C., Leticevscaia, O., and Peelen, V. M. (2020). Auditory and semantic cues facilitate decoding of visual object category in meg. *Cerebral Cortex*, 30(2):597–606.

Butera, I. M., Stevenson, R. A., Mangus, B. D., Woynaroski, T. G., Gifford, R. H., and Wallace, M. T. (2018). Audiovisual temporal processing in postlingually deafened adults with cochlear implants. *Scientific Reports*, 8(1):1–12.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents-eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, 13(6):407–420.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10(12).

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., and Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, 102(5):1076–1087.e8.

Capitani, E., Laiacona, M., Mahon, B., and Caramazza, A. (2003). *What are the facts of semantic category-specific deficits? A critical review of the clinical evidence*, volume 20.

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10):1–19.

Carlson, T. A., Hoogendoorn, H., Kanai, R., Mesik, J., and Turrett, J. (2011). High temporal resolution decoding of object. *Journal of Vision*, 11(2011):1–17.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., and Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, 26(1):132–142.

Cecere, R., Gross, J., Willis, A., and Thut, G. (2017). Being first matters: topographical representational similarity analysis of erp signals reveals separate networks for audiovisual temporal binding depending on the leading sense. *The Journal of Neuroscience*, pages 2926–16.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016a). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(June):1–13.

Cichy, R. M. and Pantazis, D. (2017). Multivariate pattern analysis of meg and eeg: A comparison of representational structure in time and space. *NeuroImage*, 158(July):441–454.

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462.

Cichy, R. M., Pantazis, D., and Oliva, A. (2016b). Similarity-based fusion of meg and fmri reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, 26(8):3563–3579.

De Lucia, M., Tzovara, A., Bernasconi, F., Spierer, L., and Murray, M. M. (2012). Auditory perceptual decision-making based on semantic categorization of environmental sounds. *NeuroImage*, 60(3):1704–1715.

Demarchi, G., Sanchez, G., and Weisz, N. (2019). Automatic and feature-specific prediction-related neural activity in the human auditory system. *Nature Communications*, 10(1):1–11.

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473.

Dubey, A. and Ray, S. (2016). Spatial spread of local field potential is band-pass in the primary visual cortex. *Journal of Neurophysiology*, 116(4):1986–1999.

Dunovan, K. E., Tremel, J. J., and Wheeler, M. E. (2014). Prior probability and feature predictability interactively bias perceptual decisions. *Neuropsychologia*, 61(1):210–221.

Einevoll, G. T., Pettersen, K. H., Devor, A., Ulbert, I., Halgren, E., and Dale, A. M. (2007). Laminar population analysis: Estimating firing rates and evoked synaptic activity from multielectrode recordings in rat barrel cortex. *Journal of Neurophysiology*, 97(3):2174–2190.

Elsayed, G. F., Papernot, N., Shankar, S., Kurakin, A., Cheung, B., Goodfellow, I., and Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in Neural Information Processing Systems*, 2018-Decem:3910–3920.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(24):429–433.

Etard, O. and Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 39(29):5750–5759.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

Fries, P. (2015). Rhythms for cognition: Communication through coherence. *Neuron*, 88(1):220–235.

Fujioka, T., Ross, B., Kakigi, R., Pantev, C., and Trainor, L. J. (2006). One year of musical training affects development of auditory cortical-evoked fields in young children. *Brain*, 129(10):2593–2608.

Galindo-Leon, E. E., Stitt, I., Pieper, F., Stieglitz, T., Engler, G., and Engel, A. K. (2019). Context-specific modulation of intrinsic coupling modes shapes multisensory processing. *Science Advances*, 5(4):eaar7633.

Gau, R., Bazin, P. L., Trampel, R., Turner, R., and Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. *eLife*, 9:1–26.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. nature neuroscience 2, 6 (. *June*, 2(6):568–573.

Ghazanfar, A. A. and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6):278–285.

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., and Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, 23(9):2025–2037.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–11.

Gratiy, S. L., Halnes, G., Denman, D., Hawrylycz, M. J., Koch, C., Einevoll, G. T., and Anastassiou, C. A. (2017). From maxwell's equations to the theory of current-source density analysis. *European Journal of Neuroscience*, 45(8):1013–1023.

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., and Carlson, T. A. (2017). Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions. *Journal of cognitive neuroscience*, 29(12):1995–2010.

Güçlü, U., Thielen, J., Hanke, M., and Van Gerven, M. A. (2016). Brains on beats. In *Advances in Neural Information Processing Systems*, pages 2109–2117.

Güçlü, U. and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the

complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Güçlü, U. and van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–17.

Huang, L. and Barth, M. (2009). Tightly-coupled lidar and computer vision integration for vehicle detection. *IEEE Intelligent Vehicles Symposium, Proceedings*, pages 604–609.

Hubel, D. N. and Wiesel, T. N. (1962). And functional architecture in the cat ' s visual cortex from the neurophysiolojy laboratory , department of pharmacology central nervous system is the great diversity of its cell types and inter- receptive fields of a more complex type (part i) and to. *Journal of Physiology*, 160(1):106–154.

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

Jackson, R. E. and Calvillo, D. P. (2013). Evolutionary psychology. *Evolutionary Psychology*, 11(5):1011–1026.

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., and Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife*, 8:1–12.

Jung, Y., Larsen, B., and Walther, D. B. (2018). Modality-independent coding of scene categories in prefrontal cortex. *Journal of Neuroscience*, 38(26):5969–5981.

Kajikawa, Y. and Schroeder, C. E. (2011). How local is the local field potential? *Neuron*, 72(5):847–858.

Kajikawa, Y. and Schroeder, C. E. (2015). Generation of field potentials and modulation of their dynamics through volume integration of cortical activity. *Journal of Neurophysiology*, 113(1):339–351.

Kajikawa, Y., Smiley, J. F., and Schroeder, C. E. (2017). Primary generators of visually evoked field potentials recorded in the macaque auditory cortex. *Journal of Neuroscience*, 37(42):10139–10153.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(11):4302–11.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983.

Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., and Beauchamp, M. S. (2019). The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *eLife*, 8:1–19.

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron*, 48(2):373–384.

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, 18(7):1560–1574. In monkeys, Kayser et al. (2008) found that visual modulation in auditory cortex was sensitive to stimulus asynchrony: i.e., visual stimuli had significant effects only when presented 20–80 ms be- fore the auditory stimuli. This timing difference is consistent with the delayed processing of visual

stimuli compared to other sensory modalities. In macaque monkeys, the visual response latency in V1 is in the range of 20–30 ms, compared to shortest sensory response latencies of about 10 ms in primary auditory cortex, and about 6 ms in primary somatosensory cortex (Schroeder et al., 1998; Sch- roeder and Foxe, 2002; Musacchia and Schroeder, 2009). Visual re- sponses in association areas STP and the intraparietal sulcus occur only slightly later than V1, at about 25 ms (Schroeder and Foxe, 2002). Thus the approximate time frame of 20–80 ms visual-audi- tory disparity of described by Kayser et al. (2008) is at least consis- tent with the possibility that auditory cortex is modulated by connections with very early stages of cortical visual processing, but it does not exclude the possibility that this input comes from downstream association areas.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-haignere, V. S., Mcdermott, J. H., Kell, A. J. E., Yamins, D. L. K., Shook, E. N., and Norman-haignere, V. S. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):1–15.

Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11).

Kiar, L., Zeman, A., and Op de Beeck, H. (2021). Training for object recognition with increasing spatial frequency : A comparison of deep learning with human vision . *bioRxiv*.

Kietzmann, T. C., Mcclure, P., and Kriegeskorte, N. (2019). Oxford research encyclopedia of neuroscience deep neural networks in computational neuroscience explaining brain information processing requires complex , task-performing models. (January):1–29.

King, J. R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210.

Kleppmann, B., Bizer, C., Yaqub, E., Temme, F., Schlunder, P., Arnu, D., and Klinkenberg, R. (2018). Spp:spatial pyramid pooling. *CEUR Workshop Proceedings*, 2191:191–194.

Kok, P., Jehee, J. F. M., and de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2):265–270.

Kolinsky, R., Fery, P., Messina, D., Peretz, I., Evinck, S., Ventura, P., and Morais, J. (2002). The fur of the crocodile and the mooing sheep: A study of a patient with a category-specific impairment for biological things. *Cognitive Neuropsychology*, 19(4):301–342.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):4.

Kriegeskorte, N., Mur, M., Ruff, D. A., and Kiani, R. (2008b). Matching categorical object representations in inferior temporal cortex of mand and monkey. *Neuron*, 60(6):1126–1141.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances In Neural Information Processing Systems*, pages 1–9.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, 12(4):1–26.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9).

Lakatos, P., O'Connell, M. N., Barczak, A., Mills, A., Javitt, D. C., and Schroeder, C. E. (2009). The leading sense: Supramodal control of neurophysiological context by attention. *Neuron*, 64(3):419–430.

Lalla, L., Rueda Orozco, P. E., Jurado-Parras, M. T., Brovelli, A., and Robbe, D. (2017). Local or not local: Investigating the nature of striatal theta oscillations in behaving rats. *eNeuro*, 4(5).

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., and Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, 158(4):405–414.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., and Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, 166(3-4):289–297.

Laws, K. R. (2000). Category-specific naming errors in normal subjects: The influence of evolution and experience. *Brain and Language*, 75(1):123–133.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. *Shape, Contour and Grouping in Computer Vision*, pages 319–345.

Leong, A. T., Dong, C. M., Gao, P. P., Chan, R. W., To, A., Sanes, D. H., and Wu, E. X. (2018). Optogenetic auditory fmri reveals the effects of visual cortical inputs on auditory midbrain response. *Scientific Reports*, 8(1):1–11.

Leski, S., Lindén, H., Tetzlaff, T., Pettersen, K. H., and Einevoll, G. T. (2013). Frequency dependence of signal power and spatial reach of the local field potential. *PLoS Computational Biology*, 9(7).

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A. Y., Tal, I., Haegens, S., Melloni, L., Knight, R. T., and Schroeder, C. E. (2020). Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. *Science advances*, (August):1–13.

Li, Y. and Olson, E. B. (2011). Structure tensors for general purpose lidar feature extraction. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 1869–1874.

Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., and Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features.

Lindén, H., Tetzlaff, T., Potjans, T. C., Pettersen, K. H., Grün, S., Diesmann, M., and Einevoll, G. T. (2011). Modeling the spatial reach of the lfp. *Neuron*, 72(5):859–872.

Livingstone, M. S., Vincent, J. L., Arcaro, M. J., Srihasam, K., Schade, P. F., and Savage, T. (2017). Development of the macaque face-patch system. *Nature Communications*, 8.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, 7(2):230–242.

Lugo, E., Doti, R., and Faubert, J. (2008). Ubiquitous crossmodal stochastic resonance in humans: Auditory noise facilitates tactile, visual and proprioceptive sensations. *PLoS ONE*, 3(8).

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., and Caramazza, A. (2009). Category-specific organization in the human brain does not require visual experience. *Neuron*, 63(3):397–405.

Maier, A., Wilke, M., Aura, C., Zhu, C., Ye, F. Q., and Leopold, D. A. (2008). Divergence of fmri and neural signals in v1 during perceptual suppression in the awake monkey. *Nature Neuroscience*, 11(10):1193–1200.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco.

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J. P., Maeder, P. P., Clarke, S., and Meuli, R. A. (2007). Multisensory interactions within human primary cortices revealed by bold dynamics. *Cerebral Cortex*, 17(7):1672–1679.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, 1(2):185–196.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences of the United States of America*, 118(8):1–9.

Mercier, M. R., Molholm, S., Fiebelkorn, I. C., Butler, J. S., Schwartz, T. H., and Foxe, J. J. (2015). Neuro-oscillatory phase alignment drives speeded multisensory response times: An electro-corticographic investigation. *Journal of Neuroscience*, 35(22):8546–8557.

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 7(10):3215–3229.

Millet, J. and King, J.-R. (2021). Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech.

Mitzdorf, U. (1985). Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and eeg phenomena. *Physiological Reviews*, 65(1):37–100.

Morgan, M. L., DeAngelis, G. C., and Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron*, 59(4):662–673.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Neuroscience: Coupling between neuronal firing, field potentials, and fmri in human auditory cortex. *Science*, 309(5736):951–954.

Murray, M. M. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience*, 26(4):1293–1302.

Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2005). Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cerebral Cortex*, 15(7):963–974.

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598–16603.

Neymotin, S. A., Lytton, W. W., Olypher, V. A., and Fenton, A. A. (2011). Measuring the quality of neuronal identification in ensemble recordings. *Journal of Neuroscience*, 31(45):16398–16409.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., and Malach, R. (2007). Coupling between neuronal firing rate, gamma lfp, and bold fmri is related to interneuronal correlations. *Current Biology*, 17(15):1275–1285.

Ozker, M., Schepers, I. M., Magnotti, J. F., Yoshor, D., and Beauchamp, M. S. (2017). A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. *Journal of Cog*, 29(6):1044–1060.

O'Connell, T. P. and Chun, M. M. (2018). Predicting eye movement patterns from fmri responses to natural scenes. *Nature Communications*, 9(1).

Parise, V. C. and Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, 7:1–9.

Perkins, K. L. (2006). Cell-attached voltage-clamp and current-clamp recording and stimulation techniques in brain slices. *Journal of Neuroscience Methods*, 154(1-2):1–18.

Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., and Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature Neuroscience*, 5(8):805–811.

Pike, R. (1973). Response latency models for signal detection. *Psychological review*, 80(1):53–68.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, 41(9):1017–1023.

Poggio, T. and Riesenhuber, M. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., and Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60-200 hz) in macaque local field potentials and their potential implications in electrocorticography. *Journal of Neuroscience*, 28(45):11526–11536.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier.

Ritchie, J. B. and Op De Beeck, H. (2018). Using neural distance to predict reaction time for categorizing the animacy, shape, and abstract properties of objects. pages 1–19.

Ritchie, J. B., Tovar, D. A., and Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology*, 11(6).

Rohe, T. and Noppeney, U. (2018). Reliability-weighted integration of audiovisual signals can be modulated by top-down attention. *Eneuro*, 5(1):ENEURO.0315–17.2018.

Romei, V., Gross, J., and Thut, G. (2012). Sounds reset rhythms of visual cortex and corresponding human visual perception. *Current Biology*, 22(9):807–813.

Sato, T. K., Nauhaus, I., and Carandini, M. (2012). Traveling waves in visual cortex. *Neuron*, 75(2):218–229.

Schaefer, M. K., Kössl, M., and Hechavarría, J. C. (2017). Laminar differences in response to simple and spectro-temporally complex sounds in the primary auditory cortex of ketamine-anesthetized gerbils. *PLoS ONE*, 12(8):1–28.

Scheeringa, R. and Fries, P. (2019). Cortical layers, rhythms and bold signals. *NeuroImage*, 197(October 2017):689–698.

Schepers, I. M., Schneider, T. R., Hipp, J. F., Engel, A. K., and Senkowski, D. (2013). Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. *NeuroImage*, 70:101–112.

Schiller, J., Major, G., Koester, H. J., and Schiller, Y. (2000). Nmda spikes in basal dendrites. *Nature*, 1261(1997):285–289.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007.

Self, M. W., van Kerkoerle, T., Goebel, R., and Roelfsema, P. R. (2017). Benchmarking laminar fmri: Neuronal spiking and synaptic activity during top-down and bottom-up processing in the different layers of cortex. *NeuroImage*, (March):1–12.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:618–626.

Sherstinsky, A. (2018). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *arXiv*, 404(March):1–43.

Simion, F., Regolin, L., and Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences of the United States of America*, 105(2):809–813.

Simon, D. M. and Wallace, M. T. (2017). Rhythmic modulation of entrained auditory oscillations by visual inputs. *Brain Topography*, 30(5):565–578.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., Hill, C., and Arbor, A. (2014). Going deeper with convolutions. pages 1–9.

Teller, D. Y. (1984). Linking propositions. *Vision Research*, 24(10):1233–1246.

Tolooshams, B., Song, A. H., Temereanca, S., and Ba, D. (2019). Convolutional dictionary learning based auto-encoders for natural exponential-family distributions.

Tort, A. B., Kramer, M. A., Thorn, C., Gibson, D. J., Kubota, Y., Graybiel, A. M., and Kopell, N. J. (2008). Dynamic cross-frequency couplings of local field potential oscillations in rat striatum and hippocampus during performance of a t-maze task. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20517–20522.

Tovar, D., Murray, M., and Wallace, M. (2020). Selective enhancement of object representations through multisensory integration. *Journal of Neuroscience*, 40(29):5604–5615.

Traub, R. D. and Bibbig, A. (2000). A model of high-frequency ripples in the hippocampus based on synaptic coupling plus axon-axon gap junctions between pyramidal neurons. *Journal of Neuroscience*, 20(6):2086–2093.

van den Hurk, J., Van Baelen, M., and Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22):E4501–E4510.

Van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M. A., Poort, J., Van Der Togt, C., and Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(40):14332–14341.

Vetter, P., Smith, F. W., and Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Current Biology*, 24(11):1256–1262.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., and Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences of the United States of America*, 115(44):11333–11338.

Vogler, J. N. and Titchener, K. (2011). Cross-modal conflicts in object recognition: Determining the influence of object category. *Experimental Brain Research*, 214(4):597–605.

von Nicolai, C., Engler, G., Sharott, A., Engel, A. K., Moll, C. K., and Siegel, M. (2014). Corticostriatal coordination through coherent phase-amplitude coupling. *Journal of Neuroscience*, 34(17):5938–5948.

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology*, 80(2):1006–1010.

Wang, Y., Zhang, J., Zou, J., Luo, H., and Ding, N. (2019). Prior knowledge guides speech segregation in human auditory cortex. *Cerebral Cortex*, 29(4):1561–1571.

Warrington, E. K. and Mccarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, 110(5):1273–1296.

Xing, D., Yeh, C. I., and Shapley, R. M. (2009). Spatial spread of the local field potential and its laminar variation in visual cortex. *Journal of Neuroscience*, 29(37):11540–11549.

Xu, Y. and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 12(2065):1–16.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yuval-Greenberg, S. and Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research*, 193(4):603–614.

Zanos, T. P., Mineault, P. J., Nasiotis, K. T., Guitton, D., and Pack, C. C. (2015). A sensorimotor role for traveling waves in primate visual cortex. *Neuron*, 85(3):615–627.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1):818–833.

Zeitler, M., Fries, P., and Gielen, S. (2006). Assessing neuronal coherence with single-unit, multi-unit, and local field potentials. *Neural Computation*, 18(9):2256–2281.

Zhou, Z. and Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, 10(1).

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, 118(3).

Part 1

"You're nothing but a pack of neurons"

-Francis Crick

Chapter 2

Stimulus Feature-Specific Information Flow Along the Columnar Cortical Microcircuit Revealed by Multivariate Laminar Spiking Analysis

The contents of this chapter are adapted from

Tovar, D. A., Westerberg, J. A., Cox, M. A., Dougherty, K., Carlson, T. A., Wallace, M. T., & Maier, A. (2020). Stimulus Feature-Specific Information Flow Along the Columnar Cortical Microcircuit Revealed by Multivariate Laminar Spiking Analysis. Frontiers in Systems Neuroscience, 14, 1–14.

2.1 Abstract

Most of the mammalian neocortex is comprised of a highly similar anatomical structure, consisting of a granular cell layer between superficial and deep layers. Even so, different cortical areas process different information. Taken together, this suggests that cortex features a canonical functional microcircuit that supports region-specific information processing. For example, the primate primary visual cortex (V1) combines the two eyes' signals, extracts stimulus orientation, and integrates contextual information such as visual stimulation history. These processes co-occur during the same laminar stimulation sequence that is triggered by the onset of visual stimuli. Yet, we still know little regarding the laminar processing differences that are specific to each of these types of stimulus information. Univariate analysis techniques have provided great insight by examining one electrode at a time or by studying average responses across multiple electrodes. Here we focus on multivariate statistics to examine response patterns across electrodes instead. Specifically, we applied multivariate pattern analysis (MVPA) to linear multielectrode array recordings of laminar spiking responses to decode information regarding the eye-of-origin, stimulus orientation, and stimulus repetition. MVPA differs from conventional univariate approaches in that it examines patterns of neural activity across simultaneously recorded electrode sites. We

were curious whether this added dimensionality could reveal neural processes on the population level that are challenging to detect when measuring brain activity without the context of neighboring recording sites. We found that eye-of-origin information was decodable for the entire duration of stimulus presentation, but diminished in the deepest layers of V1. Conversely, orientation information was transient and equally pronounced along all layers. More importantly, using time-resolved MVPA, we were able to evaluate laminar response properties beyond those yielded by univariate analyses. Specifically, we performed a time generalization analysis by training a classifier at one point of the neural response and testing its performance throughout the remaining period of stimulation. Using this technique, we demonstrate repeating (reverberating) patterns of neural activity that have not previously been observed using standard univariate approaches.

2.2 Introduction

Certain anatomical motifs are repeated across disparate brain areas with wide-ranging functions. The mammalian neocortex is one such example as it predominantly features the same laminar structure. A popular model for cortical function resting upon this stereotypical structure is the canonical cortical microcircuit (CCM: Douglas et al., 1989; Douglas and Martin, 1991; Bastos et al., 2012). The CCM gives rise to a series of distinct, yet overlapping, activation steps that are spatially segregated between the superficial (supragranular), deep (infragranular), and middle (granular) layers of cortex (Rockland and Pandya, 1979; Rockland and Virga, 1989; Callaway, 1998; Binzegger et al., 2004; Douglas and Martin, 2004). According to this model, ascending (feedforward) signals from parts of the brain that are closer to the sensory periphery terminate in the middle layers of cortical areas while descending (feedback) signals from downstream areas target the layers above and below (Rockland and Pandya, 1979; Rockland and Virga, 1989; Felleman and Van Essen, 1991, but see Self et al., 2013).

Since the CCM applies virtually ubiquitously across neocortex, an improved under-

standing of the laminar cortical processing chain is bound to translate into an improved understanding of cortical processing more generally (Hubel and Wiesel, 1977; Douglas et al., 1989; Felleman and Van Essen, 1991; Douglas and Martin, 2004; Bastos et al., 2012). Our knowledge of laminar neural activity in primates has grown greatly over the last decade thanks to the prevalence of linear electrode arrays (Schroeder et al., 1998; Xing et al., 2009, 2012; Burns et al., 2010; Buffalo et al., 2011; Kajikawa and Schroeder, 2011; Maier et al., 2011, 2014; Hansen et al., 2012; Spaak et al., 2012; Smith et al., 2013; Bastos et al., 2014, 2018; Van Kerkoerle et al., 2014; Nandy et al., 2017; Cox et al., 2019a,b; Westerberg et al., 2019; Dougherty et al., 2019a; Gieselmann and Thiele, 2020). Yet, our knowledge about laminar neuronal activation remains limited (e.g., Mignard and Malpeli, 1991). Recent studies demonstrated that—matching predictions by the CCM—there are two distinct sequences of laminar activation for feedforward and feedback activation, respectively (Maier, 2013; Van Kerkoerle et al., 2014, 2017; Cox et al., 2019a). Much less is known about the different types of feedforward processes that occur along cortical layers. Specifically, we still know little about how one and the same feedforward sweep of neural activation across cortical layers entails multiple streams of stimulus-specific information that manifest differently across space and time.

Our knowledge regarding laminar cortical processing is bound to rapidly increase since there have been notable advances in microelectrode technology. Specifically, the increase in simultaneously placed electrodes and the associated increase dimensionality of laminar neurophysiological data obtained by second generation laminar arrays is rapidly approaching those of other techniques such as fMRI (Jun et al., 2017; Steinmetz et al., 2018; Musk and Neuralink, 2019). Yet, laminar recordings are usually analyzed using the same univariate techniques that have been established for single electrodes, rather than utilizing the additional, contextual information provided by neighboring electrode contacts in a multivariate fashion.

There are several statistical approaches that quantify information distributed across

neighboring measurements in the brain, directly capturing neuronal interactions on the population level. Specifically, machine-learning based multivariate pattern classification analysis (MVPA) has proven fruitful in systems neuroscience (Haxby et al., 2001; Kriegeskorte and Bandettini, 2007; Kriegeskorte et al., 2008; Kriegeskorte and Kreiman, 2012; Rutishauser et al., 2018; Kamitani & Tong, 2005). More recently, time-resolved MVPA has emerged as a powerful technique to study the time courses with which information processing occurs across the brain (Carlson et al., 2013; Cichy and Pantazis, 2017; Tovar et al., 2020). While time-resolved MVPA has been applied to multielectrode recordings (Goddard et al., 2017), to date no study to our knowledge probed whether this technique can reveal aspects of laminar cortical activation that are opaque to univariate analyses. For instance, through time generalization, which is achieved by training a classifier at a specific time point—such as early in the neuronal response to a stimulus—then testing it throughout the remainder of the response, one can search for repeating patterns of neural activity across electrodes that might be invisible when analyzing single channels in isolation.

Here we use time-resolved MVPA to analyze the pattern of spiking activity across 24 and 32 channel (first generation) linear multielectrode array recordings in primate primary visual cortex (V1). Instead of relying on the average response across all electrode channels or only examining one channel at a time, MVPA uses patterns of activity across neighboring channels to classify neuronal responses. We use both time-resolved MVPA and an MVPA-based "searchlight" analysis commonly used for neuroimaging data to map how information regarding stimulus orientation, eye-of-origin, and stimulus history differentially flows within the laminar activation sequence of V1. We found that MVPA can be utilized effectively despite the relatively low channel counts of first generation laminar linear arrays. We then explored time-generalization, as this analysis provides insight that cannot be gained from more conventional, univariate approaches that are blind to patterns of activity that span multiple electrodes. This analysis revealed repeating patterns in neuronal activity that entailed information about whether a stimulus had previously been shown or

not, which we had not observed in a prior study that had relied on univariate analyses exclusively (Westerberg et al., 2019). We discuss these findings and their implications for the advent of massively increased channel counts for linear multielectrode arrays that are rapidly gaining prominence (Jun et al., 2017; Steinmetz et al., 2018; Musk and Neuralink, 2019).

2.3 Materials and Methods

2.3.1 Animal Care and Surgical Procedures

Data were collected from two macaque monkeys [Macaca radiata, one female (designated Monkey 1) and one male (designated Monkey 2)]. All procedures were in compliance with regulations set forth by the Association for the Assessment and Accreditation of Laboratory Animal Care (AALAC), approved by the Vanderbilt University Institutional Animal Care and Use Committee, and followed National Institutes of Health guidelines. A detailed description of the surgical procedures can be found in previous publications (Westerberg et al., 2019, 2020a,b). Briefly, in a series of surgeries, each monkey was implanted with a custom MRI-compatible headholder and recording chamber over perifoveal V1 concurrent with a craniotomy.

2.3.2 Behavioral Paradigm

In each recording session, monkeys viewed a 20" CRT monitor (Diamond Plus 2020u, Mitsubishi Electric Inc.) operating at 60 or 85 Hz. Monkeys passively fixated within a one-degree radius around a central fixation dot and viewed stimuli through a custom mirror stereoscope so that stimuli could be viewed monocularly or binocularly (Figure 2.1A). To eliminate potential response differences due to binocular disparity, prior to the main tasks, a mirror calibration task was performed. In this task, monkeys shifted gaze to a series of stimuli positioned across the visual display and held fixation at each position to receive fluid reward. Each stimulus was presented to only one eye at a time. This resulted in two maps of fixation positions, one for the set of stimuli presented to each eye. The stereoscope

was then adjusted if differences were observed in those maps (e.g., the maps were not completely overlapping). Stimuli were generated using MonkeyLogic (Asaad et al., 2013; Hwang et al., 2019) via MATLAB (R2012, R2014a, The Mathworks, Inc.) running on a computer using a Nvidia graphics card. Following 300 ms of fixation, monkeys viewed five sequentially presented stimuli for 200 ms each, with a 200 ms inter-stimulus interval (ISI). If fixation was maintained throughout the five presentations, the monkey was rewarded with juice and relieved of the fixation constraint for an inter-trial interval (ITI). If the monkey broke fixation during trial performance, the presentation was eliminated from analysis and the monkey experienced a short timeout (1–5 s) before starting the next trial. Each stimulus in the presentation sequence was a sinusoidal bar grating of equivalent size, spatial frequency, and phase, with variable orientation and eye-of-origin (Figure 2.1B). For each recording session, the stimuli were optimized for the measured neural activity evaluated by listening to the multi-unit activity (MUA) during exposure to a wide variety of stimuli. We selected stimulus parameters that evoked the greatest neural response. For a more detailed description of the paradigm, as well as further information on stimulus optimization and receptive field mapping (Supplementary Figure 2.6), see previous publications (Cox et al., 2013, 2019a,b; Dougherty et al., 2019a; Westerberg et al., 2019).

2.3.3 Neurophysiological Procedure

All data used in this paper are available upon request from the communicating author, pending approval by Vanderbilt University. During task performance, broadband (0.5 Hz12.207 kHz) intracranial voltage measurements were taken at a sampling rate of 30 kHz and amplified, filtered, digitized using a 128-channel Cerebus™ Neural Signal Processing System (NSP, Blackrock Microsystems LLC). Neuronal data was downsampled offline to 1 kHz, following low-pass filtering with an anti-aliasing filter. Gaze position was recorded at 1 kHz (NIDAQ PCI-6229, National Instruments) using an infrared light sensitive camera and commercially available eye tracking software (Eye Link II, SR Research

Ltd.; iView, SensoMotoric Instruments). Recordings took place inside an electromagnetic radio frequency-shielded booth and were performed using one or two acute laminar multielectrode arrays with 24 or 32 contacts with 0.1 mm electrode spacing and impedances ranging between 0.2 and 0.8 megaohms at 1 kHz (U-Probe, Plexon, Inc.; Vector Array™, NeuroNexus). Electrodes were connected to the NSP using analog headstages. In each recording, the electrode array(s) were introduced into dorsal V1 through the intact dura mater using a chamber-mounted microdrive (custom modification of a Narishige International Inc. Micromanipulator) and adjusted such that the majority of recording contacts spanned the cortical sheet. This procedure was repeated across the 61 experimental sessions (n = 13 for monkey I34).

2.3.4 Receptive Field Mapping

Since achieving single-unit isolation on every channel is difficult, we instead opted to estimate the local population spiking response by quantifying the time-varying activity in the spiking frequency range (multi-unit activity, MUA) as we wanted to ensure overlapping receptive fields along the cortical depth. Verifying overlapping receptive fields provides confidence that the activity we are recording across columns originates from the same cortical location rather than spanning adjacent columns (i.e., that the electrode penetration was orthogonal to cortex). Monkeys performed a visual fixation task where a visual stimulus was presented repeatedly in the contralateral visual hemifield – relative to the position of the electrode array. Up to five stimuli were presented on each trial for 200 ms with a 200 ms interstimulus interval. Stimulus size and positioned varied between recording sessions, but each session usually consisted of a "coarse" receptive field mapping task followed by a more focused version once an estimation for the exact position was found. We mapped receptive fields using a reverse-correlation technique (Supplementary Figure 2.6) which resulted in 3-dimensional receptive field matrices where 2 dimensions corresponded to visual space and the third, response magnitude (Cox et al., 2013). Only sessions where the recep-

tive field matrices were overlapping along cortical depth were included for further analysis. Additionally, this procedure determined the position where the stimulus was positioned to stimulate the column receptive field for the main task (see section Behavioral Paradigm).

2.3.5 Laminar Alignment

Current source density (CSD) in response to brief visual stimulation was used to find the boundary between the granular and infragranular compartments of V1 as per previously documented methods (Schroeder et al., 1998; Maier et al., 2010; Maier, 2013; Ninomiya et al., 2015; Cox et al., 2019a,b; Dougherty et al., 2019a; Westerberg et al., 2019). Only sessions that were found to be perpendicular to the cortical surface were included in analysis (see section Receptive Field Mapping). Additional neurophysiological criteria were used, such as well-defined patterns of LFP power spectral density (Van Kerkoerle et al., 2014; Bastos et al., 2018; Westerberg et al., 2019), signal correlations between LFP recorded on differing channels (Westerberg et al., 2019), and latency (Self et al., 2013) of stimulus-evoked MUA. The granular to supragranular boundary was set to 0.5 mm above the granular to infragranular boundary (Figure 2.1C). Supplementary Figure 2.7 demonstrates the reliability of these functional markers following alignment of all sessions. Both extracranial to intracranial and gray matter to white matter boundaries were determined by finding the pair of recording electrodes where no multiunit response to visual stimuli was observed on one channel and a significant response was observed on the other (Cox et al., 2019b; Westerberg et al., 2019). Recording channels positioned between these pairs all showed significant responses. That is, we found no instances of a lack of response on a channel determined to be within the gray matter. The L2/3–L4 boundary was set to 0.5 mm above the L4–L5 boundary as we do not have a reliable functional marker and that distance is consistent with histological studies of V1 laminar structure (see Cox et al., 2019b; Westerberg et al., 2019 for details).

2.3.6 Data Preprocessing

All contiguous recording channels found to be within the gray matter were taken and multiunit signals were computed. Channels in the gray matter were found by determining first whether a visual response could be evoked on the channel and second, whether a receptive field was present for the multiunit and/or LFP activity through a previously described receptive field mapping paradigm (Westerberg et al., 2019). If the channel was found to be in the gray matter, the broadband neural signal recorded at that channel was then band-pass filtered between 500 and 5,000 Hz, rectified, and low-pass filtered at 200 Hz using Butterworth filters (Self et al., 2013; Shapcott et al., 2016; Westerberg et al., 2020a). These derived neural signals, with no further filtering of the multiunit activity, were then used in performing both the univariate and multivariate analyses (Figure 2.1D).

2.3.7 Multivariate Pattern Analysis

To track how sensory information from different stimulus features are processed within this laminar microcircuit, we applied multivariate pattern analysis (MVPA) using CoSMoMVPA (Oosterhof et al., 2016) to the MUA of each of the three laminar compartments (Figure 2.1E, left-most panel). To do so, we assembled two-dimensional neuronal response matrices (NRMs) that contained the millisecond-by-millisecond population spiking response at each electrode channel as a function of trials. Each row/electrode in the NRM can be thought of as a separate axis forming a multidimensional space whose dimensionality is determined by the number of electrodes. Each stimulus presentation will elicit a different response across each of the dimensions. The specific stimulus features we tested comprised of grating orientation, the eye that the stimuli were presented to (eye-of-origin) and the relative position of each stimulus within the stimulation sequence (Figure 2.1F). We next randomly divided trials within sessions to perform a 4-fold cross-validation procedure. In this procedure, 3/4 of the data is used to train an MVPA classifier (Figure 2.1E, second-to-left panel). The remaining 1/4 of the NRMs are used to determine classifier per-

formance. To classify a given stimulus feature, a different hyperplane or set of hyperplanes (as is the case with the orientation where we have four orientations) is used to distinguish stimulus feature on a trial by trial basis. The decoding accuracy is the number of trials over the total number of trials that classifier is able to correctly identify for each session. We performed this computation separately within each recording session on a millisecond-by-millisecond basis, evaluating the accuracy of classifier performance as a function of time (Figure 2.1E, second-to-rightmost panel). The resulting time courses of decoding accuracy for each laminar compartment were then pooled together and compared to a randomized trial shuffle control to determine statistical significance (Figure 2.1E, rightmost panel). To correct for multiple comparisons, we used the false discovery rate (FDR) adjusted p-values with = 0.01. For each of the decoding distinctions, the subsets were balanced, such that both training subsets and testing subsets contained the same number of trials for each stimulus category.

For orientation decoding, all recording sessions were used for analysis. However, some recording sessions included orientation presentations that were not shown in other recording sessions (i.e., 22.5° in one recording session and 30° in another sessions). Therefore, orientation presentations were binned into four categories: 0–44°, 45–89°, 90–134°, and 135–179°. For trial repetition decoding, the five stimuli presentations for a given trial were grouped as either the first presentation or as a repetition. To have an equal number of first presentations and repetitions, we randomly subsampled from the repetitions to match the number of first presentations.

For each stimulus feature, we also performed a time generalization analysis (Carlson et al., 2011; King and Dehaene, 2014) which uses a similar decoding procedure described, with one notable exception — the classifier is trained on the information at one time point for each stimulus feature and the model is subsequently tested on all timepoints. This procedure is repeated across all timepoints resulting in a 2D "time generalization matrix" that plots training time against decoding time to gain insight into how information at specific

timepoints evolve throughout the time course. Lastly, to determine the effects of repeated stimuli presentations on orientation and eye of origin decoding, we further divided the repetition subset of data into balanced eye of origin subsets and balanced orientation subsets. We then again performed a 4-fold classification using a linear discriminant analysis classifier.

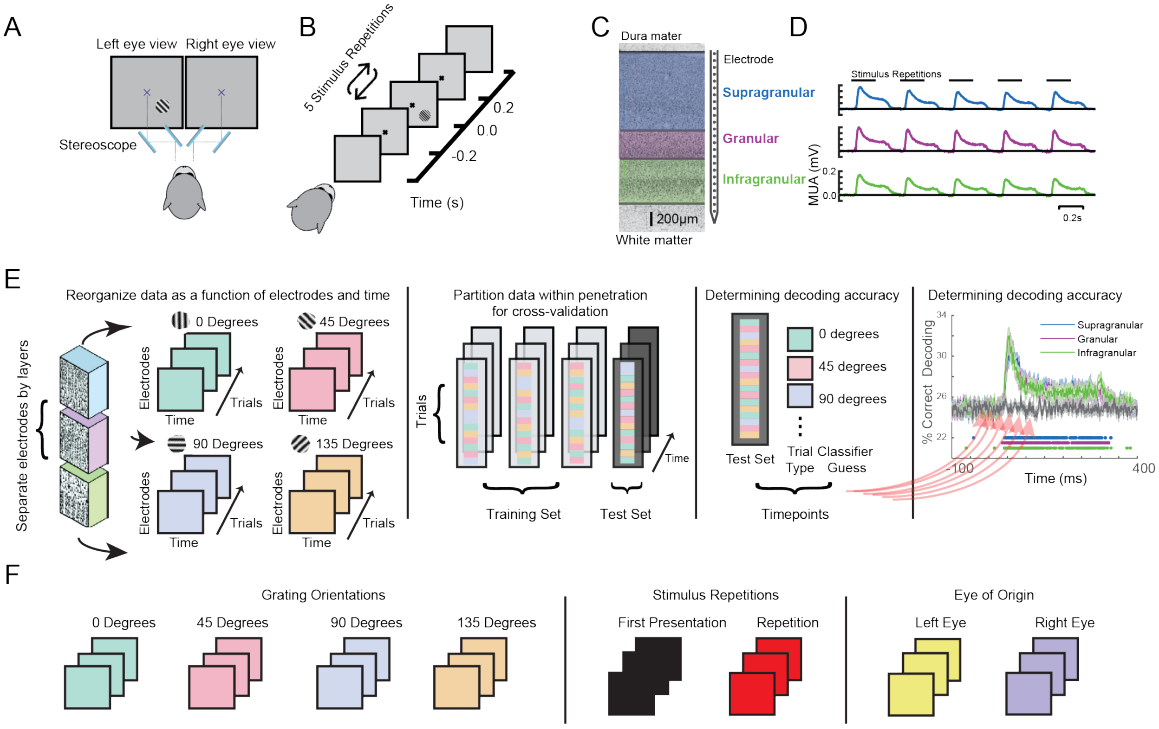

Figure 2.1: Experimental setup, paradigm, preprocessing, and analysis. (A) Monkeys were positioned in front of a monitor and tasked to passively fixate a central dot through a custom mirror stereoscope. (B) Monkeys were shown a series of five grating stimuli of randomly varying orientations and ocular configuration with all other parameters were held constant. (C) Linear multicontact array recording laminar neuronal responses at 100 micron spatial resolution spanning through visual cortex. (D) Grand average multiunit spiking responses (MUA) to the stimulus sequence for all three main laminar compartments (both animals, all sessions). (E) Schematic of multivariate pattern analysis (MVPA). Population spiking responses (MUA) from each laminar compartment were reorganized as a function of electrode contact and time. A classifier was trained at each timepoint using linear discriminant analysis and 4-fold cross validation. (F) Decoding analysis was separately performed for grating orientations, stimulus history (initial stimulus vs. repetitions), and eye-of-origin.

2.4 Results

2.4.1 Stimulus specific information within neural activation of the CCM

Before investigating each stimulus feature in isolation, we evaluated whether the grand average spiking response to our stimuli matched predictions from the CCM (Figure 2.2A). To do so, we spatially aligned the spiking data from each recording session to the layer 4C/5 boundary. Using these aligned datasets, we computed the grand average spiking response to all stimuli as a function of cortical depth and time (Figure 2.2B). The resulting laminar profile of activation was consistent with both the expectations set by the CCM and previous studies of laminar visual activation in that layer 4C activity preceded that of the other layers (Mitzdorf, 1985; Schroeder et al., 1998; Maier et al., 2010; Spaak et al., 2012; Van Kerkoerle et al., 2014). Interestingly, however, both the supragranular and infragranular layers responded virtually simultaneously, which might either be explained by (i) V1's idiosyncratic laminar connections [i.e., there are also, less pronounced, geniculate projections outside layer 4C (Callaway, 1998)], (ii) limitations of the CCM model itself (e.g., Godlove et al., 2014; Ninomiya et al., 2015), or both. This pattern of sensory activation occurs regardless of stimulus feature, raising the question of how stimulus-specific information is extracted within this activation sequence. To answer this question, we applied MVPA using a "moving searchlight" analysis (Etzel et al., 2013). Specifically, we limited both our training and test data sets to three neighboring electrode channels, performed MVPA over time, and then repeated the process after moving this "searchlight" 0.1 mm deeper along the electrode array. In this analysis a classifier is trained and tested for each timepoint of the response, in 1 ms increments (Figure 2.2C). No spatial or temporal smoothing were added.

We first focused on the eye-of-origin for each stimulus presentation. While V1 harbors both neurons that respond to one or both eyes, most of the neurons that respond to one eye only (monocular neurons) are located in the middle, granular layers (Hubel and Wiesel, 1977; Dougherty et al., 2019a). This finding is consistent with neuroanatomy, as the gran-

ular layers receive the bulk of (monocular, eye-specific) inputs from the lateral geniculate nucleus of the thalamus (LGN) that connects eye and cortex (Casagrande and Boyd, 1996). A long-standing hypothesis is that the eye-specific inputs in the middle layers are merged to a combined (binocular) response in the layers above, even though most V1 neurons maintain preference for one eye over the other (Hubel and Wiesel, 1972; Ohzawa and Freeman, 1986; Prince et al., 2002; Read and Cumming, 2004). Neurons in the uppermost layers of V1 project to neurons in V1's lower layers, so if the upper layers form a combined binocular signal, this signal should be present in the lower layers as well (Hubel and Wiesel, 1972; Cox et al., 2019b; Dougherty et al., 2019a). However, based on several other pieces of empirical evidence, an alternative hypothesis postulates that the two eyes' signals are interacting at or before LGN responses arrive in the middle layers of V1 (see Dougherty et al., 2019b for review).

Using MVPA, we found information regarding eye-of-origin initially followed the CCM profile of general activation, with neurons reliably indicating whether a stimulus was shown to left or right eye in the middle layers, followed by the upper layers of V1. This eye-specific information largely diminished once neuronal activation reached the lower layers of V1 (Figure 2.2C, left panel). These timing differences can clearly be seen for a layer-specific MVPA using all electrode channels within the middle, upper and lower layers of V1, respectively (Figure 2.2D). We utilized this analysis to perform several statistical comparisons. First, we compared decoding performance on a millisecond-by-millisecond basis against a randomized trial shuffle control. Second, we compared decoding across laminar compartments. Decoding of eye-of-origin first emerged in the middle layers (29 ms), followed by the upper (40 ms) and lower layers (40 ms). Decoding which eye the stimuli were shown to was comparable between middle and upper layers but significantly reduced in the lower layers, suggesting that eye-specific information is largely preserved when granular neurons project to neurons in the layers above. However, decoding of eye-of-origin is relatively poor in the lower layers of V1, suggesting that, at least on the multiunit-level, there is

significant binocular convergence after activation reaches the upper layers of cortex. This finding demonstrates that eye-of-origin is more robustly represented in supragranular compared to infragranular layers.

Next, we computed the laminar evolution of stimulus orientation information. A common notion regarding the functional layout of V1 states that orientation selectivity (tuning) is less pronounced in the middle layers of V1 (Hubel and Wiesel, 1972, 1977; Ringach et al., 2002). Several authors have since challenged this idea, arguing that V1 already receives orientation-biased inputs (Daniels et al., 1977; Vidyasagar and Urbas, 1982; Leventhal and Schall, 1983; Smith et al., 1990; Pugh et al., 2000; Xu et al., 2002). We thus wondered what the laminar profile of MVPA-based decoding of stimulus orientation across V1 layers might be.

We binned our grating stimuli into four groups (0°, 45°, 90°, and 135°, respectively) and trained a classifier to discriminate between them (Figure 2.2C). Interestingly, we found that information regarding stimulus orientation was more transient than information regarding of eye-of-origin. Moreover, the laminar profile was strikingly different: the center of the granular layers discriminated relatively poorly between gratings of varying orientation, and neurons in the layers above and below did so without any significant temporal delay. Closer inspection of the layer-resolved decoding (Figure 2.2D), collapsed across time, revealed that there was no significant difference between any of the laminar compartments (bar plots). These results seem to suggest that stimulus orientation information is extracted almost uniformly across V1 layers. However, visual inspection reveals clear differentiation within the middle layers, which is lost when collapsing this layer into a single measure. This heterogeneous pattern within the granular layers might at least be partially explained by the fact that the middle layers host several sublayers that each receive separate inputs from the LGN (Casagrande and Boyd, 1996), although it is not immediately clear how the granular sublayers relate to the specific pattern we found.

Given that V1 is known to modulate its responses depending on contextual cues, such

as the behavioral state of the animal or stimulus history (Van Kerkoerle et al., 2014; Cox et al., 2019a; Westerberg et al., 2019), we next examined how stimulus history affects the laminar flow of stimulus-specific information. To do so, we first studied the laminar flow of information of whether a stimulus was novel or preceded by another stimulus in the stimulation sequence. We found that this information regarding stimulus history yielded yet another pattern of laminar information flow (Figure 2.2C). We found that the bulk of information regarding stimulus history resided outside the granular input layers. This finding was also apparent in layer-specific MVPA (Figure 2.2D). These results are in line with earlier work showing that V1 granular layers are least affected by the adaptive effects of repeated visual stimulation (Westerberg et al., 2019).

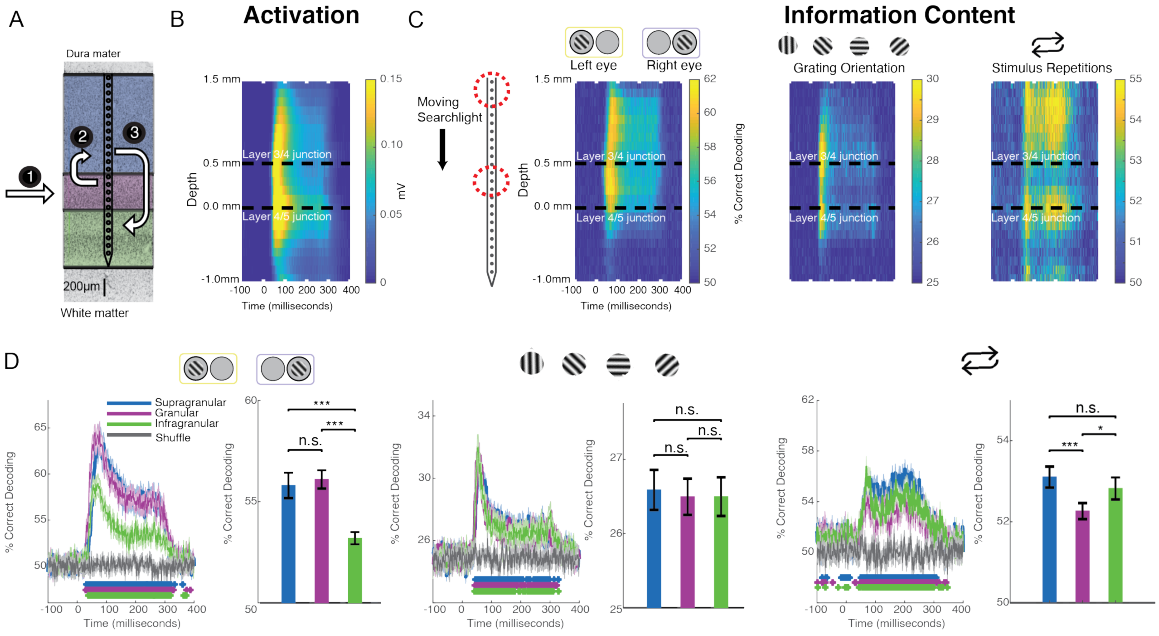

Figure 2.2: Stimulus feature-specific information within neural activation of the CCM. (A) Canonical microcircuit model (CCM) of neural activation in V1. Feedforward activation initially excites the middle layers before reaching upper and lower layers of cortex. (B) Grand average laminar MUA profile to all stimulus presentations along the depth of the electrode (all sessions, both monkeys). (C) Decoding performance using a "moving searchlight" along the electrode array for eye of origin (leftmost panel), grating orientation (middle panel), and stimulus repetition (rightmost panel). (D) Time series of MVPA decoding for eye of origin (leftmost panel), grating orientation (middle panel), and stimulus repetitions (rightmost panel). Graphs show decoding accuracy as a function of time and laminar compartment, together with a randomized shuffled control as a baseline. Significance is indicated with colored asterisks above the abscissa using Wilcoxon signed-rank test, FDR corrected, $q < 0.01$. Bar plots to the right indicate time-averaged statistics of the data with Wilcoxon signed-rank test P values (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) above the plots.

2.4.2 Quantifying Differences Between Spatiotemporal Searchlight Maps.

We next quantified the visual difference we observed between the spatiotemporal maps for the stimulus-specific information (Figure 2.3). Since we were primarily interested in relative decoding performance throughout the cortical columns, we normalized each channel (electrode contact) by subtracting mean decoding performance across channels for each individual timepoint in the time series for each stimulus feature. We then calculated the Euclidean distance between each of our stimulus feature at each timepoint. These results were then compared to a shuffled label control where we similarly normalized our electrodes at each timepoint and then calculated the Euclidean distance (Figure 2.3B). Here, we find that the spatiotemporal differences between eye of origin, orientation, and stimulus history are all higher than the differences found in the respective shuffled label control. Eye-of-origin, which was more readily decoded in the granular layers was distinct from the decoding of stimulus orientation and repetition, which both lead to higher decoding in superficial and deeper layers. To statistically compare the differences across space and time, we next converted the searchlight matrices into one-dimensional vectors and then normalized across channels before conducting a pairwise signed rank test. Using this approach, we found significant decoding differences between eye of origin and orientation ($p < 0.001$), eye-of-origin ($p < 0.001$) and repetition ($p < 0.001$), and orientation and stimulus history ($p < 0.001$). As expected, there were no significant differences between the shuffled label controls. These decoding differences between stimulus features indicate that processing these stimulus features occurs distinctly but simultaneously with the laminar microcircuit.

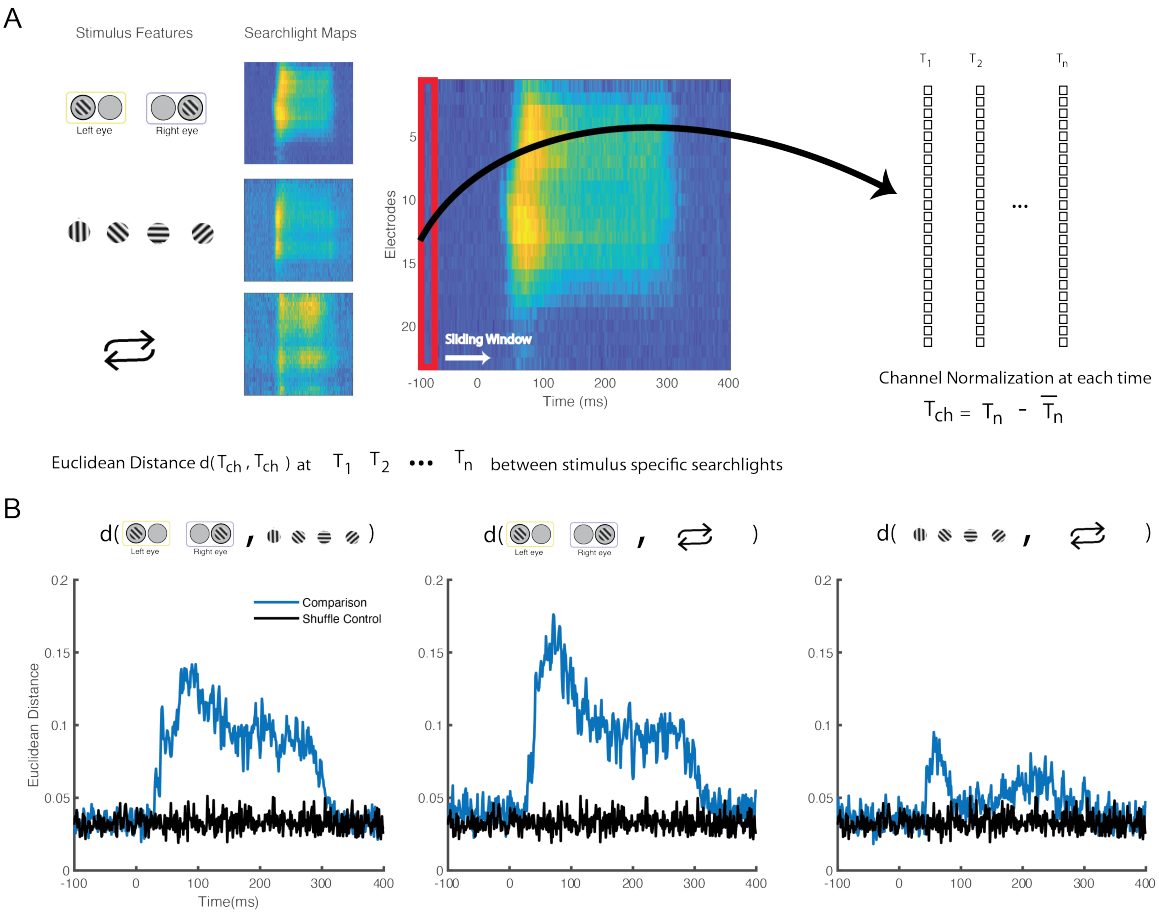

Figure 2.3: Statistical comparison of columnar flow of stimulus feature-specific information. (A) Schematic for comparison between stimulus-feature specific searchlight analyses. Decoding results from the searchlight analyses for each of the stimulus features, normalized across all the channels for each individual timepoint from 100 ms prior to stimulus presentation to 400 ms after stimulus presentation (B) Euclidean distance of the normalized decoding values calculated between each stimulus feature. A shuffled control where stimulus labels have been shuffled prior to channel normalization and Euclidean distance calculation is shown for comparison.

2.4.3 Temporal Dynamics of Stimulus Information Using Time Generalization

To further investigate how feature information evolves over time (see also: Ringach et al., 1997, 2002, 2003; Bair et al., 2002; Smith et al., 2006; Shapley et al., 2007), we decoded neuronal data based on a classifier that was trained for another time period of the same neuronal response ("time generalization") (Carlson et al., 2011; King and Dehaene, 2014). The result of this analysis is a 2D "time generalization matrix" that plots training time against decoding time. Figure 2.4A illustrates several possible outcomes for generalization matrices. It is possible, for example, that there is little to no generalization between a classifier trained at one time and tested on the remaining time of a neuronal response. In other words, spiking might be constantly changing in a way that any information used to discriminate between stimuli is specific to each individual point in time of the neuronal response ("unique states"). In contrast, if the information used to discriminate between stimuli were static across the neuronal response, we would expect a square-like pattern ("sustained"). This analysis can also show information decaying over time ("information decay"). An asymmetric pattern occurs because a classifier trained on lower signal-to-noise ratio (SNR) data generalizes better to higher SNR data than the converse (van den Hurk and Op de Beeck, 2019). Lastly, information might reoccur at a later time point of a response ("recurrence").

We performed time generalization analysis for the decoding of eye-of-origin, stimulus orientation as well as stimulation history within each laminar compartment (Figure 2.3 and Supplementary Figure 2.7). Decoding eye-of-origin was mostly sustained but also exhibited some information decay within each laminar compartment (Figure 2.4). Decoding of stimulus orientation, in contrast, was less sustained. Interestingly, whether or not a stimulus preceded or succeeded other stimuli showed a very different pattern. Specifically, the time generalization matrix was suggestive of recurrent processing, in that the initial information emerges, weakens and then re-emerges at a later time point. This reactivation pattern was most prominent in the supragranular and infragranular layers (Figure 2.4).

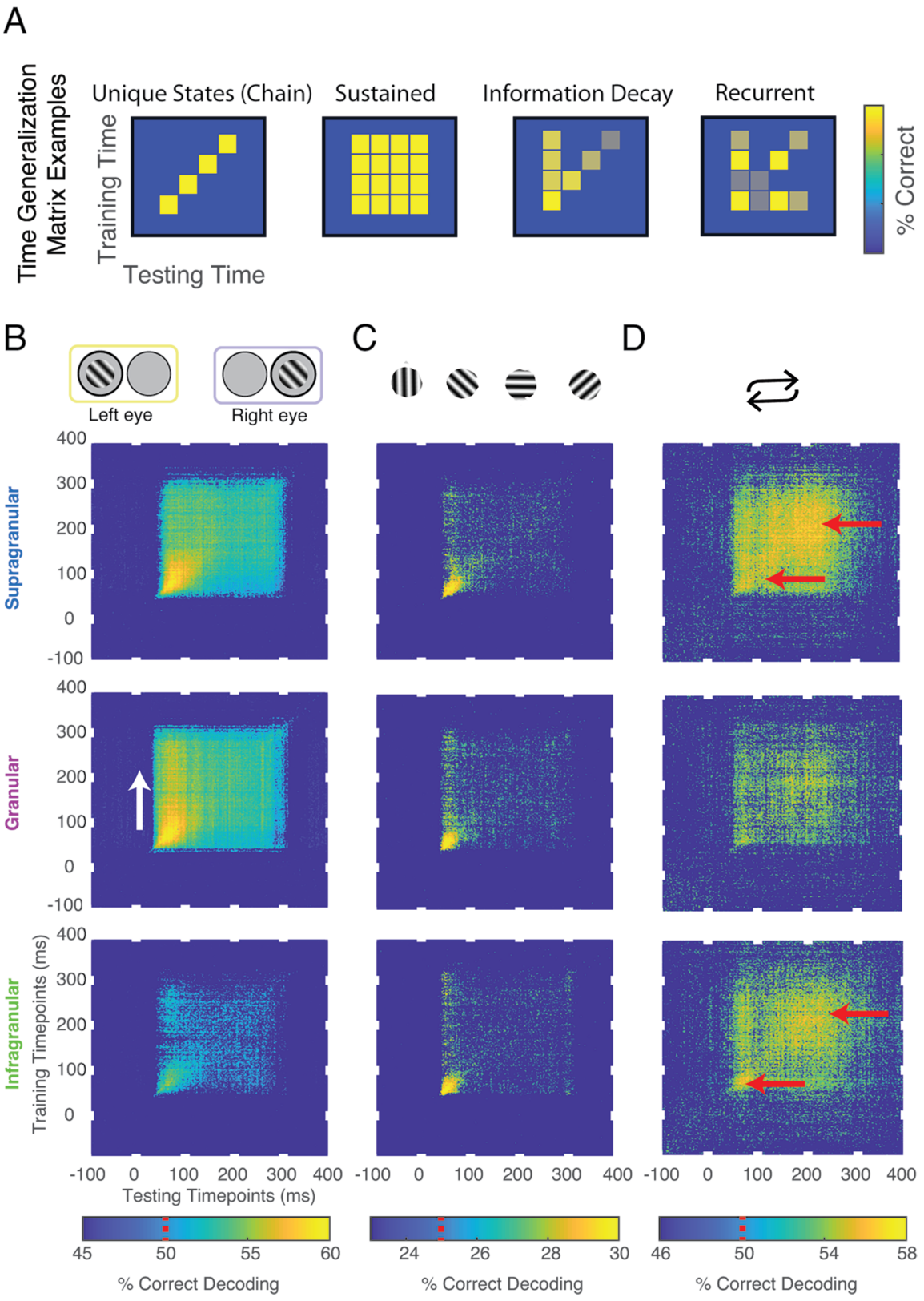

Figure 2.4: Temporal dynamics of stimulus information using time generalization. (A) Cartoon models of possible results. (B) Significant time generalization results, FDR corrected for multiple comparisons, $q < 0.025$, for: (B) Eye-of-origin, (C) Orientation, (D) Stimulus repetitions (see Methods for details). Chance decoding level is indicated on each color bar by a red line. Red and white arrows are added for emphasis.

To further investigate how the temporal dynamics for each of the stimulus features varies within compartments. We combined the searchlight and time generalization analyses (Figure 2.5 and Supplementary Video 2.8). Using this approach, we found that the electrode-specific time generalization matrices were generally representative of their respective compartments. However, within compartments there was notable heterogeneity. For example, for eye of origin decoding, time generalization was comparable across contiguous electrodes. In contrast, for decoding stimulus history (repetition), the reactivation pattern noted in Figure 2.5 waxes and wanes even within laminar compartments. These results provide evidence for the notion that sub-layers within laminar compartments differentially process distinct stimulus features.

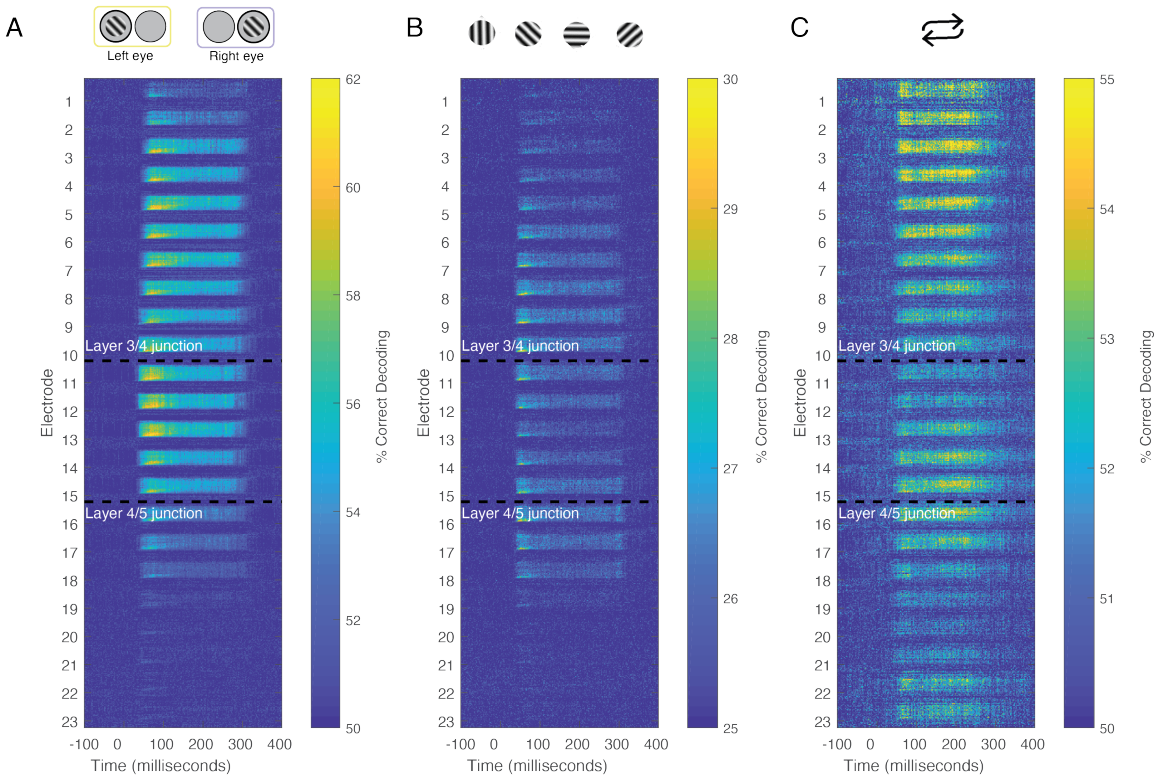

Figure 2.5: Combined time generalization and moving searchlight analysis along the depth of the linear electrode array. (We performed this analysis for each of the main stimulus features analyzed in this paper: Stimulus (A) eye-of-origin, (B) orientation and (C) repetition. Each sub-panel shows a series of time generalization plots ranging from 100 ms before stimulus to presentation to 400 ms post stimulus presentation using a moving searchlight of three electrodes and two electrodes at the end of the electrode array.

2.5 Discussion

Recent studies using linear multielectrode arrays in V1 have successfully contrasted externally evoked feedforward activation with internally generated feedback (Spaak et al., 2012; Maier, 2013; Van Kerkoerle et al., 2014, 2017). These results are encouraging as they demonstrate that the flow of neural activation across cortical layers is highly informative regarding the context of neuronal activation – an important insight that is largely absent in single electrode recordings. In this study we went beyond these earlier findings by showing how the build-up of cortical laminar activation contains several parallel streams for information specific to stimulus features that are difficult to trace using univariate analyses, even when laminar data has been obtained.

2.5.1 Drawing Insight From Multivariate Spiking Profiles

In recent work, layer-specific processes are often grouped to perform univariate analyses to investigate differences between layers (see Westerberg et al., 2019 for example). This is because we often consider cortical processes that follow a model known as the canonical cortical microcircuit (see Bastos et al., 2012 for review). This model hypothesizes three functional compartments in granular cortex: a feedforward recipient granular compartment sandwiched between supragranular and infragranular compartments. While this model has provided powerful insight into cortical function, we know that even within layers there can be degree of heterogeneity in the distribution of neurons. That is, neuron "A" might exist in layer 2 of cortex where neuron "B" exists in layer 3. While both neurons are present in the supragranular compartment and their activity might reflect the same process, the information they carry might vary in meaningful ways. MVPA incorporates information across all channels comprising a predefined laminar compartment. This allows a more integrative approach in evaluating the activity of laminar compartments than previous approaches. Namely, previous work considers independent channels from a laminar compartment representative of the compartment's overall activation state (Westerberg et al.,

2019). However, information might be encoded in the dynamics within a layer that would be lost in univariate analyses.

Another advancement afforded by the MVPA approach is by being able to generalize information states across time. The time generalization analysis allows us to track patterns of information encoding. That is, by evaluating decoding performance by training and testing the classifier at different time periods, we can observe how information processing is remaining consistent or evolving. A stable representation of a feature will not only be decodable at the timepoint in which a classifier is trained, but also at later timepoints. Meanwhile, with a dynamic representation, a classifier will not generalize far beyond the trained time (Carlson et al., 2011; King and Dehaene, 2014; Mohsenzadeh et al., 2018). Furthermore, we can infer how certain stimuli features vary in time and match potential models of neural encoding found across a number of studies (for review see King and Dehaene, 2014).

2.5.2 Implications for the Circuitry of Binocular Combination, Orientation Representation, and Repetition Suppression

The analyses performed here further our understanding of several processes along the V1 laminar microcircuit. First to consider is the laminar profile of binocular combination. Through our analyses, we found that visual signals of each eye are more strongly integrated once they reach the deep layers. We found a drastic reduction in eye-specific information in the lower layers of V1, suggesting the information regarding eye-of-origin are largely resolved prior to the lower layers. This pattern is in line with earlier reports, locating the bulk of V1 binocular neurons in both the upper and lower layers (Hubel and Wiesel, 1977). This apparent paradox might be explained by a recent finding that a large fraction of monocular V1 neurons are sensitive to both eyes (Dougherty et al., 2019a). Thus, a neuron's preference for one or the other eye may not necessarily be predictive of how it responds to binocular stimulation (see also Read and Cumming, 2004). Furthermore, eye-specific in-

formation also seemed to decrease in both the searchlight decoding and time generalization results, indicating that it is more readily dispensed by V1's CCM compared to other types of stimulus information, which seems in line with the fact that eye-of-origin information is of low behavioral relevance (Blake and Cormack, 1979; Solomon and Morgan, 1999; Schwartzkopf et al., 2010). While our findings regarding the representation of eye information the lower layers requires more direct testing to reconcile with previous work, our other finding that each eye's stream of information stays largely separate until visual activation reaches the upper layers of V1 are compatible with hypotheses regarding the origins of binocular combination.

Our results also revealed a fine-grained spatiotemporal laminar pattern of orientation tuning, with some but not all sublayers of granular layer 4 exhibiting less sensitivity to stimulus orientation than the superficial and deep layers of V1. Although it is not immediately clear how the specific pattern produced by MVPA relates to the magno- and parvocellular recipient sublayers, our finding seems to be generally in line with the idea that V1 receives at least some LGN inputs that are somewhat "biased" toward certain stimulus orientations, with further processing within V1 producing the more discerning orientation tuning that characterizes this area.

With respect to the circuitry of adaptation in V1, it is interesting to note that stimulus repetition yielded a unique signature of time generalization in the feedback-recipient layers of V1. Previous work suggested that adaptive changes largely arise from changes in feedback activation in V1 (Westerberg et al., 2019). The temporal features of this time generalization pattern are somewhat reminiscent of prior descriptions of feedback modulation in V1 (Van Kerkoerle et al., 2014). However, our finding goes beyond the demonstration of a secondary peak in activation by revealing that the information content within this activation is specific to contextual information.

2.5.3 Sources for Feature-Specific Activation Patterns in V1

It is interesting to speculate as to the source of these differences in layer-specific information flow. Could it be that differences arise through differences in processing local to V1 or is another brain area affecting feature-specific change in the V1 laminar microcircuit? Previous work has begun to investigate such questions. For example, investigation into the origins of adaptation resulting from visual repetition suggests that the reduction in neural responses in V1 associated with visual repetition comes about through a reduction in the feedback activity to the V1 laminar microcircuit rather than through changes in feedforward processing local to V1 (Westerberg et al., 2019). This is in contrast to the process of binocular combination which is largely thought to be accomplished even prior to the feedforward activation of the supragranular layers of V1. It is through these differences in activation that might elicit the observed differences in information flow along the layers. Further investigation, perhaps through causal inactivation of feedback connections to V1 (Nurminen et al., 2018), would shed light on whether feedback activation is indeed necessary for the observed patterns of information flow described here.

2.5.4 Toward Ultra-High-Resolution Laminar Neurophysiology

We are on the cusp of a revolution in primate neurophysiology that will allow for massively increased insights into the function of mesoscopic neural circuits (Jun et al., 2017; Steinmetz et al., 2018; Musk and Neuralink, 2019). Modern recording technologies have advanced to allow for the simultaneous recording from thousands of channels. This substantial advance in resolution of data allows for the interrogation of data through novel analytical methods. With increased resolution of data comes the ability to investigate data in more integrative approaches. MVPA has proven highly useful in the functional imaging literature where large multichannel datasets have been commonplace for decades. Through the analyses demonstrated here, we propose these same analyses as useful approaches to investigating ultra-high-resolution neurophysiology as these recording techniques become

more and more common.

2.6 References

Asaad, W. F., Santhanam, N., Mcclellan, S., and Freedman, D. J. (2013). High-performance execution of psychophysical tasks with complex visual stimuli in matlab. *Journal of Neurophysiology*, 109(1):249–260.

Bastos, A. M., Loonis, R., Kornblith, S., Lundqvist, M., and Miller, E. K. (2018). Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 115(5):1117–1122.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Binzegger, T., Douglas, R. J., and Martin, K. A. (2004). A quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience*, 24(39):8441–8453.

Blake, R. and Cormack, R. H. (1979). On utrocular discrimination. *Perception Psychophysics*, 26(1):53–68.

Callaway, E. M. (1998). Prenatal development of layer-specific local circuits in primary visual cortex of the macaque monkey. *Journal of Neuroscience*, 18(4):1505–1527.

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10):1–19.

Carlson, T. A., Hoogendoorn, H., Kanai, R., Mesik, J., and Turrett, J. (2011). High temporal resolution decoding of object. *Journal of Vision*, 11(2011):1–17.

Casagrande, V. A. and Boyd, J. D. (1996). The neural architecture of binocular vision. *Eye*, 10(2):153–160.

Cichy, R. M. and Pantazis, D. (2017). Multivariate pattern analysis of meg and eeg: A comparison of representational structure in time and space. *NeuroImage*, 158(July):441–454.

Cox, M. A., Dougherty, K., Adams, G. K., Reavis, E. A., Westerberg, J. A., Moore, B. S., Leopold, D. A., and Maier, A. (2019). Spiking suppression precedes cued attentional enhancement of neural responses in primary visual cortex. *Cerebral Cortex*, 29(1):77–90.

Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A., and Maier, A. (2013). Receptive field focus of visual area v4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 110(42):17095–17100.

Daniels, J. D., Norman, J. L., and Pettigrew, J. D. (1977). Biases for oriented moving bars in lateral geniculate nucleus neurons of normal and stripe-reared cats. *Experimental Brain Research*, 29(2):155–172.

Dougherty, K., Cox, M. A., Westerberg, J. A., and Maier, A. (2019). Binocular modulation of monocular v1 neurons. *Current Biology*, 29(3):381–391.e4.

Douglas, R. J. and Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, 27(1):419–451.

Douglas, R. J., Martin, K. A., and Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural Computation*, 1(4):480–488.

Etzel, J. A., Zacks, J. M., and Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, 78.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierachical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

Goddard, E., Solomon, S. G., and Carlson, T. A. (2017). Dynamic population codes of multiplexed stimulus features in primate area mt. *Journal of Neurophysiology*, 118(1):203–218.

Haxby, V. J., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distrubuted and overlapping representations of face and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

Hubel, D. H. and Wiesel, T. N. (1972). Laminar and columnar distribution of geniculocortical fibers in the macaque monkey - hubel - 2004 - journal of comparative neurology - wiley online library. *Journal of Comparative Neurology*, 146(4):421–450.

Hubel, D. H. and Wiesel, T. N. (1977). Ferrier lecture. functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London - Biological Sciences*, 190(1130):1–59.

Hwang, J., Mitz, A. R., and Murray, E. A. (2019). Nimh monkeylogic: Behavioral control and data acquisition in matlab. *Journal of Neuroscience Methods*, 323:13–21.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685.

King, J. R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210.

Kriegeskorte, N. and Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fmri. *NeuroImage*, 38(4):649–662.

Kriegeskorte, N. and Kreiman, G. (2012). *Visual Population Codes: Toward a Common multivariate Framework for Cell Recording and Functional Imaging*.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis

- connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):4.

Leventhal, A. G. and Schall, J. D. (1983). Structural basis of orientation sensitivity of cat retinal ganglion cells. *Journal of Comparative Neurology*, 220(4):465–475.

Maier, A. (2013). Neuroscience: The cortical layering of visual processing. *Current Biology*, 23(21).

Maier, A., Adams, G. K., Aura, C., and Leopold, D. A. (2010). Distinct superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. *Frontiers in Systems Neuroscience*, 4.

Mignard, M. and Malpeli, J. G. (1991). Paths of information flow through visual cortex. *Science*, 251(4998):1249–1251.

Mitzdorf, U. (1985). Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and eeg phenomena. *Physiological Reviews*, 65(1):37–100.

Mohsenzadeh, Y., Qin, S., Cichy, R. M., and Pantazis, D. (2018). Ultra-rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *eLife*, 7:e36329.

Ninomiya, T., Dougherty, K., Godlove, D. C., Schall, J. D., and Maier, A. (2015). Microcircuitry of agranular frontal cortex: Contrasting laminar connectivity between occipital and frontal areas. *Journal of Neurophysiology*, 113(9):3242–3255.

Ohzawa, I. and Freeman, R. D. (1986). The binocular organization of simple cells in the cat's visual cortex. *Journal of neurophysiology*, 56(1):221–42.

Oosterhof, N. N., Connolly, A. C., and Haxby, V. J. (2016). Cosmomvpa: Multi-modal

multivariate pattern analysis of neuroimaging data in matlab/gnu octave. *Frontiers in Neuroinformatics*, 10(JUL):27.

Prince, S. J., Cumming, B. G., and Parker, A. J. (2002). Range and mechanism of encoding of horizontal disparity in macaque v1. *Journal of Neurophysiology*, 87(1):209–221.

Pugh, M. C., Ringach, D. L., Shapley, R., and Shelley, M. J. (2000). Computational modeling of orientation tuning dynamics in monkey primary visual cortex. *Journal of Computational Neuroscience*, 8(2):143–159.

Read, J. C. A. and Cumming, B. G. (2004). Ocular dominance predicts neither strength nor class of disparity selectivity with random-dot stimuli in primate v1. *Journal of Neurophysiology*, 91(3):1271–1281.

Rockland, K. S. and Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, 179(1):3–20.

Rockland, K. S. and Virga, A. (1989). Terminal arbors of individual "feedback" axons projecting from area v2 to v1 in the macaque monkey: A study using immunohistochemistry of anterogradely transported phaseolus vulgaris-leucoagglutinin. *Journal of Comparative Neurology*, 285(1):54–72.

Schroeder, C., Mehta, A., and Givre, S. J. (1998). A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. *Cerebral Cortex*, 8(7):575–592.

Self, M. W., van Kerkoerle, T., Supèr, H., and Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area v1 in figure-ground segregation. *Current biology : CB*, 23(21):2121–2129.

Shapcott, K. A., Schmiedt, J. T., Saunders, R. C., Maier, A., Leopold, D. A., and Schmid,

M. C. (2016). Correlated activity of cortical neurons survives extensive removal of feedforward sensory input. *Scientific Reports*, 6:1–8.

Smith, E. L., Chino, Y. M., William, H., Ridder, R., Kitagawa, K., and Langston, A. (1990). Orientation bias of neurons in the lateral geniculate nucleus of macaque monkeys. *Visual Neuroscience*, 5(6):525–545.

Spaak, E., Bonnefond, M., Maier, A., Leopold, D. A., and Jensen, O. (2012). Layer-specific entrainment of gamma-band neural activity by the alpha rhythm in monkey visual cortex. *Current Biology*, 22(24):2313–2318.

Tovar, D., Murray, M., and Wallace, M. (2020). Selective enhancement of object representations through multisensory integration. *Journal of Neuroscience*, 40(29):5604–5615.

van den Hurk, J. and Op de Beeck, H. (2019). Generalization asymmetry in multivariate cross-classification : When representation a generalizes better to representation b than b to a. *bioRxiv*.

Van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M. A., Poort, J., Van Der Togt, C., and Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(40):14332–14341.

van Kerkoerle, T., Self, M. W., and Roelfsema, P. R. (2017). Erratum: Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nature communications*, 8:15555.

Vidyasagar, T. R. and Urbas, V. J. (1982). Orientation sensitivity of cat lgn neurones with and without inputs from visual cortical areas 17 and 18. *Experimental Brain Research*, 46(2):157–169.

Westerberg, J. A., Cox, M. A., Dougherty, K., and Maier, A. (2019). V1 microcircuit dynamics: altered signal propagation suggests intracortical origins for adaptation in response to visual repetition. *Journal of neurophysiology*, 121(5):1938–1952.

Xu, X., Ichida, J., Shostak, Y., Bonds, A. B., and Casagrande, V. A. (2002). Are primate lateral geniculate nucleus (lgn) cells really sensitive to orientation or direction? *Visual Neuroscience*, 19(1):97–108.

2.7 Supplemental Figures

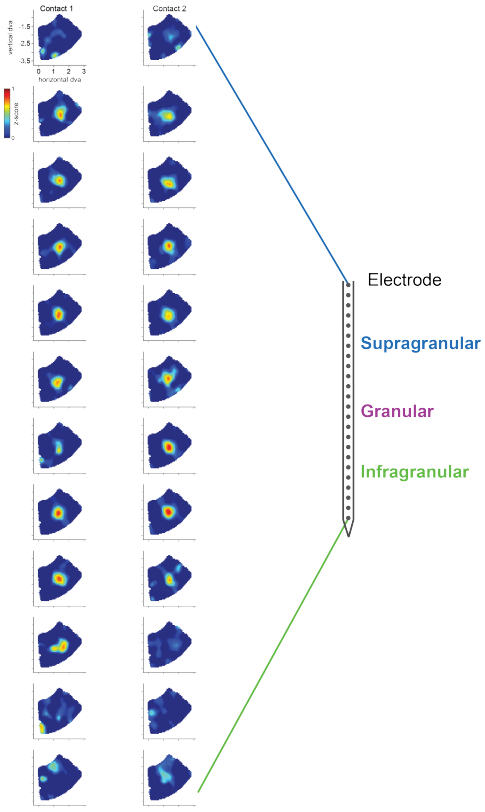

Figure 2.6: Receptive field mapping. For each contact of the linear multielectrode array, we computed the magnitude of MUA spiking responses as a function of visual field stimulation using a reverse-correlation technique (see Methods). Colored plots to the left show averaged neuronal response in units of standard deviation as a function of angle and magnitude in visual degrees. Panels are arranged in descending order with each column representing neighboring channels on the electrode array so that each row represents the electrode channel that is 200 microns below the channel above. Note that the receptive field locations deviate little between the top and the bottom of the array, indicating that the electrode was inserted perpendicularly to the cortical surface. The topmost and bottommost channels of the array produced no visual responses as these electrode channels reached outside the cortical thickness.

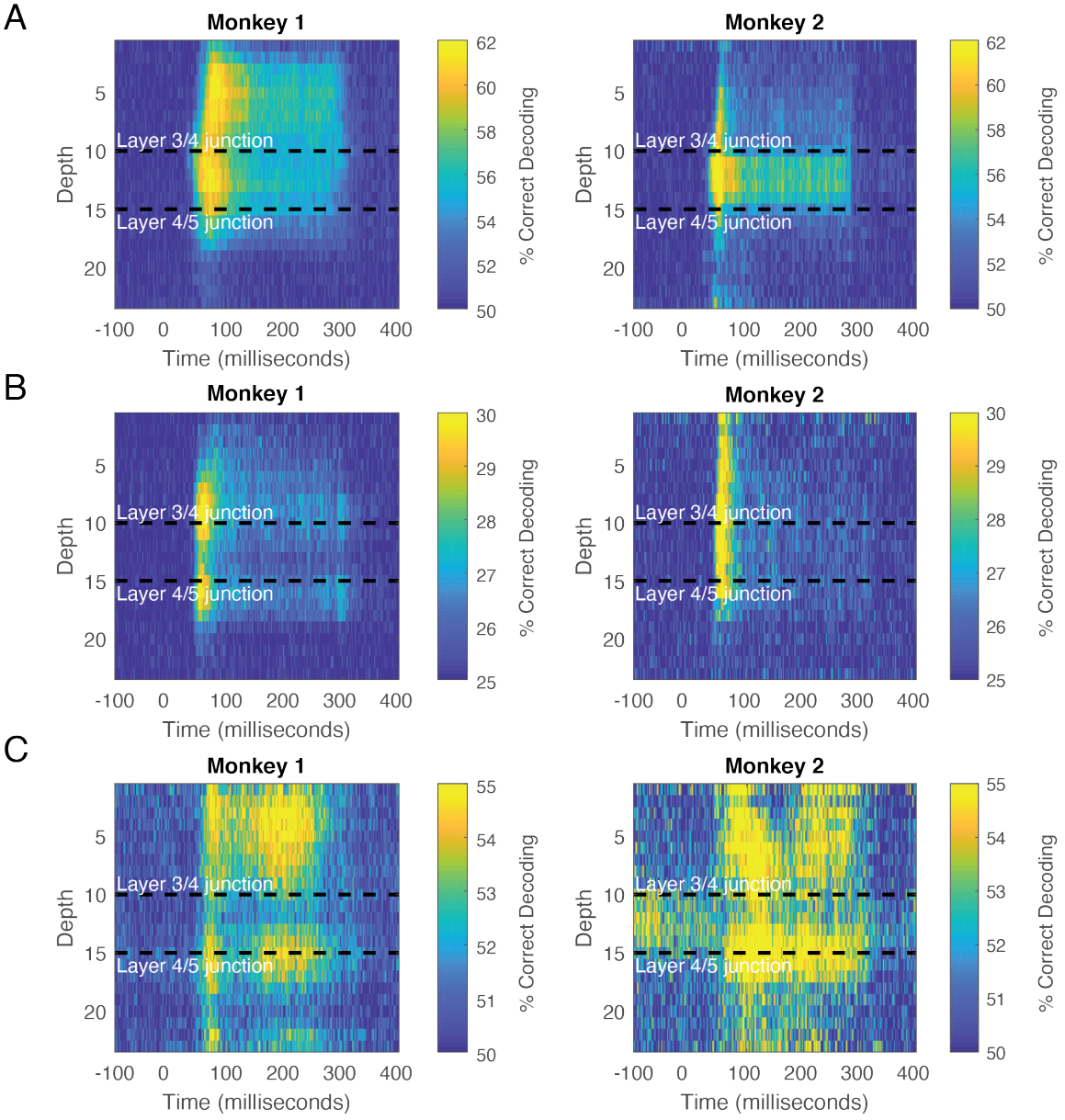

Figure 2.7: Supplemental searchlight analysis separated by monkey. Decoding performance using a moving searchlight along the electrode array for (A) eye of origin, (B) grating orientation, and (C) stimulus repetition. Monkey 1 (left panel) and Monkey 2 (right panel).

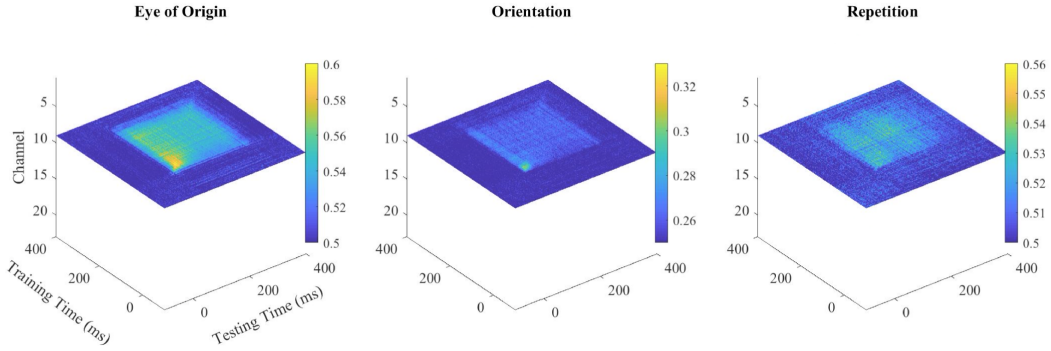

Figure 2.8: Video of Combined Time Generalization and Searchlight Analysis. Video can be found here: https://ndownloader.figstatic.com/files/25628039

Chapter 3

Volume conduction masks feature information in locally generated LFP

3.1 Abstract

Local field potentials (LFP) are low-frequency voltage fluctuations which reflect neural coding within and across brain areas. However, unlike neuronal spiking which is highly localized, LFP is spatially non-specific – what is measured at one location is not necessarily generated there. This volume-conducted component of the LFP might therefore interfere with accurate measurement, and subsequent interpretation, of the information conveyed in the locally generated low-frequency signal. We sought to uncover whether information embedded in locally generated low-frequency activity was masked by the volume-conducted signals. Monkeys viewed sequences of multifeatured stimuli while laminar recordings were performed in area V1. We compared information content of volume-conducted and locally generated LFP through spatiotemporal multivariate pattern analysis of cortical columns. Volume-conducted vs. locally generated information dissociated in two important ways. For stimulus features (orientation and eye-of-origin), locally generated LFP held more information. Conversely, the volume conducted signal was more informative with respect to temporal context (stimulus position in sequence). These relationships were layer specific. We further explored these relationships with respect to frequency bands. This revealed distinct patterns of shared information between frequency bands which differed for volume-conducted and locally generated signals. These findings reveal low-frequency neural activity generated at the level of laminar cortical microcircuits encode information and display cross-frequency relationships which are masked by volume-conducted activity.

3.2 Introduction

The local field potential (LFP) is a complex, far-reaching signal comprising transmembrane potentials arising from incoming synaptic inputs (Buzsáki, Anastassiou, & Koch, 2012; Mitzdorf, 1985), sodium currents (Ray, Crone, Niebur, Franaszczuk, & Hsiao, 2008),

calcium currents (Schiller, Major, Koester, & Schiller, 2000), and gap junctions (Traub & Bibbig, 2000). LFP captures graded potentials in addition to the all-or-none response of the action potential, effectively providing a larger population response (Bijanzadeh, Nurminen, Merlin, Clark, & Angelucci, 2018). As a result, the information contained in LFPs often complements what is found from action potentials (Leszczyński et al., 2020; Mineault, Zanos, & Pack, 2013), and in some cases explains behavioral responses better than action potentials (e.g., Pesaran, Pezaris, Sahani, Mitra, & Andersen, 2002). Additionally, LFPs show more consistency across recording sessions than population spiking activity, as they are not affected as much by the position of the electrode relative to the recorded population (Bédard, Kröger, & Destexhe, 2004). This stability can be partially explained by the LFPs proclivity to diffuse through space. However, this otherwise helpful property can be problematic for identifying the structures involved in generating or receiving neural signals. There are a number of known instances where a brain structure thought to have been involved in producing a neural signal, was in fact the result of volume conducted LFPs from nearby structures (Bertone-Cueto et al., 2020; Kajikawa, Smiley, & Schroeder, 2017; Lalla, Rueda Orozco, Jurado-Parras, Brovelli, & Robbe, 2017). Accounting for this volume conduction is important given that stimulus features are processed preferentially across sensory cortical areas, within areal maps comprised of columns, and even across the layers of a cortical column (Tovar et al., 2020).

Variability exists in reports of the extent to which LFP volume conducts through neural tissue. Reports range from a few hundred micrometers (Katzner et al., 2009; Xing, Yeh, & Shapley, 2009) to a few millimeters horizontally and centimeters vertically (Kajikawa & Schroeder, 2011, 2015; Kreiman et al., 2006; Nauhaus, Busse, Carandini, & Ringach, 2009). A number of factors contribute to this variability. Cell morphology is a considerable factor, with modeling showing that pyramidal cells, due to their asymmetry, have larger LFP spread than any other cell type (Lindén et al., 2011). Complicating matters further, the spontaneous correlation between cells at rest as well as during stimulus presentation,

affected by factors such as the presence of horizontal cells and the particular stimulus features encoded, can affect passive spread by an order of magnitude (Leski, Lindén, Tetzlaff, Pettersen, & Einevoll, 2013; Lindén et al., 2011; Rosenbaum, Smith, Kohn, Rubin, & Doiron, 2017). Lastly, volume conducted passive spread from the activation observed when neural activity in one brain area elicits activity in neighboring brain areas, propagating a traveling wave (Sato, Nauhaus, & Carandini, 2012; Zanos, Mineault, Nasiotis, Guitton, & Pack, 2015). These factors can make triangulation of a source for LFP difficult.

However, the implications of the volume conducted component can be investigated through careful transformation of the LFP signal. Current source density (CSD), the second spatial derivative of the LFP (Mitzdorf, 1985), estimates localized synaptic activations comprising the spatially non-specific LFP. Interestingly, the CSD seems to be a more complex signal than LFP and population spiking, requiring more principle components to explain signal variance (Einevoll et al., 2007; Schaefer, Kössl, & Hechavarría, 2017). Importantly, the CSD can be re-summed into LFP_{Cal} – an estimate of locally generated LFP (at the columnar microcircuit scale) minimizing contamination by volume conduction. However, investigation using the LFP_{Cal} signal is limited. The LFP_{Cal} has primarily been used to quantify the amount of volume conduction in the original LFP signal (Kajikawa & Schroeder, 2011). This recalculated signal has not been used to study how feature processing is affected by volume conduction.

Here, we used multivariate pattern analysis, exploiting information captured by the laminar response variability, to study the information present in the volume conducted and locally generated LFP in V1 during visual presentation of multifeatured stimuli. Since different stimulus features are uniquely processed across brain areas and as such might have different degrees of volume conduction, we studied stimulus features primarily localized to V1 as well as stimulus features that may also be processed outside of V1. We find that decoding performance for features processed within V1 suffer from volume conduction effects while decoding performance for stimulus features processed outside of V1 appear

enhanced by volume conduction. Additionally, we found that volume conduction differentially affected stimulus features information across frequency bands associated with varying degrees of feedforward and feedback processes (Bastos et al., 2012; Bastos et al., 2015; Belitski et al., 2008; Peter et al., 2019; Van Kerkoerle et al., 2014). Our findings demonstrate that volume conducted signals can mask information relayed in locally generated low-frequency signals. Moreover, these findings depend on the feature being processed and are differentially impacted with respect to the frequency measured.

3.3 Methods

3.3.1 Animal care and surgical procedures

Procedures were in accordance with National Institutes of Health Guidelines, Association for Assessment and Accreditation of Laboratory Animal Care Guide for the Care and Use of Laboratory Animals, and approved by the Vanderbilt Institutional Animal Care and Use Committee following United States Department of Agriculture and Public Health Services policies. Two macaque monkeys (Macaca radiata: monkey E48 [male], monkey I34 [female]) underwent a series of surgeries implanting MR compatible head posts and cranial recording chambers positioned over one hemisphere of V1. A craniotomy was performed concurrent with the location of the recording chamber. All surgical procedures were performed under general anesthesia. Anesthetic induction was performed with ketamine (5-25 mg/kg). Monkeys were then catheterized and intubated. Surgeries were performed under aseptic conditions. N2O/O2, isoflurane (1-5%) anesthesia was used. Vital signs were monitored continuously. Postoperative antibiotics and analgesics were administered. Additional descriptions of animal care and surgical procedures can be found elsewhere (Westerberg, Cox, Dougherty, & Maier, 2019; Westerberg, Maier, & Schall, 2020; Westerberg, Maier, Woodman, & Schall, 2020).

3.3.2 Magnetic resonance imaging

Magnetic resonance (MR) imaging was used to guide recording chamber implant surgeries as well as to guide linear electrode array penetrations. All MR scans were conducted with animals under general anesthesia per the procedures described in Animal care and surgical procedures. Scans were obtained using a Philips 3T MR scanner. T1-weighted 3D MPRAGE scans were acquired with a 32-channel head coil equipped for sense imaging. Images were acquired using 0.5 mm isotropic voxel resolution with the following parameters: repetition 5 s, echo 2.5 ms, and flip angle 7°.

3.3.3 Visual display and stimuli

Monkeys viewed stimuli presented on a 20" CRT monitor at 60 or 85 Hz. Stimuli were presented through a custom mirror stereoscope allowing for monocular or binocular presentation of stimuli (Figure 3.1A) (Cox et al., 2019b; Dougherty et al., 2019, 2021). Prior to performance of the main task, monkeys performed a stereoscope calibration task as to eliminate any potential confound of binocular disparity (see Tovar and Westerberg et al., 2020). Each recording session (N = 61, monkey E48: 48, monkey I34: 13) comprised several hundred trials each with sequences of 1-5 stimulus presentations (Figure 3.1B). Stimulus displays were generated using MonkeyLogic (Asaad & Eskandar, 2008; Hwang, Mitz, & Murray, 2019). Monkeys initiated trials by fixating within 1 degree of visual angle (dva) of a central fixation dot. Following a short (300 ms) fixation period, stimuli appeared in sequence (1-5 stimuli) each for 200 ms with a 200 ms inter-stimulus interval. If monkeys maintained fixation throughout the stimulus sequence, they received a juice reward before an inter-trial interval ensued. If monkeys failed to maintain fixation throughout the trial, they received a brief timeout before the inter-trial interval. Each stimulus presented was a sinusoidal bar grating stimulus localized to the visual receptive field of the cortical column (see Receptive field mapping below). The stimuli maintained the same size, spatial frequency, and phase throughout each recording session (optimized for the column multi-unit

activity, see Cox et al., 2013, 2019a, 2019b), but had variable orientation and eye-of-origin. Stimuli were only presented monocularly.

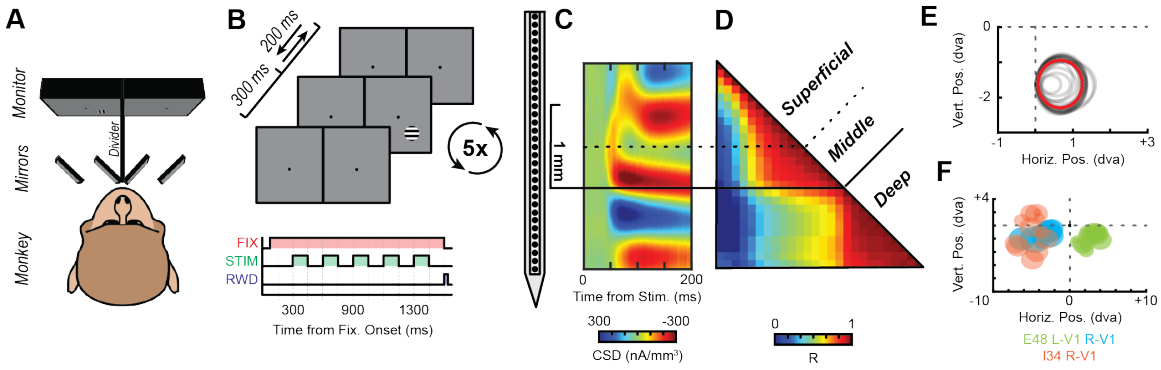

Figure 3.1: Experimental setup and laminar alignment. A. Stereoscope setup for stimulus presentation. Monkey is positioned behind a series of mirrors which provide stimulation to each of the eyes independently. B. Stimulus presentation sequence. Monkeys fixated for 300 ms after which a stimulus was presented to the receptive field of the V1 column for 200 ms. 5 stimuli were presented sequentially with 200 ms inter-stimulus intervals while monkeys maintained central fixation. C. Laminar alignment across sessions using current source density (CSD). For each session (n=61), CSD was computed and the L4/5 boundary identified as the early sink with an accompanying deeper source. Color plot shows session average CSD response to stimulus in the column RF following alignment. D. LFP correlations were also used to confirm laminar alignment. Color plot shows profile of contact-by-contact correlations across time (512 ms moving window) during each session then averaged across all sessions following laminar alignment. E. Representative RF mapping for a single session. Each gray circle represents the estimated multiunit RF for a recording site identified to be in cortex. The red circle indicates the average along contacts. All circles overlap indicating perpendicular penetration. F. Average column RF for each session where color indicates the monkey-hemisphere combination where the column was recorded.

3.3.4 Neurophysiological procedure

Broadband (0.5-12.207 kHz) neurophysiological signal was recorded during task performance. Signals were amplified and digitized at 30 kHz using a 128-channel Cerebus Neural Signal Processing System (Blackrock Microsystems). LFP signals were downsampled to 1 kHz. All neural recordings were performed using 24- or 32-channel linear microelectrode arrays with 0.1 mm interelectrode spacing (S-probe, U-Probe, V-Probe – Plexon; Vector array – NeuroNexus) positioned orthogonal to the cortical surface in dorsal V1. Microelectrode recording contacts had impedances between 0.2-0.8 MOhms. Electrode arrays were held in position using a custom Narishige micromanipulator. Electrode arrays were interfaced with the amplifier system using the Blackrock analog headstage. Gaze was measured binocularly at 1 kHz using an Eyelink system (SensoMotoric Instruments). All recordings took place in a radio frequency-isolated booth.

3.3.5 Receptive field mapping

Receptive field mapping was performed during the neural recordings to ascertain the receptive field of the cortical column being recorded from as well as to confirm an orthogonal electrode array penetration into V1 (Figure 3.1E-F). Multiunit and local field potential activity was measured while a series of stimuli were presented in the contralateral lower quadrant of the visual hemifield relative to the position of the recording chamber. Monkeys fixated a central fixation dot while 1-5 stimuli were presented. Successful maintenance of fixation throughout the stimuli presentations yielded a juice reward. Qualitative (auditory evaluation) and quantitative assessment of the visual responses was performed online. More detailed description of the procedure is detailed elsewhere (Cox et al., 2013). Monkeys proceeded to perform the main task described in Section 3.3.3 if there was an observable receptive field which was consistent along cortical depth, indicative of an orthogonal presentation. The measured receptive field in this task was used as the location for stimuli presentation in the main task.

3.3.6 Current source density and laminar alignment

Current source density (CSD) served to identify the location of the electrode relative to the layers of V1 (Mitzdorf, 1985). The spatiotemporal profile of CSD has a distinct pattern which allows for the reliable identification of the boundary between the granular input layers and the infragranular layers of V1 (Schroeder et al., 1998). To compute the CSD from the LFP, we used previously describe procedure (Nicholson & Freeman, 1975):

$$CSD(t,d) = -\sigma\left(\frac{x(t,d-z) + x(t,d+z) - 2x(t,d)}{z^2}\right) \tag{3.1}$$

Where the CSD at timepoint t and at cortical depth d is the sum of voltages x at electrodes immediately above and below (z is the interelectrode distance) minus 2 times the voltage at d divided by the interelectrode-distance-squared. That yields the voltage local to d. To transform the voltage to current, we multiplied that by - σ, where σ is a previously reported estimate of the conductivity of cortex (Logothetis, Kayser, & Oeltermann, 2007). In addition to using CSD for laminar alignment (Figure 3.1C), we confirmed positioning of the electrode array relative to the layers by identifying reliable patterns in the correlations between LFP across electrodes (Figure 3.1D) and in the LFP power spectral density (PSD, Figure 3.3) through previously reported means (Maier, Adams, Aura, & Leopold, 2010; Westerberg et al., 2019).

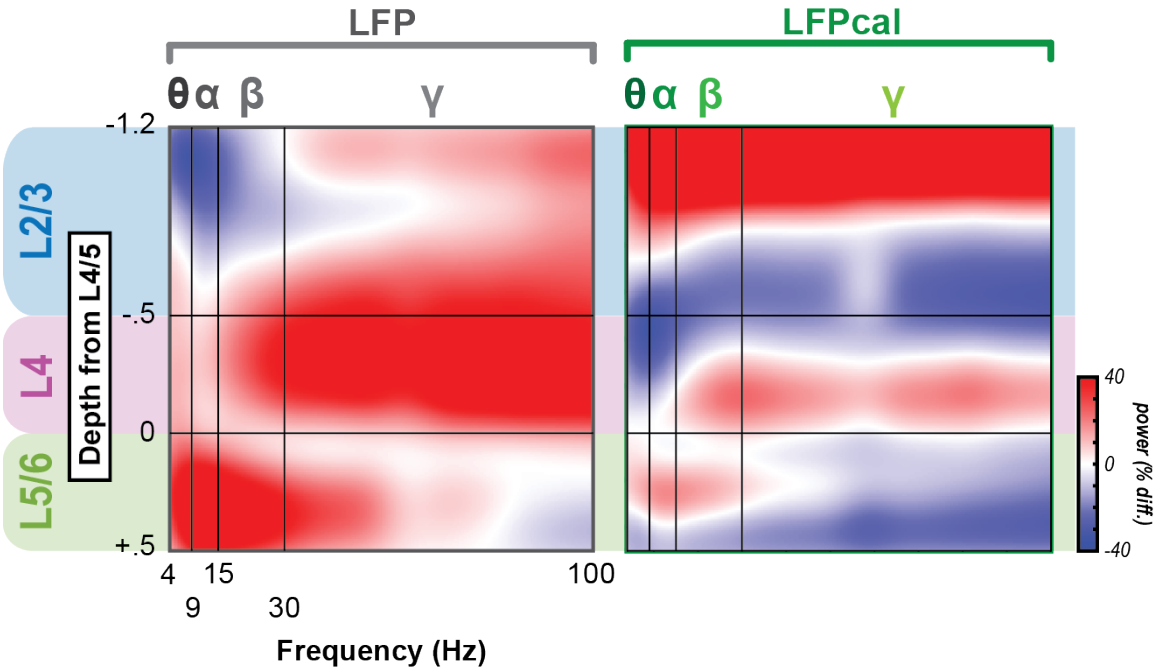

Figure 3.2: Laminar power spectral density for volume conducted and locally generated LFP signals. Left column shows raw LFP signal and right column, the LFP_{Cal}. PSD normalized by finding the average power for each frequency along depth and then power for each contact for each frequency was taken as the percent difference from the column average. Profiles were first normalized at the session level then averaged across sessions (n=61). Ordinates are the depth relative to the L4/5 boundary and abscissa, the frequency. Red indicates greater that column average power in that frequency at that depth. Blue indicates lower than column average power. The transformation of LFP to LFP_{Cal} modifies the laminar profile of field potential power.

3.3.7 Locally generated LFP recalculation

The configuration of microelectrodes in the linear array provides the opportunity to recalculate the low-frequency LFP signal without contamination of volume-conduction. We calculated the locally generated component of the LFP from the measured laminar CSD (LFP_{Cal}) using a previously described model (Kajikawa & Schroeder, 2011,2015; Nicholson & Llinas, 1971):

$$LFP_{Cal}(d_j, t) = A \sum_j \frac{CSD(d_j, t)}{\sqrt{h^2 + |d_j - d_i|^2}} \tag{3.2}$$

where LFP_{Cal} at depth i (d_i) for each timepoint t is taken as the sum of CSD at depths j (d_j) for each timepoint divided by the Euclidean distance to account for the attenuating impact of local currents on distant field potentials. The factor A acts only as a scaling factor and we cannot accurately estimate the magnitude of the one-dimensional CSD-derived waveform, so we eliminate this parameter from the calculation. This omission is consistent with previous reports (Kajikawa and Schroeder, 2011) and limits our comparisons of volume-conducted LFP and the locally generated LFP_{Cal} to only shape of the waveforms. However, magnitude differences can be observed between conditions for the volume-conducted and locally generated LFP, independently. Also, for our purposes, we set h, the displacement distance of the center of mass of CSD from the array electrode assuming vertically aligned CSD components, to 0 as we assume that our observed CSD and the recalculated LFP are colocalized.

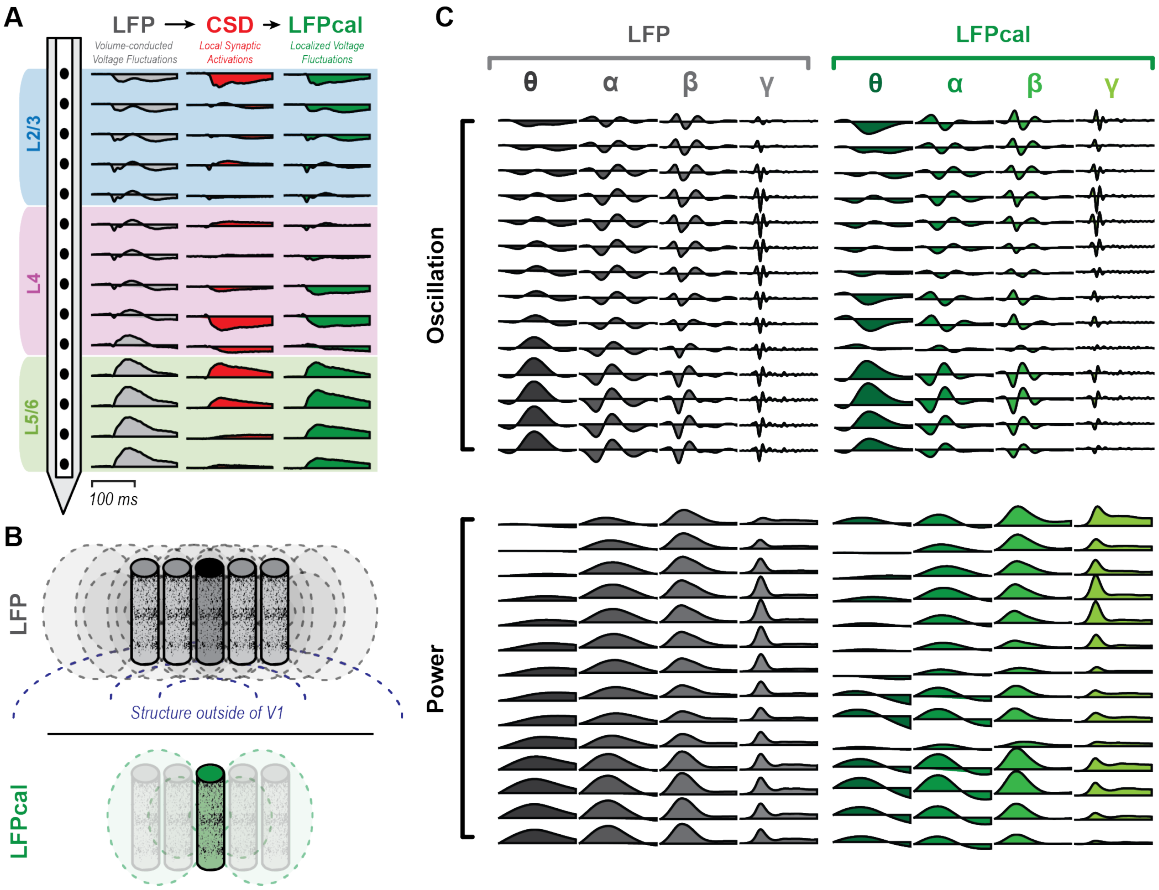

Figure 3.3: Calculation of locally-generated LFP from volume-conducted signal for a representative session. A. Procedure for finding LFP_{Cal}. Raw LFP signal is taken at each electrode contact. CSD is computed as the second spatial derivative of the LFP signal. LFP_{Cal} is then calculated from the CSD as the sum of field potentials generated at each of the electrodes in the column accounting for the attenuation of magnitude with respect to distance from contact. Blue, purple, and green background indicate the laminar compartment from which the data was taken from. B. Cartoon exemplifying the concept behind the LFP_{Cal} procedure. Cylinders represent columns. Dashed lines represent field potentials. LFP signals stem from both locally-generated and volume-conducted signals including LFP generated in nearby cortical columns as well as deeper neural structures. The LFP_{Cal} procedure attenuates or eliminates the signals generated outside of the recorded column to estimate the LFP that is locally-generated in isolation. C. LFP frequency bands comprising both the raw LFP signal as well as the LFP_{Cal} can be filtered out.

3.3.8 LFP frequency analysis

Once the volume conducted LFP was recalculated and locally generated LFP_{Cal} were isolated, we performed a filtering step (where necessary) to investigate differences that might exist with respect to component frequency bands. Filtering was done using a bidirectional bandpass 2nd order Butterworth filter (Maier et al., 2011). Filtering was performed on the raw neurophysiological signal prior to extracting trials. Power spectral density was calculated on the raw neurophysiological signal through a Fourier transform.

3.4 Results

3.4.1 Dissociated information in volume conducted versus locally generated LFP

LFP data was recorded using linear microelectrode arrays affording laminar localization and alignment (Figure 3.1C). Moreover, we can recompute the locally generated LFP – broadband and in distinct frequency bands – from the volume conducted signal using CSD as an intermediary. This process is detailed for an example session in Figure 3.3. The volume conducted and locally generated LFP signals will hereafter be referred to as LFP and LFP_{Cal}, respectively, to reflect the underlying measurement type/derivation. We assure the transformation between LFP and LFP_{Cal} alters the spatial profile of LFP by evaluating the power spectral density (PSD) along the layers of cortex (Figure 3.3). This demonstrates a difference – at least in the spectral power and content – between the volume conducted and locally generated LFP. However, we were interested in evaluating the information content in the LFP between these spatial scales. This derivation affords that opportunity. To compare how stimulus features differed and evolved over time and space for the LFP and LFP_{Cal} signals, we extracted information regarding stimulus features by performing a moving searchlight analysis (Etzel, Zacks, & Braver, 2013; Tovar et al., 2020) for each of the signals. In this analysis, we trained and tested a linear discriminant analysis (LDA) classifier for each session using one electrode and its immediate neighboring electrodes at each timepoint, iteratively repeating the process until we have performed the analysis for the

entire laminar probe across the stimulus interval [-100 ms to 400 ms]. This analysis creates spatiotemporal maps of feature information flow within the V1 microcircuit for each session. To isolate the information that is present in the LFP signal due to volume conduction, we subtracted the stimulus feature information found in the LFP_{Cal} from the LFP searchlight results. This volume conducted signal can be coming from cortical columns immediately surrounding the electrodes, as well as structures outside of V1 altogether (Bertone-Cueto et al., 2020; Kajikawa & Schroeder, 2011; Kajikawa et al., 2017). For the LFP and LFP_{Cal} spatiotemporal maps, we used Wilcoxon signed rank tests to evaluate for significance against chance decoding. Chance decoding is the percentage that would be obtained if the classifier guessed stimuli labels randomly, with FDR correction for multiple comparisons over time and space. For the LFP and LFP_{Cal} differences, significance was evaluated against zero.

For eye-of-origin (Figure 3.4A), the spatiotemporal decoding maps between LFP and LFP_{Cal} varied primarily during the initial transient response. In the LFP, eye-of-origin information first emerged in the granular layer and quickly spread to the supragranular layer, but was greatly diminished in the infragranular layers. Conversely, for the LFP_{Cal} signal, information was present in all layers, including in the infragranular layers near L4/L5 border. Initially, it was somewhat surprising to find information in infragranular layers of the LFP_{Cal} signal, given our own previous findings of reduced eye-of-origin information in infragranular neural spikes along with others (Dougherty, Cox, Westerberg, & Maier, 2019; Hubel & Wiesel, 1972; Tovar et al., 2020). The finding, however, may be a product of LFP and CSD signals largely representing synaptic inputs (Buzsáki et al., 2012; Mitzdorf, 1985) while spikes represent output signals, contributing to them having different spatiotemporal profiles (Leszczyński et al., 2020). Nevertheless, there was more information in the LFP_{Cal} signal throughout the transient response when subtracting LFP_{Cal} from the LFP signal, with the most prominent difference seen in the infragranular layer at the transient and shortly following the transient response. Together, these findings show that the volume

conducted signals from surrounding cortical columns and other areas in the brain add noise to signals containing eye-of-origin information.

The LFP_{Cal} signal also contained more orientation information than the LFP signal (Figure 3.4B), varying most notably during the transient response. The LFP signal information was uneven across layers, localizing primarily to the supragranular layers near the L3/L4 border. Meanwhile, the LFP_{Cal} transient response contained orientation information evenly throughout the compartments. The orientation information in the LFP_{Cal} also appeared to be more prolonged than the LFP. These differences become evident when subtracting the two signals, with the most sustained differences observed in the granular layers. These results show that in addition to eye-of origin, volume conduction obscures orientation information in the V1 microcircuit. Additionally, they support the overarching idea that more signal is not necessarily better, with reduced LFP signals derived from CSD containing more information.

However, for stimulus history, the pattern of LFP_{Cal} containing more information than LFP was broken (Figure 3.4C). Additionally, stimulus history information is distributed differently for the LFP and LFP_{Cal} signals. For LFP, stimulus history is most prominently found in the granular and supragranular layers. However, for the LFP_{Cal} signal, stimulus history information is found in the supragranular layers and infragranular layer. The LFP_{Cal} spatiotemporal profile is more consistent with previous reports of where the feedback recipient layers in V1 are located (Van Kerkoerle et al., 2014; Westerberg et al., 2019). Despite the different spatiotemporal profiles, overall there was more stimulus history information in the LFP signal both before and after stimulus presentation, especially in the granular and supragranular layers. These results show that while stimulus history is present in local signals, information is found to an even greater degree in distant signals – including those that are likely coming from structures downstream of V1. Together, these results show how volume conduction selectively increases or decreases stimulus information at the recording site depending on how local and distant brain areas process stimulus features.

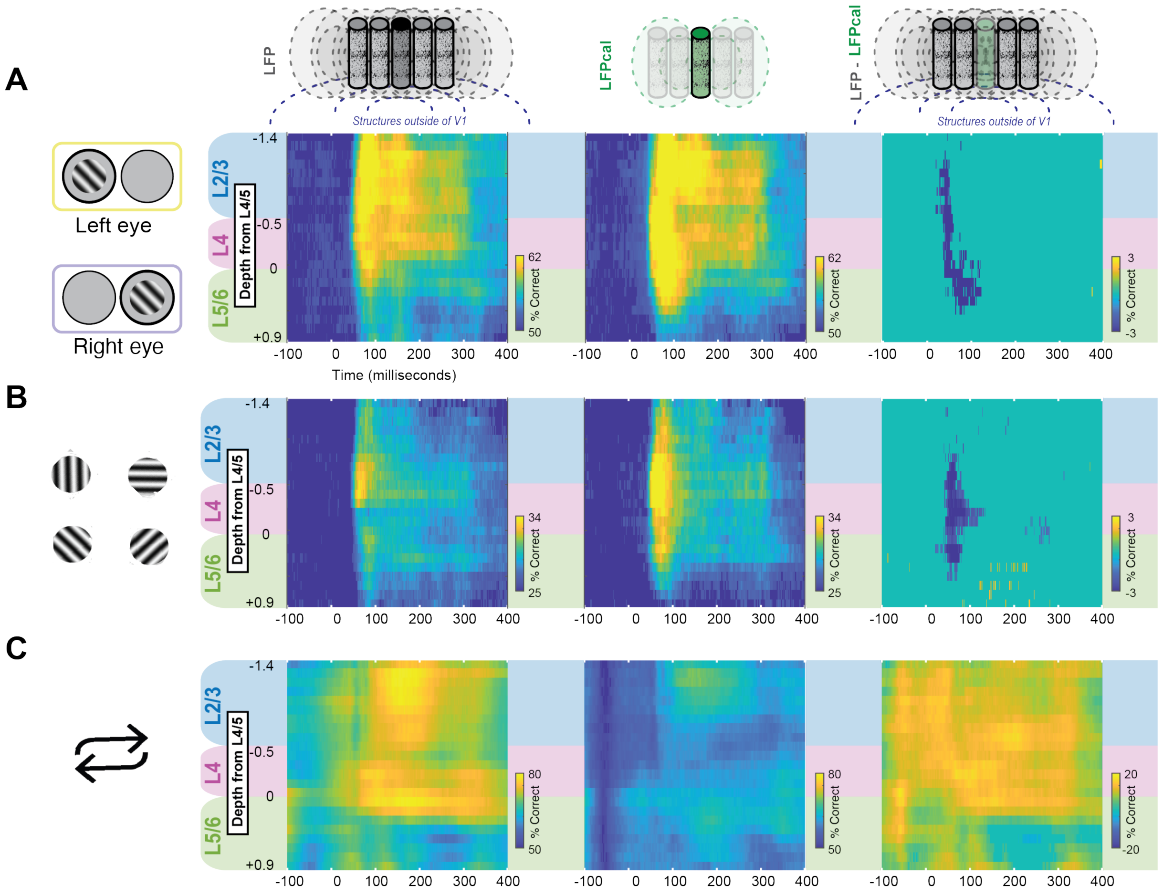

Figure 3.4: More information in the reduced LFP_{Cal} signal depending on stimulus feature. A 3-electrode searchlight along the laminar probe was used to decode (A) eye-of-origin, (B) orientation, and (C) stimulus repetitions. Left column shows results for the LFP signal, middle column shows the LFP_{Cal} decoding results and right column shows the LFP_{Cal} searchlight maps subtracted from the LFP searchlight maps. All results tested for significance against chance decoding: 50% for eye-of-origin and stimulus repetitions, 25% for orientation, and 0% for the searchlight map differences, FDR corrected q=0.05. Ordinates are the depth relative to the L4/5 boundary and abscissa, the time in milliseconds. Blue indicates lower decoding performance and yellow higher decoding performance.

3.4.2 Unique frequency patterns evolve for stimulus features over time

Next, we investigate how the relative distribution of information regarding stimulus features in different LFP frequency bands is affected by volume conduction. To this end, we used a novel frequency generalization analysis on the localized CSD signal using electrodes from all laminar compartments. In this analysis we began by constructing a time frequency spectrogram, collapsing across trials regardless of stimulus features. This time-frequency spectrogram was used to identify key epochs of interest that capture the evolution of the neural response to stimulus presentation. Specifically, we selected epochs representing: (1) time window prior to stimulus presentation, (2) the transient peak following stimulus presentation, (3) sustained response, (4) stimulus offset, and (5) time window after stimulus presentation. Note that the signal is centered at these times, but there was a degree of temporal imprecision, as we found signal prior to stimulus presentation in the spectrogram (Figure 3.5A). Thus, the relative relationship between frequencies for different stimulus features is our metric of interest. In Figure 3.6B, we show a schematic explaining the frequency generalization analysis. Briefly, we trained a classifier at one frequency band and then tested the classifier at all remaining frequency bands to reveal how well different types of stimulus feature information at a particular frequency band generalized across frequency bands. We iteratively repeated this process until we had constructed a matrix in which all frequency bands were used for training and testing. This analysis was done separately for the selected epochs of interest.

In Figure 3.5C, we show various cartoon models of matrix patterns that might arise from the frequency generalization analysis. Patterns found through this analysis have unique interpretations. For a shared broadband pattern, the information at one frequency is equally shared across all frequency bands, creating a square like pattern. For unique narrowband, each frequency bands contains unique information regarding the stimulus feature and thus information will not generalize across frequencies, creating a diagonal along the matrix. Conversely, if information is shared and contained within distinct bands, we would expect

that two distinct squares would emerge within the generalization matrix. Lastly, we show what the matrix information might be contained and shared amongst the low frequency bands or high frequency bands. While these are examples of discrete model patterns, it is more likely that stimulus information will adopt combinations of these matrix patterns.

In Figures 3.5D-F, we show the frequency generalization matrices at various times for eye-of-origin, orientation, and stimulus history evolving over time. All matrices were thresholded for significance using FDR correction for multiple comparisons, $q<0.05$. During the pre-stimulus period, eye-of-origin and orientation is localized to a narrowband 5-7 Hz. For stimulus history, the decoding information generalizes from 5 Hz all the way to 65 Hz, highlighting the baseline shifts that may be informing the decoding differences in the stimulus history. Additionally, there is significant pre-stimulus decoding for both eye-of-origin and orientation to a localized narrowband 5-7 Hz. However, this isolated low frequency information most likely reflects the temporal imprecision of the method for low frequency band. At the transient peak, marked differences emerge in the frequency pattern across stimulus features. Eye-of-origin and orientation both show a shared broadband response, but orientation information shows more shared information across frequencies. Stimulus history on the other hand contains distinct low frequency and high frequency bands of information delineated approximately around low gamma (>30 Hz.) For the sustained period of activity, eye-of-origin has distinct bands of information, one unique narrowband from 5-25 Hz and then another shared broadband pattern of information from 45-85 Hz. Orientation information on the other hands was largely localized to the lower frequency bands at <25 Hz. Stimulus history does not considerably change from what was observed in the transient peak. At stimulus offset, eye-of-origin and orientation information is decreased and diffuse across frequency bands, with modest differences between frequency bands lower and higher than 20 Hz. Stimulus history is similarly diffuse, without distinct bands. Following stimulus offset, stimulus feature information dissipates with only some remnant eye-of-origin and stimulus history information in the low frequency bands

(>20 Hz). Together, these results demonstrated how feature information contained in different frequency bands varies dramatically over the course of a stimulus presentation, and how this may subsequently potentially affect volume conduction.

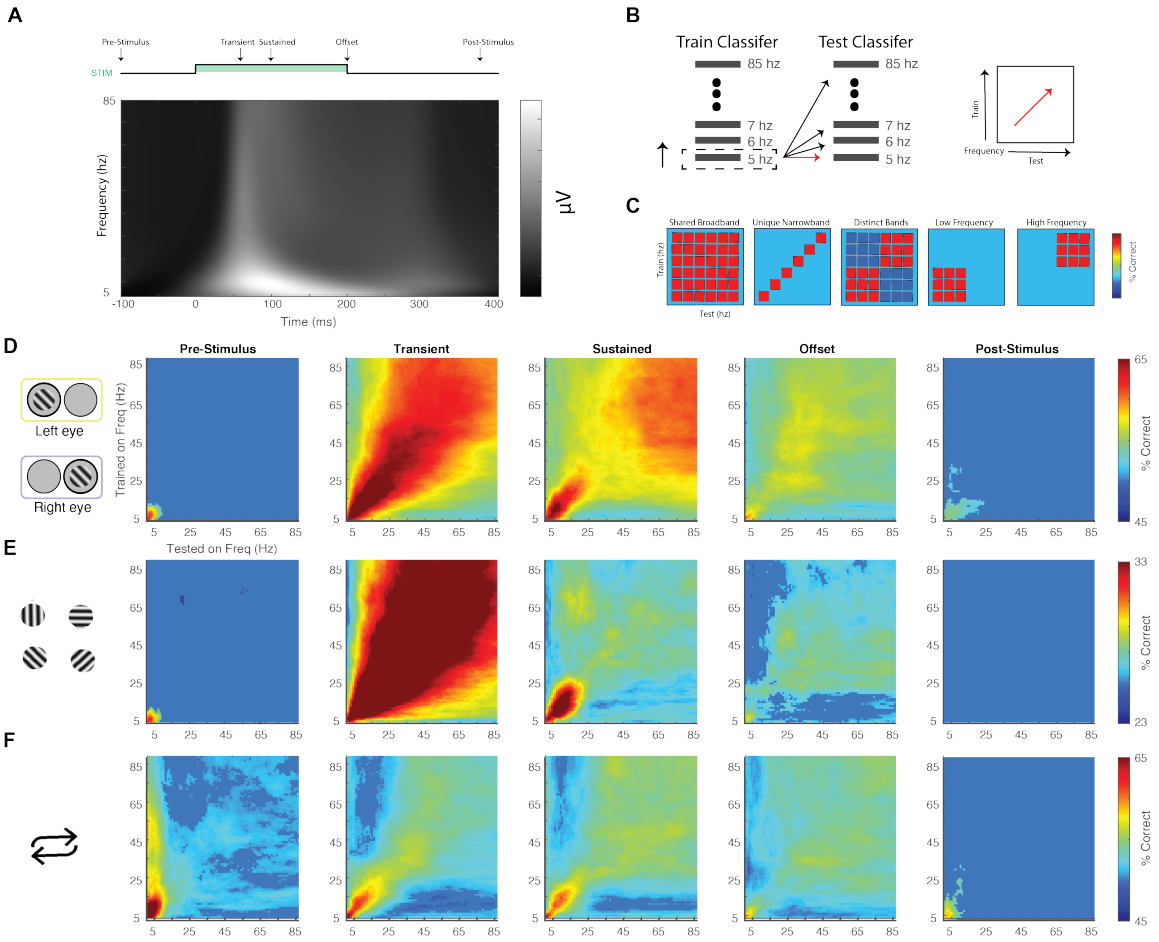

Figure 3.5: Frequency generalization of stimulus information evolving over time. (A) CSD full wave rectified time frequency spectrogram of mean stimulus locked responses. Stimulus timeline is shown above for reference, indicating five key timepoints of interest (B) Schematic of time frequency generalization procedure. An LDA classifier was iteratively trained on a frequency bin and tested on all frequency bins, repeating the process until all frequency bins are used for training and testing to create a frequency generalization matrix. This procedure was done for the five key timepoints of interest. (C) Cartoon models of possible results. (D-F) Frequency generalization matrices FDR corrected for multiple comparisons, $q<0.05$, for: (D) Eye-of-origin, (E) Orientation, (F) Stimulus repetitions.

3.4.3 Relative information found in LFP and LFP_{Cal} signals vary by stimulus features across frequency bands

We next directly tested how volume conduction effects varied across stimulus features along different frequency bands. Using the LFP Power and LFP_{Cal} Power signals (Figure 3.2C), we once again employed a moving searchlight analysis to construct spatiotemporal maps of stimulus features. All results were thresholded by significance, FDR corrected across electrodes and time, $q<0.05$. Beginning at theta (Figure 3.6A), a frequency band associated with feedforward activity (Bastos et al., 2015), there was more information in the localized signal in cortical layers associated with the initial feedforward volley. For eye-of-origin and orientation, these features first emerge in the granular layer (Casagrande & Boyd, 1996; Hubel & Wiesel, 1972), while stimulus history first emerges prominently in the supragranular layer (Tovar et al., 2020; Westerberg et al., 2019). During the sustained response, we noticed that this pattern changed for eye-of-origin and orientation with more information contained in the LFP signal than LFP_{Cal} signal outside of the initial feedforward volley – supragranular and infragranular layers. Meanwhile, there is more stimulus history information in the LFP signal within the granular and infragranular layer both before and after stimulus presentation. Given that there may be information regarding stimulus history, prior to presentation, it is not surprising to find significant decoding. However, we do note that there is some temporal imprecision with significant decoding for eye-of-origin and orientation, but the relative temporal relationship between layers and between the LFP and LFP_{Cal} signals is preserved. In total, we find that in theta, layers that are not associated with the initial volley of sensory information are the layers in which volume conduction effects predominate.

For alpha (Figure 3.6B), there was more eye-of-origin information found in the localized LFP_{Cal} signal, for the initial feedforward sweep in the granular layer as well as the supragranular layer for the sustained response. These results are consistent with recent work (Gieselmann & Thiele, 2020) showing alpha is associated with feedforward in ad-

dition to the more commonly associated feedback processing (Buffalo, Fries, Landman, Buschman, & Desimone, 2011; Van Kerkoerle et al., 2014) depending on the stimulus feature. By removing volume conduction, the eye-of-origin information in the feedforward sweep becomes more apparent. For orientation, there was more localized information along the transient, spread out equally across layers during the initial feedforward sweep. These results suggest that the volume conducted signal adds noise for orientation information, as decoding improves for the reduced LFP_{Cal} signal. For stimulus history, the LFP signal predominates throughout the spatiotemporal map. Overall, these results demonstrate the utility of the LFP_{Cal} in highlighting the role alpha may have in feedforward processing for select stimuli features.

In beta (Figure 3.6C), the LFP_{Cal} signal from the supragranular and infragranular layers near the 4/5 border contained more eye-of-origin information during the sustained response, mirroring what was observed alpha frequency. Similarly, but to a much lesser extent, low gamma and high gamma eye-of-origin information showed differences during the sustained response (Figure 3.6D-E). Low gamma, but not high gamma showed more information in the LFP signal than the LFP_{Cal} signal. Meanwhile, differences in orientation information between the LFP and LFP_{Cal} were minimal from beta to high gamma (Figure 3.6C-F). For stimulus history, there was more information found in the LFP signal than LFP_{Cal} signal from beta to high gamma (Figure 3.6D-E), but interestingly became increasingly localized to the infragranular layer at higher frequencies. These results are consistent with previous studies that have found that volume conduction effects are most prominent in the infragranular layers (Kajikawa & Schroeder, 2015). Overall, across frequencies, different spatiotemporal profiles emerged for the LFP signal and CSD derived LFP_{Cal} signal depending on the stimulus feature, in turn leading to different degrees of volume conduction.

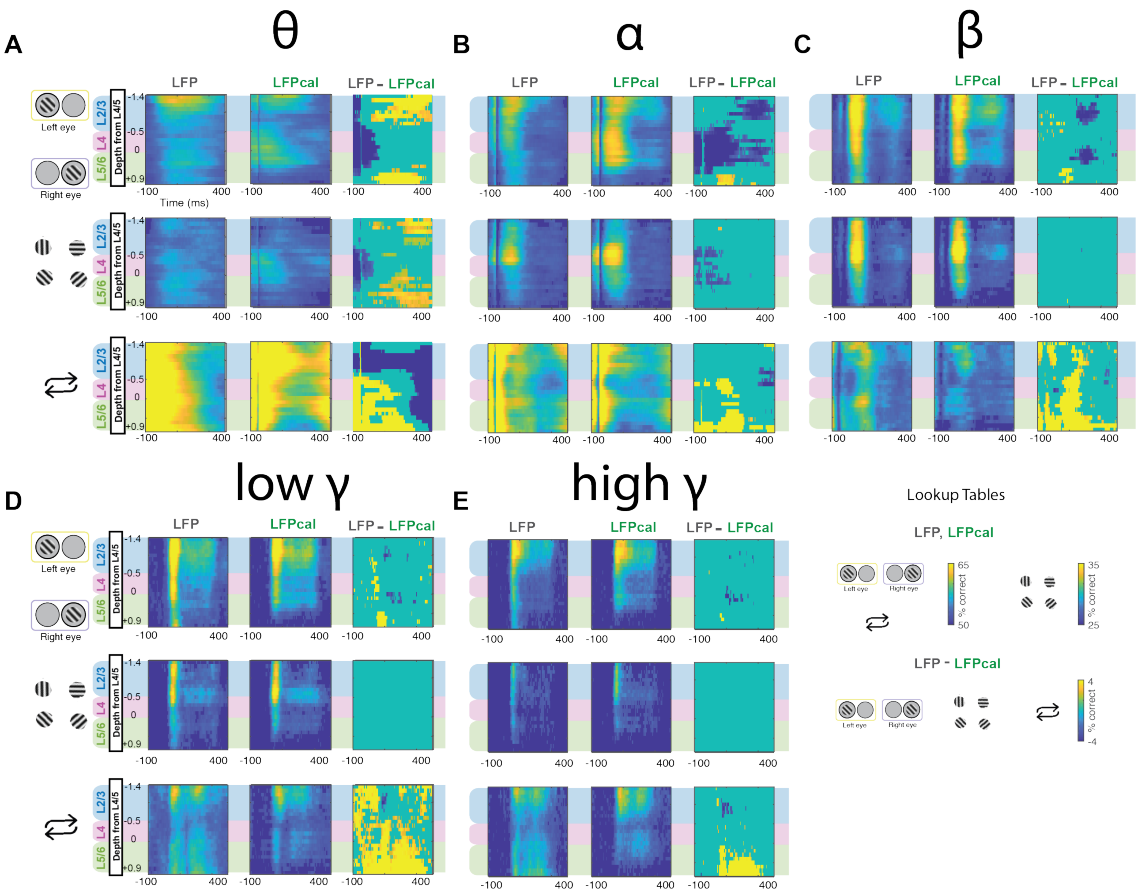

Figure 3.6: Relative information found in LFP and LFP_{Cal} signals vary by stimulus features across frequency bands. A 3-electrode searchlight along the laminar probe was used to decode Eye-of-Origin, Orientation, and Stimulus Repetitions for (A) theta 4-8 hz, (B) alpha 8-15hz, (C) beta 15-30hz, (D) low gamma 30-60hz, and (E) high gamma 60-100hz. LFP, LFP_{Cal}, and difference maps tested for significance from chance decoding, FDR corrected, q=0.05.

3.5 Discussion

In this study, we used MVPA to extract information regarding stimulus features from volume conducted LFP and the localized LFP_{Cal} signal. By analyzing more than just the magnitude of these respective signals, we were able to show that added volume conduction in the LFP signal reduced the amount of information found for some stimuli features but enhanced information in other features. We then explored how stimulus feature information might vary along frequency in local signals, finding different stimulus features exhibited drastically different patterns of shared information in the time course of a stimulus presentation. In turn, we showed that the stimulus feature differences in frequency led to differing volume conduction effects across frequency bands. Together, these results provide an analytical framework that in addition to informing investigation in V1, can be applied to any brain area to decipher whether feature information is contained to a local circuit or outside of it.

Within the literature there's been a longstanding debate about how localized the local field potential is (Kajikawa & Schroeder, 2011, 2015; Katzner et al., 2009; Mineault et al., 2013; Xing et al., 2009). The results of our work suggest that perhaps the question that we should be asking is what type of stimulus information is local within the LFP. This question becomes increasingly important when we consider that both experimental and modeling studies have shown that specific properties of the LFP signal, primarily correlation of the synaptic inputs, can have an order of magnitude difference in the degree of volume conduction (Buffalo et al., 2011; Leski et al., 2013; Lindén et al., 2011; Rosenbaum et al., 2017). Not surprisingly, one of the biggest factors that influences whether the synaptic inputs are correlated is the content of the stimulus being processed (Peter et al., 2019). Beyond determining the extent by which the signal is able to physically propagate, the specific stimulus being processed determines which parts of the brain are active. The relative distribution of brain activity can considerably influence volume conduction effects, as volume conduction is most prominently observed when distant signals are stronger than

the local signals from the recording site (Kajikawa et al., 2017). Importantly, even simple stimuli, such as the orientation gratings used in this study can contain wildly different spatiotemporal feature maps within a cortical column (Tovar et al., 2020) and may conceivably vary even more across brain areas. As the complexity of stimuli increases, the number of possible stimuli features exponentially increases, and the possible different types of volume conduction across stimulus features exponentially rises with it as well. Thus, it becomes increasingly important to characterize the information present in local and distant signals.

The debate regarding the extent of volume conduction spread also extends into frequency space. Here, the question is whether volume conduction effects are spread evenly across frequencies (Kajikawa & Schroeder, 2011, 2015), or whether volume conduction is more prominently seen in some frequency bands but not others (Leski et al., 2013). Much like the broader question regarding LFP spread, the discrepancy between studies might be explained by the specific information contained within different frequency bands. Differences between frequency bands have been highlighted by a number of studies investigating feedforward and feedback activity (Bastos et al., 2015; Van Kerkoerle et al., 2014), signal synchrony along frequencies (Buffalo et al., 2011), or shared mutual information using information theory approaches (Belitski et al., 2008; Kayser, Montemurro, Logothetis, & Panzeri, 2009). However, stimuli features such as the specific size of stimuli or if attention is directed towards the stimuli, can influence synchronization within frequency bands (Buffalo et al., 2011; Ferro, van Kempen, Boyd, Panzeri, & Thiele, 2021; Gieselmann & Thiele, 2020). In the current study, we show that within stimulus presentations, shared information between frequency bands changes completely depending on both the stimulus feature and the particular time epoch within the neural response. Our analysis focused on only local processes with the assumption that if shared information between frequencies changed locally, frequency "information profiles" will change across brain areas. As a result, whether volume conduction is uniform or localized to a frequency band depends on the local brain area and surrounding neighbors processes the particular feature being studied.

A valid question that might arise from our study is why quantify information present in LFP at all? If volume conduction contaminates information present at electrodes, why not simply quantify everything using CSD or a CSD derived signal like the LFP_{Cal}. However, when volume conduction is isolated from the localized signal, it can be informative of processes that are arising in areas outside of the local circuit. In the current study, for example, we found that stimulus history information was more prominently found outside of V1 than within the V1 microcircuit. This finding is consistent with what is known about the role of many brain areas including the visual pulvinar, middle and inferior temporal gyri, and frontal gyri in repetition suppression (Kaas & Lyon, 2007; Wig, Buckner, & Schacter, 2009). However, there are a number of circumstances where potential brain areas might encode stimulus features that have not been as well studied. Recently, convolutional neural networks (CNNs) have been used to model the ventral visual stream with remarkable accuracy(Kar & DiCarlo, 2020; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Schrimpf et al., 2018; Yamins & DiCarlo, 2016). However, while we can visualize the features captured by CNN layers (Bashivan, Kar, & DiCarlo, 2019), the features are not anchored in cognitive theories and experiments specifically targeting those features. In this circumstance, isolating local signals and volume conducted signal would help guide whether a particular CNN layer matched the area being recorded or if the CNN features are likely to be found outside of the recording site.

In total, we have presented the utility of using MVPA to extract feature specific information from local and distant signals in the V1 microcircuit. The different spatiotemporal profiles between LFP and LFP_{Cal} for eye-of-origin, orientation, and stimulus history highlights the importance of accounting for possible contamination from distant signals. By focusing on stimulus features, rather than activation, our results also help reconcile the conflicting findings from previous studies quantifying volume conduction in the LFP signal as a whole, as well as how volume conduction is affected by frequency. Lastly, our study provides a method with potential practical applications. For example, if it were only

possible to record from an early sensory area and late motor area, but not an intermediate area, comparing between LFP and LFP_{Cal} would provide clues on how information transforms from early to late brain area. This added flexibility can be invaluable in understanding cognitive processes when the number of recording sites is limited by practical or theoretical constraints.

3.6 References

Asaad, W. F. and Eskandar, E. N. (2008). A flexible software tool for temporally-precise behavioral control in matlab. *Journal of Neuroscience Methods*, 174(2):245–258.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439).

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., DeWeerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401.

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2008). Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *Journal of Neuroscience*, 28(22):5696–5709.

Bertone-Cueto, N. I., Makarova, J., Mosqueira, A., García-Violini, D., Sánchez-Peña, R., Herreras, O., Belluscio, M., and Piriz, J. (2020). Volume-conducted origin of the field potential at the lateral habenula. *Frontiers in Systems Neuroscience*, 13(January).

Bijanzadeh, M., Nurminen, L., Merlin, S., Clark, A. M., and Angelucci, A. (2018). Distinct

laminar processing of local and global context in primate primary visual cortex. *Neuron*, 100(1):259–274.e4.

Buffalo, E. A., Fries, P., Landman, R., Buschman, T. J., and Desimone, R. (2011). Laminar differences in gamma and alpha coherence in the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11262–11267.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents-eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, 13(6):407–420.

Bédard, C., Kröger, H., and Destexhe, A. (2004). Modeling extracellular field potentials and the frequency-filtering properties of extracellular space. *Biophysical Journal*, 86(3):1829–1842.

Carlson, T. A., Hoogendoorn, H., Kanai, R., and Turrett, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, 11:245–258.

Casagrande, V. A. and Boyd, J. D. (1996). The neural architecture of binocular vision. *Eye*, 10(2):153–160.

Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A., and Maier, A. (2013). Receptive field focus of visual area v4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 110(42):17095–17100.

Dougherty, K., Cox, M. A., Westerberg, J. A., and Maier, A. (2019). Binocular modulation of monocular v1 neurons. *Current Biology*, 29(3):381–391.e4.

Einevoll, G. T., Pettersen, K. H., Devor, A., Ulbert, I., Halgren, E., and Dale, A. M. (2007). Laminar population analysis: Estimating firing rates and evoked synaptic activity from multielectrode recordings in rat barrel cortex. *Journal of Neurophysiology*, 97(3):2174–2190.

Etzel, J. A., Zacks, J. M., and Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, 78.

Ferro, D., van Kempen, J., Boyd, M., Panzeri, S., and Thiele, A. (2021). Directed information exchange between cortical layers in macaque v1 and v4 and its modulation by selective attention. *PNAS*, 118(12).

Gieselmann, M. A. and Thiele, A. (2020). Stimulus dependence of directed information exchange between cortical layers in macaque v1. *bioRxiv*.

Hubel, D. H. and Wiesel, T. N. (1972). Laminar and columnar distribution of geniculocortical fibers in the macaque monkey - hubel - 2004 - journal of comparative neurology - wiley online library. *Journal of Comparative Neurology*, 146(4):421–450.

Hwang, J., Mitz, A. R., and Murray, E. A. (2019). Nimh monkeylogic: Behavioral control and data acquisition in matlab. *Journal of Neuroscience Methods*, 323:13–21.

Kaas, J. H. and Lyon, D. C. (2007). Pulvinar contributions to the dorsal and ventral streams of visual processing in primates. *Brain Research Reviews*, 55(2 SPEC. ISS.):285–296.

Kajikawa, Y. and Schroeder, C. E. (2011). How local is the local field potential? *Neuron*, 72(5):847–858.

Kajikawa, Y. and Schroeder, C. E. (2015). Generation of field potentials and modulation of their dynamics through volume integration of cortical activity. *Journal of Neurophysiology*, 113(1):339–351.

Kajikawa, Y., Smiley, J. F., and Schroeder, C. E. (2017). Primary generators of visually evoked field potentials recorded in the macaque auditory cortex. *Journal of Neuroscience*, 37(42):10139–10153.

Kar, K. and DiCarlo, J. J. (2020). Fast recurrent processing via ventrolateral prefrontal

cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, pages 1–13.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983.

Katzner, S., Nauhaus, I., Benucci, A., Bonin, V., Ringach, D. L., and Carandini, M. (2009). Local origin of field potentials in visual cortex. *Neuron*, 61(1):35–41.

Kayser, C., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron*, 61(4):597–608.

King, J. R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210.

Kreiman, G., Hung, C. P., Kraskov, A., Quiroga, R. Q., Poggio, T., and DiCarlo, J. J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron*, 49(3):433–445.

Lalla, L., Rueda Orozco, P. E., Jurado-Parras, M. T., Brovelli, A., and Robbe, D. (2017). Local or not local: Investigating the nature of striatal theta oscillations in behaving rats. *eNeuro*, 4(5).

Leski, S., Lindén, H., Tetzlaff, T., Pettersen, K. H., and Einevoll, G. T. (2013). Frequency dependence of signal power and spatial reach of the local field potential. *PLoS Computational Biology*, 9(7).

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A. Y., Tal, I., Haegens, S., Melloni, L., Knight, R. T., and Schroeder, C. E. (2020). Dissociation of broadband high-

frequency activity and neuronal firing in the neocortex. *Science advances*, (August):1–13.

Lindén, H., Tetzlaff, T., Potjans, T. C., Pettersen, K. H., Grün, S., Diesmann, M., and Einevoll, G. T. (2011). Modeling the spatial reach of the lfp. *Neuron*, 72(5):859–872.

Logothetis, N. K., Kayser, C., and Oeltermann, A. (2007). In vivo measurement of cortical impedance spectrum in monkeys: Implications for signal propagation. *Neuron*, 55(5):809–823.

Maier, A., Adams, G. K., Aura, C., and Leopold, D. A. (2010). Distinct superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. *Frontiers in Systems Neuroscience*, 4.

Mineault, P. J., Zanos, T. P., and Pack, C. C. (2013). Local field potentials reflect multiple spatial scales in v4. *Frontiers in Computational Neuroscience*, 7(MAR):1–15.

Mitzdorf, U. (1985). Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and eeg phenomena. *Physiological Reviews*, 65(1):37–100.

Nauhaus, I., Busse, L., Carandini, M., and Ringach, D. L. (2009). Stimulus contrast modulates functional connectivity in visual cortex. *Nature Neuroscience*, 12(1):70–76.

Nicholson, C. and Freeman, J. A. (1975). Theory of current source-density analysis and determination of conductivity tensor for anuran cerebellum. *Journal of neurophysiology*, 38(2):356–68.

Nicholson, C. and Llinas, R. (1971). Field potentials in the alligator cerebellum and theory of their relationship to purkinje cell dendritic spikes. *Journal of neurophysiology*, 34(4):509–531.

Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., and Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature Neuroscience*, 5(8):805–811.

Peter, A., Uran, C., Klon-Lipok, J., Roese, R., Van Stijn, S., Barnes, W., Dowdall, J. R., Singer, W., Fries, P., and Vinck, M. (2019). Surface color and predictability determine contextual modulation of v1 firing and gamma oscillations. *eLife*, 8:1–38.

Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., and Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60-200 hz) in macaque local field potentials and their potential implications in electrocorticography. *Journal of Neuroscience*, 28(45):11526–11536.

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nature Neuroscience*, 20(1):107–114.

Sato, T. K., Nauhaus, I., and Carandini, M. (2012). Traveling waves in visual cortex. *Neuron*, 75(2):218–229.

Schaefer, M. K., Kössl, M., and Hechavarría, J. C. (2017). Laminar differences in response to simple and spectro-temporally complex sounds in the primary auditory cortex of ketamine-anesthetized gerbils. *PLoS ONE*, 12(8):1–28.

Schiller, J., Major, G., Koester, H. J., and Schiller, Y. (2000). Nmda spikes in basal dendrites. *Nature*, 1261(1997):285–289.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007.

Tovar, D. A., Westerberg, J. A., Cox, M. A., Dougherty, K., Carlson, T. A., Wallace, M. T., and Maier, A. (2020). Stimulus feature-specific information flow along the columnar cortical microcircuit revealed by multivariate laminar spiking analysis. *Frontiers in Systems Neuroscience*, 14(November):1–14.

Traub, R. D. and Bibbig, A. (2000). A model of high-frequency ripples in the hippocampus based on synaptic coupling plus axon-axon gap junctions between pyramidal neurons. *Journal of Neuroscience*, 20(6):2086–2093.

Van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M. A., Poort, J., Van Der Togt, C., and Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(40):14332–14341.

Westerberg, J. A., Cox, M. A., Dougherty, K., and Maier, A. (2019). V1 microcircuit dynamics: Altered signal propagation suggests intracortical origins for adaptation in response to visual repetition. *Journal of Neurophysiology*, 121(5):1938–1952.

Westerberg, J. A., Maier, A., and Schall, J. D. (2020a). Priming of attentional selection in macaque visual cortex: Feature-based facilitation and location-based inhibition of return. *eNeuro*, 7(2).

Westerberg, J. A., Maier, A., Woodman, G. F., and Schall, J. D. (2020b). Performance monitoring during visual priming. *Journal of Cognitive Neuroscience*, 32(3):515–526.

Wig, G. S., Buckner, R. L., and Schacter, D. L. (2009). Repetition priming influences distinct brain systems: Evidence from task-evoked data and resting-state correlations. *Journal of Neurophysiology*, 101(5):2632–2648.

Xing, D., Yeh, C. I., and Shapley, R. M. (2009). Spatial spread of the local field potential and its laminar variation in visual cortex. *Journal of Neuroscience*, 29(37):11540–11549.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Zanos, T. P., Mineault, P. J., Nasiotis, K. T., Guitton, D., and Pack, C. C. (2015). A sensorimotor role for traveling waves in primate visual cortex. *Neuron*, 85(3):615–627.

Part 2

"All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason."

-Immanuel Kant

Chapter 4

Selective enhancement of object representations through multisensory integration

The contents of this chapter are adapted from

Tovar, D.A., Murray, M.M., & Wallace, M.T. (2020). Selective enhancement of object representations through multisensory integration. Journal of Neuroscience, 40(29), 5604–5615.

4.1 Abstract

Objects are the fundamental building blocks with which we construct a representation of the external world. One major distinction amongst objects is between those that are animate versus inanimate. In addition, many objects are specified by more than a single sense, yet the nature by which multisensory objects are represented by the brain remains poorly understood. Using representational similarity analysis of male and female human EEG signals, we show enhanced encoding of audiovisual objects when compared to their corresponding visual and auditory objects. Surprisingly, we discovered that the often-found processing advantages for animate objects was not evident under multisensory conditions. This was due to a greater neural enhancement of inanimate objects—which are more weakly encoded under unisensory conditions. Further analysis showed that the selective enhancement of inanimate audiovisual objects corresponded with an increase in shared representations across brain areas, suggesting that the enhancement was mediated by multisensory integration. Moreover, a distance-to-bound analysis provided critical links between neural findings and behavior. Improvements in neural decoding at the individual exemplar level for audiovisual inanimate objects predicted reaction time differences between multisensory and unisensory presentations during a go/no-go animate categorization task. Links between neural activity and behavioral measures were most evident at intervals 100-200ms and 350-500ms after stimulus presentation, corresponding to time periods

associated with sensory evidence accumulation and decision-making, respectively. Collectively, these findings provide key insights into a fundamental process the brain uses to maximize information it captures across sensory systems to perform object recognition.

4.2 Introduction

The brain is constantly bombarded with sensory information, much of which is combined to form building blocks of our perception representation of the external world. Previous multisensory literature has shown that the brain tends to optimally combine sensory information when the information between senses is equally reliable (Ernst & Banks, 2002). Furthermore, prior work has shown that the maximum gains from multisensory integration are seen when responses to the individual senses are weak (Stein & Meredith, 1993; Wallace, Ramachandran, & Stein, 2004). In large measure, these studies have focused on manipulating stimulus reliability and effectiveness through changing low-level stimulus features, such as introducing differing levels of noise, to gauge the effects on multisensory integration. However, emerging literature in vision and audition suggests that higher-level semantic features, such as the binding of stimulus elements into objects, may also play a key role in dictating reliability and effectiveness (Cappe, Thelen, Romei, Thut, & Murray, 2012; Ritchie, Tovar, & Carlson, 2015). Given that many objects are specified through their multisensory features, an open question is how might differences in object categorization lead to differences in perceptual gains from multisensory integration.

One of the major categorical distinctions between objects is animacy. In vision, animate objects offer substantial processing and perceptual advantages over inanimate objects, including being categorized faster, more consciously perceived, and found faster in search tasks (Carlson et al., 2014; Jackson & Calvillo, 2013; Lindh, Sligte, Assecondi, Shapiro, & Charest, 2019; New, Cosmides, & Tooby, 2007; Ritchie et al., 2015). Auditory studies have similarly found faster categorization times for animate objects (Vogler & Titchener, 2011; Yuval-Greenberg & Deouell, 2009). This difference may be a remnant of an evolutionary

need to rapidly recognize and process living stimuli that could pose threats or be sources of sustenance (Laws, 2000). Furthermore, many inanimate objects such as cars, trains, and cellphones have not existed long enough for there to be specialized brain areas to represent them. In contrast, a number of specialized areas exist for the processing of categories of animacy, such as faces in the fusiform face area (FFA), bodies in the extrastriate body area (EBA) and voices in the temporal voice areas (TVAs) (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; De Lucia, Clarke, & Murray, 2010; Downing, Jiang, Shuman, & Kanwisher, 2001; Kanwisher, McDermott, & Chun, 1997).

To study how perceptual differences in visual and auditory categories influence their subsequent integration as audiovisual objects, it is critical to quantify neural encoding differences between objects. Representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008) constructs a representational space quantifying relationships between stimuli with representational distance indicating the difference in their neural signatures. A greater distance in representational space signifies more distinct neural signals between stimuli, while shorter distances signify less distinct neural signals. Studies using RSA have shown that visual and auditory objects have a clear encoding distinction between animate and inanimate categories (Cichy, Pantazis, & Oliva, 2014; Giordano, McAdams, Zatorre, Kriegeskorte, & Belin, 2013; Kriegeskorte, Mur, Ruff, & Kiani, 2008), while also showing that representational space can contract if stimuli are degraded (Grootswagers, Ritchie, Wardle, Heathcote, & Carlson, 2017) or expand in cases of increased attention (Nastase et al., 2017). Although RSA has been increasingly used to study object representations, it has not been fully leveraged to examine objects as they are often represented in naturalistic setting – as multisensory entities.

In this study, we presented subjects with auditory, visual, and semantically congruent audiovisual animate and inanimate objects while we recorded high-density EEG. Our overarching hypothesis was that greater behavioral benefits would be seen for objects specified in a multisensory manner and that these gains would be accompanied by an expansion in

representational space as measured using RSA. A secondary hypothesis was that greater benefits would be observed for inanimate objects, given evidence that multisensory integration benefits are greatest for weakly effective stimuli (Ernst & Banks, 2002; Stein & Meredith, 1993; Wallace et al., 2004)

4.3 Methods

4.3.1 Participants

The experiment included 14 adults (9 males) aged 27± 4.2 years. All subjects had normal or corrected-to-normal vision and reported normal hearing. The study was conducted in accordance with the Declaration of Helsinki, and all subjects provided their informed consent to participate in the study. Each participant was compensated financially for their participation. The experimental procedures were approved by the Ethics Committee of the Vaudois University Hospital Center and University of Lausanne. Behavioral data for all subjects was used. However, EEG data for one subject was removed from further decoding analysis due to poor signal quality in the evoked potential response. Stimuli The experiment took place in a sound-attenuated chamber (Whisper room), where subjects were seated centrally in front of a 20" computer monitor (HP LP2065) and located 140 cm away from them (visual angle of objects 4°). The auditory stimuli were presented over insert earphones (Etymotic model: ER4S), and the volume was adjusted to a comfortable level (62dB). The stimuli were presented and controlled by E-Prime 2.0, and all behavioral data were recorded in conjunction with a serial response box (Psychology Software Tools, Inc.; www.pstnet.com). The auditory stimuli included 48 animate and 48 inanimate sounds from a library of 500ms-duration sounds, used in previous studies and have been evaluated in regard to their acoustics and psychoacoustics as well as brain responses as a function of semantic category (De Lucia et al., 2010; Murray, 2006; Thelen, Cappe, & Murray, 2012). The visual stimuli were semantically congruent line drawings that were taken from a standardized set (Snodgrass & Vanderwart, 1980) or obtained from an online

library (dgl.microsoft.com).

4.3.2 Experimental Design

Participants performed 10-13 experimental blocks (median 10 blocks) of a Go/No-Go task. Each block contained 1 audio, visual, and audiovisual presentation for each of the 96 stimuli exemplars, totaling 288 stimulus presentations per block. For half of the blocks, subjects were instructed to press a button when they perceived an animate object and for the other half when they perceived an inanimate object. Animate and inanimate blocks were randomized for each subject. Auditory, visual, and synchronous audiovisual stimuli were presented for 500ms, followed by a randomized interstimulus interval (ISI) ranging from 900 to 1500ms, and participants had to respond within this 1.4-2s window. Stimuli modality was randomized for each trial (see Figure 4.1 for schematic). To control for motor confounds, the block instructions alternated between indicating whether the stimuli was animate or inanimate (Grootswagers, Wardle, & Carlson, 2017). Reaction times and accuracy were measured for each response. Participants did not receive feedback during the experiment.

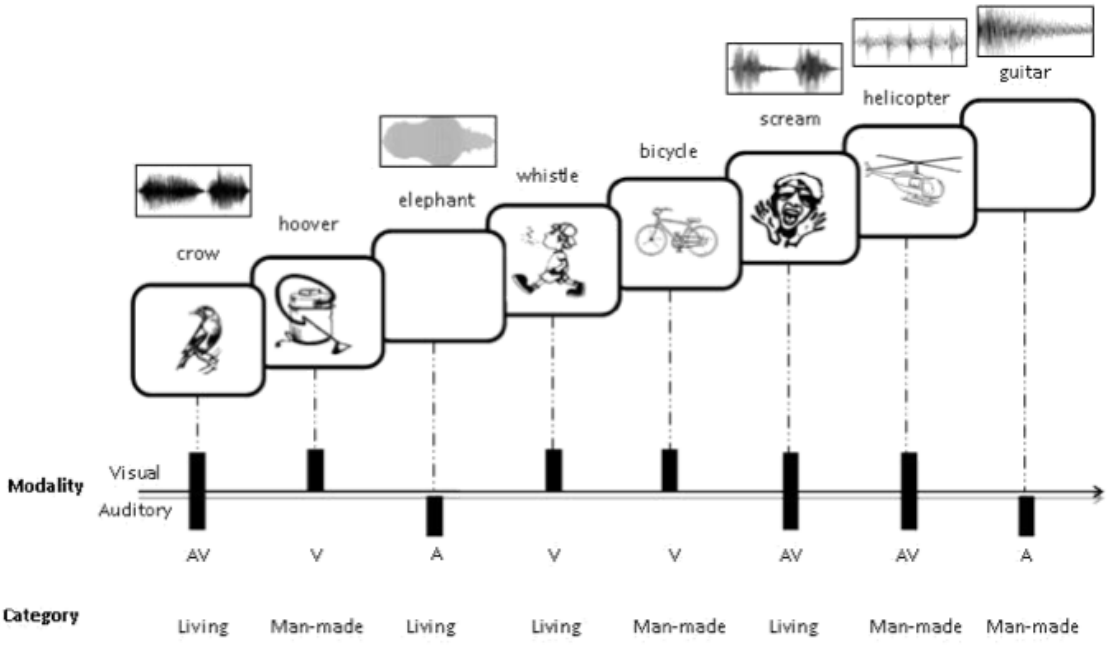

Figure 4.1: Experiment Schematic. A Go/No-Go discrimination task of animate and inanimate objects. The responses were counterbalanced such that the number of responses for animate and inanimate objects was equivalent. The stimuli consisted of 96 visual line drawings and 96 environmental sounds of common animate and inanimate objects, as well semantically congruent pairings of these objects. The sounds of animate object were non-verbal vocalizations. The stimulus duration was 500ms with a variable inter-stimulus interval of 900-1500ms.

4.3.3 Statistical Inference

All statistical inference for behavior and neural data was assessed with Bayes factors (Jeffreys, 1998; Wetzels et al., 2011) using a JZS prior (Rouder, Speckman, Sun, Morey, & Iverson, 2009), with a scale factor of 0.707. For decoding analysis, chance level decoding was estimated by randomly shuffling all trial labels for each subject once prior to classification to construct a null distribution. The probability of the group data assuming the alternative hypothesis relative to the probability of group data assuming chance level decoding was computed to calculate a Bayes Factor at each time point. Bayes Factors provide the added advantage over frequentist inference because in addition to rejecting the null hypothesis, they can provide support for the null hypothesis as well as determine whether the data is insensitive, and as a result help avoid overstating the evidence against the null hypothesis (Berger & Delampady, 1987; Edwards, Lindman, & Savage, 1963; Johnson, 2013; Sellke, Bayarri, & Berger, 2001). The theoretical differences underlying Bayesian and frequentist analysis have spurred debate on whether and how Bayes factors should be corrected for multiple comparisons (Berry & Hochberg, 1999), since they intrinsically already reduce type I errors (Gelman, Hill, & Yajima, 2012; Gelman & Tuerlinckx, 2000; Johnson, 2013). In this study, we report Bayes Factors without additional multiple comparison correction, but provide Bayes factors with varying levels of evidence, consistent with recent EEG decoding studies (Grootswagers, Robinson, & Carlson, 2019; Robinson, Grootswagers, & Carlson, 2019). Using Jeffreys' scheme, Bayes factors > 3 and >10 indicate substantial and strong evidence for the alternative hypothesis respectively, anything between 3 and 1/3 indicates insufficient evidence, and Bayes factors less than 1/3 and 1/10 indicate substantial and strong evidence for the null hypothesis (Jarosz & Wiley, 2014; Jeffreys, 1998). We further compared Bayes Factors with a cluster-based sign permutation test (Maris & Oostenveld, 2007) and found Bayes Factors to be more conservative. Therefore, we report only Bayes Factors in the Results.

4.3.4 EEG acquisition and preprocessing

Continuous EEG was acquired from 160 scalp electrodes (sampling rate at 1024 Hz) using a Biosemi ActiveTwo system. Data preprocessing was performed offline using the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011) in MATLAB. Data were filtered using a Butterworth IIR filter with 1 Hz highpass, 60 Hz lowpass, and notch at 50Hz. All channels were rereferenced to an average reference. Epochs were created for each stimulus presentation ranging from -100ms to 600ms relative to stimulus onset. Each epoch was baseline corrected using the prestimulus period.

4.3.5 Representational Similarity Analysis

Following data preprocessing, we used CoSMoMVPA (Oosterhof, Connolly, & Haxby, 2016) and custom scripts to perform cross-validated representational similarity analysis (RSA). We used a linear discriminate classifier after default regularization (0.01) with 4-fold, leave one-fold out cross validation, for all exemplar pair combinations across audio, visual, and audiovisual stimuli presentations. In this procedure, trials are randomly assigned to one of four subsets of data. Three of the four subsets (75% of the data) are then pooled together to train the classifier and then decoding accuracy is tested on the remaining subset (25% of the data). This procedure is repeated a total of four times, such that each of the subsets is tested at least once. Decoding results are reported in percent correct of classifications at each time point for each exemplar pair in the time series [-100ms 600ms]. This analysis was conducted independently to build representational dissimilarity matrices (RDM) for each subject and modality over 1 millisecond increments. The RDMs were then separated into animate exemplar pairwise comparisons, inanimate exemplar pairwise comparisons, and pairwise comparisons between categories. Using these comparison groupings, mean decoding accuracies were then calculated for each modality and subject. Significant above-chance accuracies were assessed against a randomized trial shuffle control using Bayes factors.

4.3.6 Representational Connectivity Analysis

To characterize connectivity changes for different modalities and object categories, we used a combination of a searchlight analysis and representational connectivity analysis (Kriegeskorte et al., 2008). Due to this analysis being computationally-intense, data was downsampled to 100 Hz. Electrode specific RDMs, using the same procedure describe for the RSA analysis, were built by using a moving searchlight which included the electrode of interest and every immediate adjacent electrode. Depending on the location of the electrodes, the RDMs can potentially be more descriptive of lower-level properties of the stimuli or contain higher-level object category information. Importantly, the analysis is not designed to distinguish between any particular stimulus dimension, such as animacy, but rather used to calculate the local representational geometry present at those electrodes. Electrode-specific RDMs were then correlated to each other in pairwise fashion for each electrode combination using a Spearman correlation to form a matrix of RDM correlations between electrodes. We then averaged the Spearman correlations from across all electrode comparisons to compute a mean connectivity measure. If the representational geometry is distributed across several electrodes, then the expectation is that this value would increase and if it is unique to a particular electrode, this value would decrease. This analysis was performed for visual, auditory, and audiovisual presentations. Additionally, to compare the audiovisual response to the visual and auditory response more directly, we also summed evoked responses for auditory and visual presentations for each specific exemplar and performed RCA on these trials. Note that in this calculation, the searchlight will change sizes depending on the chosen electrode and searchlights will overlap for electrodes leading to a non-zero baseline level of connectivity in neighboring RDMs, regardless of the evoked responses to stimulus presentations. Therefore, we repeated the analysis above, but shuffled all of the exemplar labels when calculating the RDMs to create a shuffled control. All connectivity measurements were compared to their respective shuffled labels control. This procedure was done for all exemplars as well as within the animate and inanimate

category along the timeseries [-100ms 600ms] to compute time-resolved representational connectivity measures.

4.3.7 Distance to Bound Analysis

To link neural representational space back to individual exemplar categorization times, we used a distance to bound analysis (for review see Ritchie & Carlson, 2016). Similar to RSA, this analysis represents individual exemplars as points in representational space. A decision boundary for animacy is then fitted using a linear discriminant analysis classifier to the representational space, defining an optimal decision boundary that separates animate and inanimate exemplars. The distance to the decision boundary is determined for each exemplar and subsequently pooled and averaged across subjects to calculate average exemplar distance across subjects for each timepoint in the timeseries [-100ms 600ms]. Next, the median exemplar reaction time, pooled across subjects, is calculated for each exemplar. We then performed a time-varying Spearman correlation between mean exemplar distance and median exemplar reaction time for each modality using a fixed-effects analysis to reduce noise and improve statistical power. The distance to bound analysis was performed across all electrodes as well as on an electrode by electrode basis using a moving searchlight.

4.3.8 Model Fitting

To account for low level visual features in our visual and auditory stimuli, we constructed model RDMs and calculated their correlations to electrode specific RDMs and the neural RDM from all electrodes. The low-level feature auditory RDM was constructed using a Welch's power spectral density (PSD) estimate for each of the 96 sounds. The resulting stimulus PSD was then organized into vectors and pairwise non-parametric spearman distance measurements were calculated for all exemplar pair combinations to form a model RDM. We then calculated the Spearman correlations between the PSD model RDM and the modality specific neural RDMs at each timepoint. An identical procedure was followed for the visual images, but instead of using PSD, image contrast was used. Note that since the

images were black and white Snodgrass images, the contrast values will be equivalent to the image intensity values. In addition to these low-level feature models, we also constructed an abstract animacy category model. The animacy category model was constructed using a 0 to indicate no differences between stimuli pairs for within animacy category exemplars and a 1 to indicate complete dissimilarity for between category exemplars. This model was then also tested across modality specific neural RDMs.

4.4 Results

4.4.1 Behavior: Advantage for Animate Objects on Unisensory but not Multisensory (i.e., Audiovisual) Presentations

Subjects were shown 48 animate and 48 inanimate auditory, visual, and audiovisual objects while they performed a go/no-go categorization task, as shown in Figure 4.1. Subjects performed near ceiling on the categorization task for objects presented in both visual (animate: 98%, inanimate: 98%) and audiovisual contexts (animate: 98%, inanimate 99%), and were less accurate for auditory presentations (animate: 86%, inanimate 87%). A two-way repeated measures ANOVA for accuracy revealed a main effect for modality $F(2,26) = 27.14$, $p=0.00$, but no main effect for animacy $F(1,26) = 0.64$, $p = 0.44$.

When examining reaction times (RTs), a two-way repeated measures ANOVA revealed main effects for modality, $F(2,26) = 238.18$, $p = 0.00$, and animacy, $F(1,26) = 10.39$, $p = 0.01$, as well as an interaction effect $F(2,26) = 3.68$, $p = .04$. We then performed post-hoc tests across sensory modalities and categories, as shown in Figure 4.2. Figure 4.2A shows median RTs for the go/no-go task across participants for the three sensory conditions. Using Bayes factors to compare median RTs across subjects, we found very strong evidence (B.F. > 30) that the auditory condition was slower than the visual and audiovisual conditions. Next, behavior was split by animate and inanimate categories to investigate the effects of animacy on RTs. Figure 4.2B shows that there was strong evidence (B.F. > 10) for faster RTs for animate objects compared to inanimate objects when presented in either

the auditory or visual modalities, consistent with the results from previous studies (Carlson et al., 2014; Murray, 2006; Vogler & Titchener, 2011; Yuval-Greenberg & Deouell, 2009). However, there was inconclusive evidence (B.F. = 0.75) for the audiovisual condition.

To further investigate this surprising lack of a difference in audiovisual performance, we created an index of sensory bias for each participant, operationalized as the difference in reaction times to the auditory and visual stimuli, and correlated this bias score to audiovisual RTs on a subject-by-subject basis using a Spearman correlation. Figure 4.2C shows that the only significant correlation between sensory bias and audiovisual RTs was for inanimate objects. The positive correlation indicates that subjects whose RTs for visual and auditory stimuli were more similar had faster multisensory RTs. Note, that these correlations included all subjects, since there were no outliers for sensory bias or audiovisual reaction times.

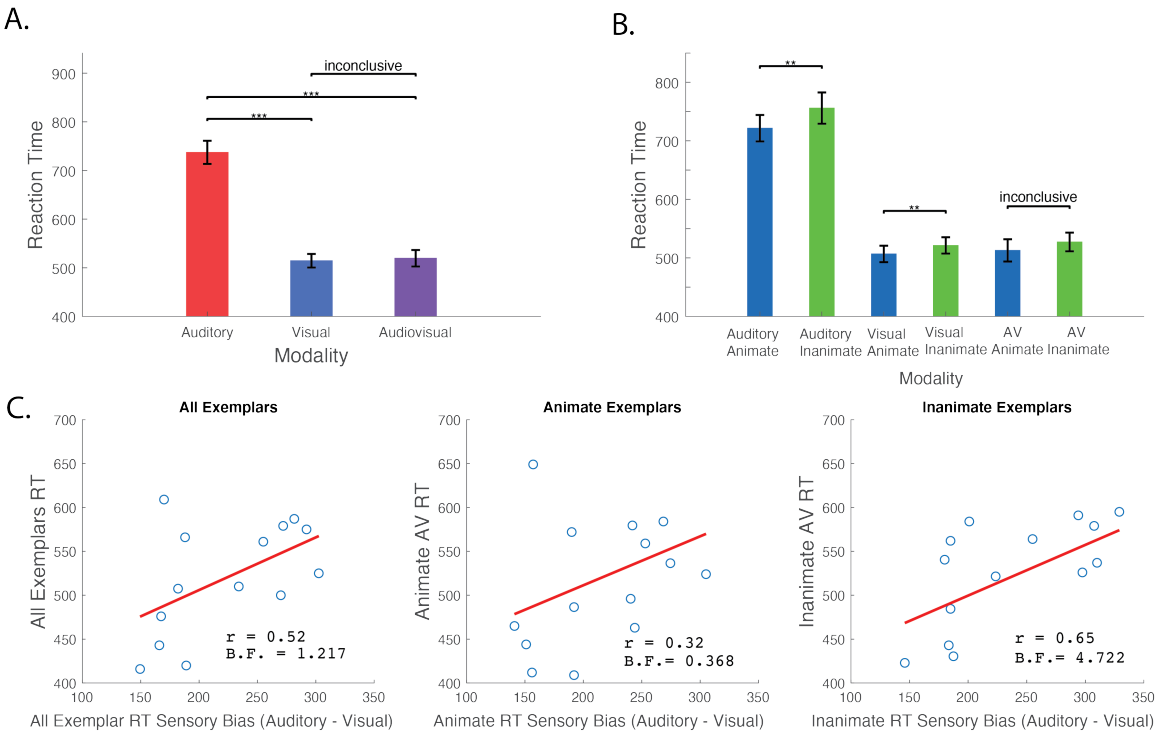

Figure 4.2: Behavior: Advantage for Animate Objects for Unisensory Presentations but not Audiovisual Presentations. (A) Reaction time (RT) results for each modality and (B) broken down by animacy. Bayes factors for substantial evidence (* B.F. > 3), strong evidence (** B.F. > 10) and very strong evidence (*** B.F. > 30) above comparisons. (C) Subject sensory bias and audiovisual RT Pearson correlations across subjects for all exemplars, only animate exemplars, and only inanimate exemplars. Sensory bias is only significantly correlated to audiovisual RT for inanimate exemplars (B.F. > 3).

4.4.2 Representational Similarity Analysis: The Influence of Sensory Modality on Between and Within Animacy Category Decoding

To investigate the neural correlates of the behavioral differences noted across conditions, we used RSA (Figure 4.3A-4.3C). Specifically, we built representational dissimilarity matrices (RDM) for each subject and modality over 1 millisecond intervals using linear discriminant analysis for each exemplar pair. From each RDM, we explored the effect of sensory modality on the distinction between animate and inanimate exemplars by calculating the mean pairwise decoding for between category pairs (e.g., dog vs. bell, dog vs. cannon). As can be seen in figure 4.3D, prior to stimulus onset, decoding is close to the shuffled label control at chance level (i.e., 50%), because the classifier does not have any meaningful neural data that will distinguish between category pairs. However, shortly after stimulus onset, decoding performance becomes significantly above the shuffled label control (B.F. > 3) across all three modalities. The latency of the onset of these decoding differences, defined as at least 20ms of sustained significant decoding (see Carlson, Tovar, Alink, & Kriegeskorte, 2013), was 183 ms for auditory, 91 ms for visual, and 65 ms for audiovisual stimulus conditions. Visual and audiovisual decoding peaked at 162ms and 154ms, respectively, with higher absolute peak decoding for audiovisual presentations (61%) compared to visual presentations (58%). Decoding of auditory stimuli was comparatively poorer, peaking at 53% at 190 ms. Note that while there were differences in significant decoding onsets, caution should be taken when comparing decoding onsets across conditions with different maximum decoding peaks (see figure 14 in Grootswagers, Wardle, & Carlson, 2017). Collectively, the results of these decoding analyses illustrate the temporal emergence of distinct neural representations for auditory, visual and audiovisual objects when subjects are performing an animacy/inanimacy categorization.

To statistically compare decoding performance across modalities, we computed the mean decoding for the interval spanning 50 to 500ms post-stimulus presentation. When comparing mean decoding values across subjects, audiovisual stimuli was significantly

higher when compared with both visual and auditory decoding (B.F. >30) and visual decoding was higher than auditory decoding. These modality focused RSA results suggest that the audiovisual presentation of an object creates a more distinct representation between animate and inanimate objects when compared to either of the corresponding unisensory presentations.

We further explored whether audiovisual presentations expanded exemplar distinctions within animacy categories by calculating the mean within category pairwise decoding accuracies (Figure 4.3E). In this analysis, onset latencies for significant decoding for auditory, visual, and audiovisual stimuli were 184 ms, 91 ms, and 79 ms, respectively. The corresponding peak decoding latencies were 189 ms, 139 ms, and 152 ms. The modality-specific comparisons for within-category decoding mirrored those seen for between-category decoding, with higher audiovisual decoding when compared with visual and auditory decoding, and higher visual decoding than auditory decoding (B.F. > 30). A comparison of between-category decoding and within-category decoding demonstrated higher between-category decoding for auditory, visual and audiovisual stimulus presentations (B.F. > 3) during the stimulus period [50-500ms]. In sum, when compared to unisensory presentations, audiovisual stimulus presentations not only expand the representational space between animacy categories, but also make exemplars within the animacy categories easier for a classifier to distinguish.

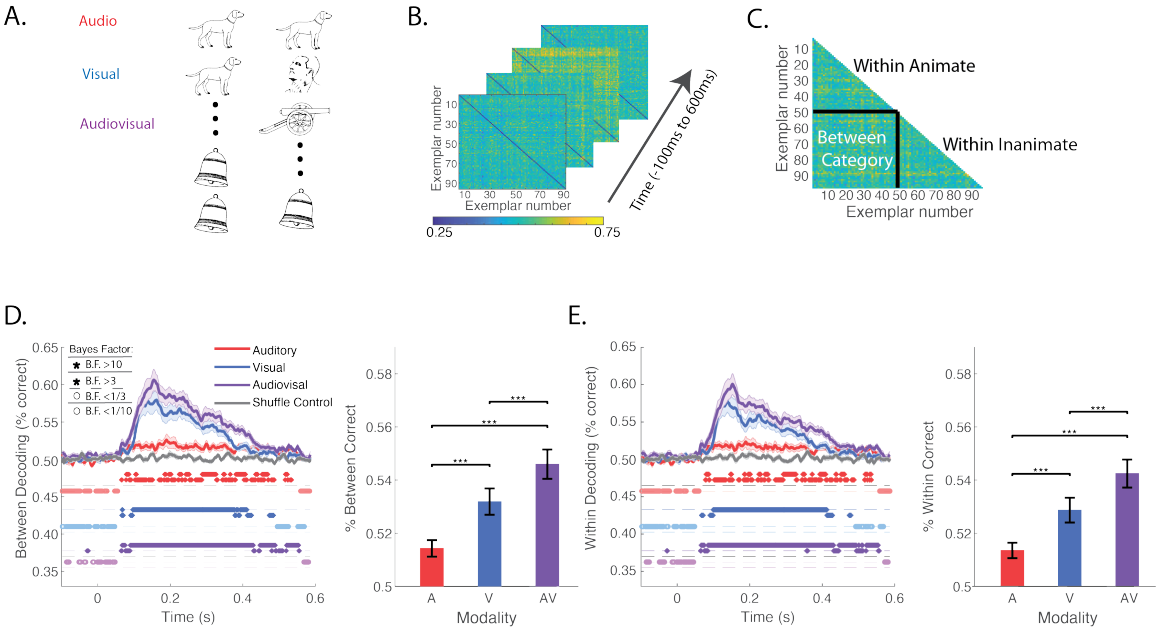

Figure 4.3: Representational Similarity Analysis: Sensory Modality Influences Between Animacy Category and Within Animacy Category Decoding. (A) RSA Schematic for pairwise decoding. Linear discriminate analysis with a 4-fold leave one-fold out cross validation was used for all exemplar pair combinations (B) Dissimilarity matrices for each of the modalities was built across time in 1 millisecond increments from pairwise exemplar classifications. (C) Mean between category and within category exemplar decoding accuracies were averaged across exemplars at each time point. (D-E) Resulting time series and summary bar plots for (D) between and (E) within categories for each of the modalities. Shaded area around lines indicates standard error across subjects. Asterisks indicate thresholded Bayes Factors for alternative and null hypothesis (see inset). Mean decoding across time (50ms to 500ms) for each modality with Bayes factors for substantial evidence (* B.F. > 3), strong evidence (** B.F. > 10) and very strong evidence (*** B.F. > 30) above comparisons.

4.4.3 Category-Specific RSA: Audiovisual Presentations Selectively Enhance Inanimate Object Decoding

We further investigated representational space broken down by animacy categories to study the neural underpinnings for the observed reaction time differences between animate and inanimate categorization (Figure 4.4). The decoding curves for animate and inanimate exemplars did not differ for auditory conditions (Figure 4.4A) with evidence for the null hypothesis present throughout the timecourse. However, this was not the case for visual exemplars, which have higher decoding performance for animate exemplars when compared with inanimate exemplars from 160 to 184 ms and from 220 to 228 ms after stimulus presentation (Figure 4.4B). Surprisingly, this difference is no longer apparent for audiovisual conditions with in fact a few sporadic timepoints with substantial evidence (B.F. >3) that inanimate objects have higher decoding than animate objects.

Since the audiovisual condition had overall higher within category pairwise decoding than the visual condition (Figure 4.3E), we additionally wanted to explore whether the lack of an animate and inanimate within-category decoding difference for audiovisual presentations was due to visual inanimate objects incurring a special benefit from audiovisual presentation. Figure 4.4D shows the difference between audiovisual decoding and visual decoding for animate and inanimate exemplars. Notably, the difference is significantly above a shuffle control subtraction of visual and audiovisual presentations for a sustained period of time extending from 137-216 ms post stimulus onset for inanimate objects (B.F. > 3) but is much sparser for animate objects without a significant sustained difference ever exceeding 20 ms

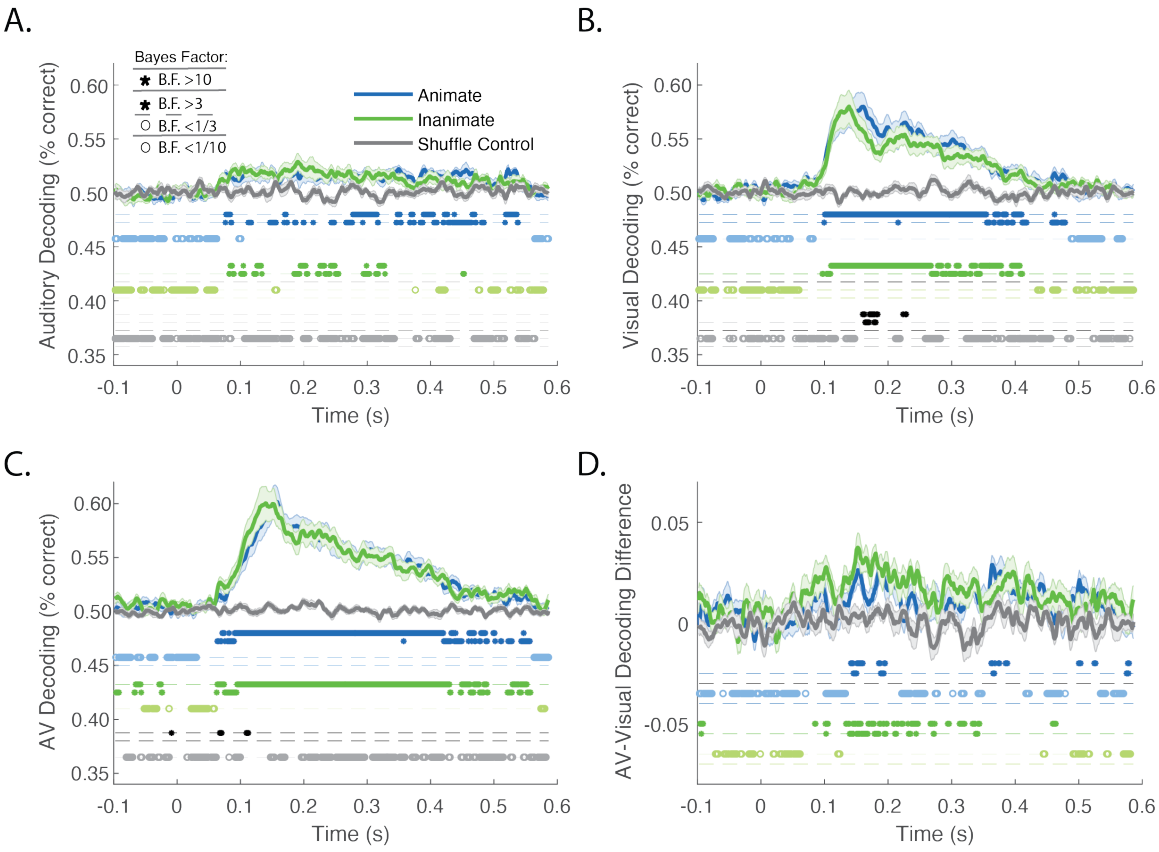

Figure 4.4: Category-Specific RSA: Audiovisual Presentations Selectively Enhance Inanimate Object Decoding. (A-C) Audio, visual and audiovisual within-category decoding for animate and inanimate exemplars. Colored asterisks indicate substantial evidence and strong evidence (see inset) compared to the shuffled control, while black asterisks indicating substantial and strong evidence for a difference between animate and inanimate objects. (D) The audiovisual-visual within category decoding difference for animate and inanimate exemplars with asterisks indicating evidence (see inset) for differences from the shuffle control.

4.4.4 Representational Connectivity Analysis: Response Patterns between Areas in the Brain are Influenced by Modality and Object Category

Given that different sensory modalities and different object classes have been shown to engage different brain networks (Braga, Hellyer, Wise, & Leech, 2017; Hillebrandt, Friston, & Blakemore, 2014), we investigated whether the pairwise decoding differences we found using RSA would also be associated with differences in mean connectivity. To carry out this analysis, we constructed electrode specific representational dissimilarity matrices (RDMs) and performed Spearman correlations across all electrode combinations to calculate a mean representational connectivity measure between electrodes. The mean representational connectivity measure is an index of how similar the representational space is between electrodes. This value is driven by two factors: spatial proximity (i.e., neighboring electrodes will have higher connectivity) and representational similarity due to stimulus features. As a control, we performed the analysis on shuffled labels for each of the respective stimulus modalities, which will account for the shared signal due to spatial proximity of neighboring electrodes, but not for the evoked responses to the specific stimuli. The shuffled control served as our comparison for all statistical comparisons.

We found that auditory, visual, and audiovisual presentations all diverged from the shuffled control (B.F. > 3), beginning at 97 ms, 107 ms, and 78 ms after stimulus presentation, respectively. Averaging across the 50-500ms stimulus period, we found that audiovisual presentations had more mean connectivity than visual presentations (B.F. >10) and auditory presentations (B.F. >30), but there was inconclusive evidence between visual and auditory connectivity (B.F. = 0.52). In addition, to compare the audiovisual response to the visual and auditory response more directly, we summed the evoked potentials for auditory and visual stimuli for each individual exemplar and then used this summed potential as input to the RCA. We found that the summed unisensory mean connectivity was significantly lower (B.F. > 30) than the mean audiovisual representational connectivity. These results suggest that shared representations across areas that lead to an increase in the mean

connectivity for audiovisual presentations is due to the simultaneous processing of auditory and visual stimuli, and not simply due to visual and auditory signals collectively activating more (or at least a more extensive set of) areas in the brain.

Similar to the RSA findings, we also found that the animate and inanimate category selectively affected connectivity measurements across the different sensory modalities. For auditory objects, connectivity diverged from the shuffled control for animate and inanimate exemplars at 156 ms. Mean connectivity over the stimulus period between groups showed substantial evidence for the null hypothesis (B.F. < 1/3), indicating no animacy difference for representational connectivity in audition. For visual objects, mean connectivity for animate objects and inanimate objects began to diverge from the shuffled control at 137 ms and 107 ms, respectively. However, visual animate exemplars had a greater mean representational connectivity than inanimate exemplars from 176-186 ms and summed over the stimulus period (B.F. >3). For audiovisual presentations inanimate objects diverge from baseline earlier at 107 ms compared to 127 ms for animate objects. In contrast to visual presentations, audiovisual animate and inanimate categories showed inconclusive evidence over the stimulus period (B.F. = 0.39). Lastly, for the summed unisensory responses, animate and inanimate objects diverged from the shuffled control at 146 ms and 107 ms, respectively. Averaged over the stimulus period there was inconclusive evidence (B.F = 0.71) for group differences. In summary, these results build off of the RSA analyses, and suggest that the presentation of objects in an audiovisual manner increase the representational connectivity when compared to when they are presented in a unisensory context, and furthermore that these connectivity measures increase to a greater extent for inanimate exemplars.

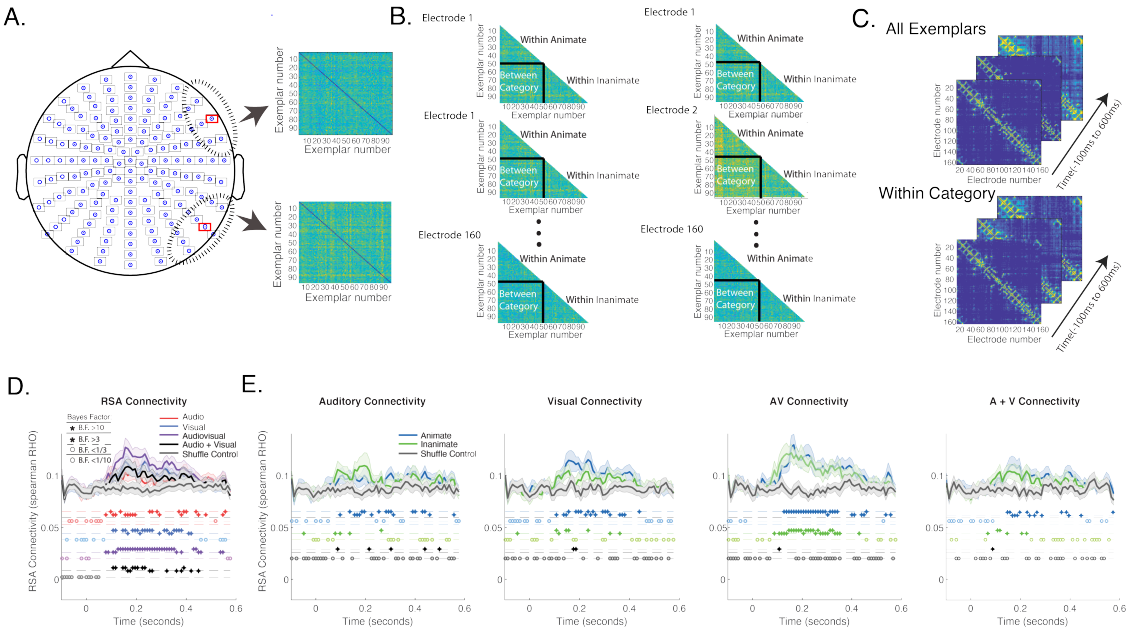

Figure 4.5: Representational Connectivity Analysis: Response Patterns between Brain Networks are Influenced by Object Category. (A) Moving searchlight to create electrode specific RDMs. The searchlight included the electrode of interest and every immediate surrounding electrode to produce an electrode specific RDM for each modality. (B) Each electrode was correlated in a pairwise fashion using a Spearman correlation. (C) This procedure was done for all exemplars as well as within the animate and inanimate category along the timecourse (-100 to 600ms) to build time-resolved electrode similarity matrices of representational space. The mean value of these matrices is the representational connectivity across all electrodes. (D-E) Representational connectivity was measured (D) across modalities and summed unisensory responses as well as (E) within the animate and inanimate categories across modalities.

4.4.5 Distance-to-Bound Analysis: Behavior can be Predicted by Exemplar Distance to the Decision Boundary in Representational Space

Having found both behavioral and neural differences between modality of presentation and animacy categories, we next considered whether the two measures were associated with one another. To do this, we computed the distance to the classifier decision boundary for all exemplars and correlated these distances with behavioral performance (i.e., reaction times). A negative correlation would denote that exemplars that are farthest away from the classifier decision boundary are those that are most rapidly categorized. Indeed, Figure 4.6A shows substantial evidence for a significant negative Spearman correlation (B.F. > 3) between representational distance and reaction time at several timepoints between 100-200 ms post-stimulus onset for both visual and audiovisual presentations and between 270-400 ms post-stimulus onset for all sensory modalities. Below the timecourse we show the results from the topographic results from applying the distance to bound analysis using a moving searchlight. We found that for visual and audiovisual presentations, occipital and temporal electrodes were most correlated to behavior for the time period spanning 100-200 ms post-stimulus onset. In contrast, frontoparietal electrodes were most correlated with behavior for the interval spanning 270-400 ms post-stimulus onset across all modalities. Figure 4.6B shows the corresponding scatter plot for the highest negative correlations in the 100-200 ms time window for visual and audiovisual presentations. These plots show that for both visual and audiovisual presentations, inanimate objects had slower categorization times than animate objects and were also closer to the decision boundary. Additionally, consistent with our behavioral and RSA results, inanimate exemplars appeared to show a greater shift along the reaction time and representational axes than animate exemplars when comparing between visual and audiovisual scatter plots.

In Figure 4.6C, we quantified this observation by using a Spearman correlation to link the reaction time difference for audiovisual versus visual exemplars with the representational difference for animate and inanimate exemplars. A negative correlation denotes: 1)

exemplars that were further away from the decisional boundary for audiovisual presentations when compared with visual presentations (positive AV-V distance value) are also the exemplars that demonstrated either more of an audiovisual RT bias (positive AV-V RT value) or less of a visual bias (negative AV-V RT value); and 2) exemplars that were further away from the decision boundary for visual presentations when compared with audiovisual presentations (negative AV-V distance value) are also the exemplars that demonstrated less of an audiovisual RT bias (positive AV-V RT value) or more of a visual bias (negative AV-V RT value). We found significant timepoints between 100-200 ms and 370-450 ms post-stimulus onset supporting the alternative hypothesis (B.F. >3) for inanimate exemplars, but only evidence for a null correlation (B.F. < 1/3) for animate exemplars. If we pool the correlations across the entire stimulus analysis epoch (50-500 ms post-stimulus) we find very strong evidence for a negative correlation for inanimate exemplars (B.F.> 30) but inconclusive evidence for animate exemplars (B.F. = 2.00). Figure 4.6D shows the corresponding scatterplot with the highest negative correlation in the 100-200 ms window for visual and audiovisual presentations at 137ms (same as figure 4.6B). Collectively, these results show associations between neural decoding differences and behavioral performance differences between audiovisual and visual stimulus presentations, but only when these stimuli are inanimate.

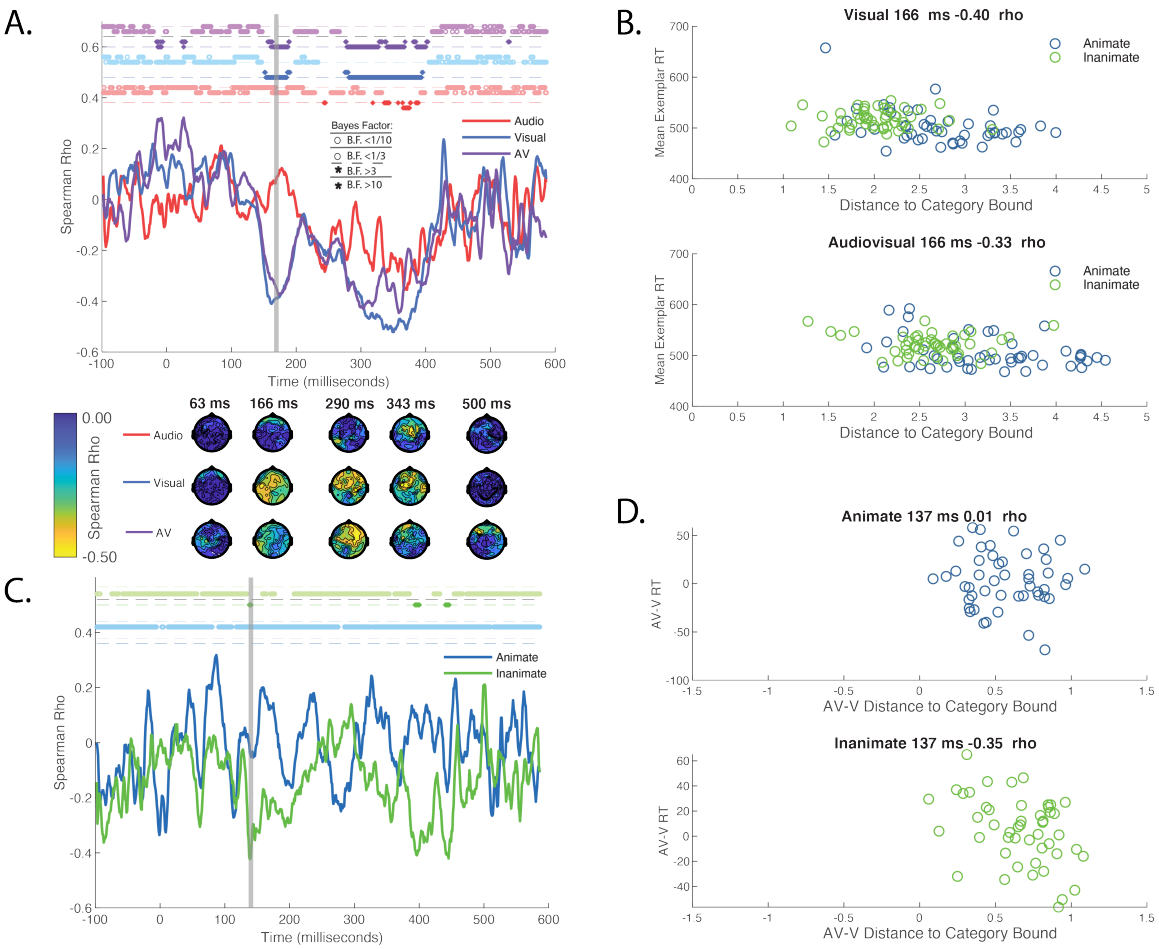

Figure 4.6: Distance-to-Bound Analysis: Behavior can be predicted by Exemplar Distance to the Decision Boundary in Representational Space. (A) Time-varying Spearman correlation between mean exemplar representational distance from animacy discriminate bound and respective average exemplar reaction time for each modality. Asterisks indicate substantial and strong evidence for the alternative hypothesis (B.F. > 3 and >10) of a correlation above 0 and null hypothesis (B.F. < 1/3 and <1/10). Below the x-axis, results from the topographic results from applying the distance to bound analysis using a moving searchlight for select timepoints (B) Scatterplot for mean exemplar visual and audiovisual representational distance and RT at a significant timepoint for both modalities. (C) Time-varying Spearman correlation between mean representational enhancement (Audiovisual-Visual distance) and median reaction time enhancement (Audiovisual-Visual RT) with asterisks indicating significant Spearman correlation (B.F. > 3) (D) Scatterplot for audiovisual representational and RT enhancement at a marginally significant timepoint for inanimate exemplars.

4.4.6 Model Testing: Abstract Category Models Predict Neural Activity Better than Low-Level Feature Models

To account for the potential contribution of low-level features to the neural RDMs, we constructed contrast dissimilarity matrices for images and power spectral density dissimilarity matrices for sounds as shown in Figure 4.7. The models were correlated using a Spearman correlation to each subject's neural RDM across channels and neighborhoods of electrodes using a moving searchlight to build topographic maps. Along the time series, we tested for significance using Bayes Factors (B.F. > 3). The contrast model and power spectrum model only had sporadic time points that had substantial evidence for the alternative hypothesis. The power spectrum model was most correlated with the auditory RDM with time points between 170-200 ms post stimulus presentation while the contrast model was most correlated with the visual RDM from 90-170 ms post stimulus presentation. Further as shown in Figure 4.7C, at early times, such as 107 ms, the occipital electrodes are most correlated with the contrast model. Similarly, for the auditory RDMs, temporoparietal electrodes correlate most with power spectrum model early at 78 ms and late in the timecourse at 400ms. In contrast, when we used an abstract model that ignored low level features and instead separated stimuli based on object animacy category, we found a significant correlation (B.F. > 3) with the visual RDMs beginning at 150ms and audiovisual RDMs at 158 ms. Occipital and temporal electrodes for visual and audiovisual presentations were most correlated to the animacy model at timepoints such as 176ms but not later at 400ms. The animacy model did not show a sustained correlation to the auditory RDM, implying that the animacy distinction is not as prominent in audition.

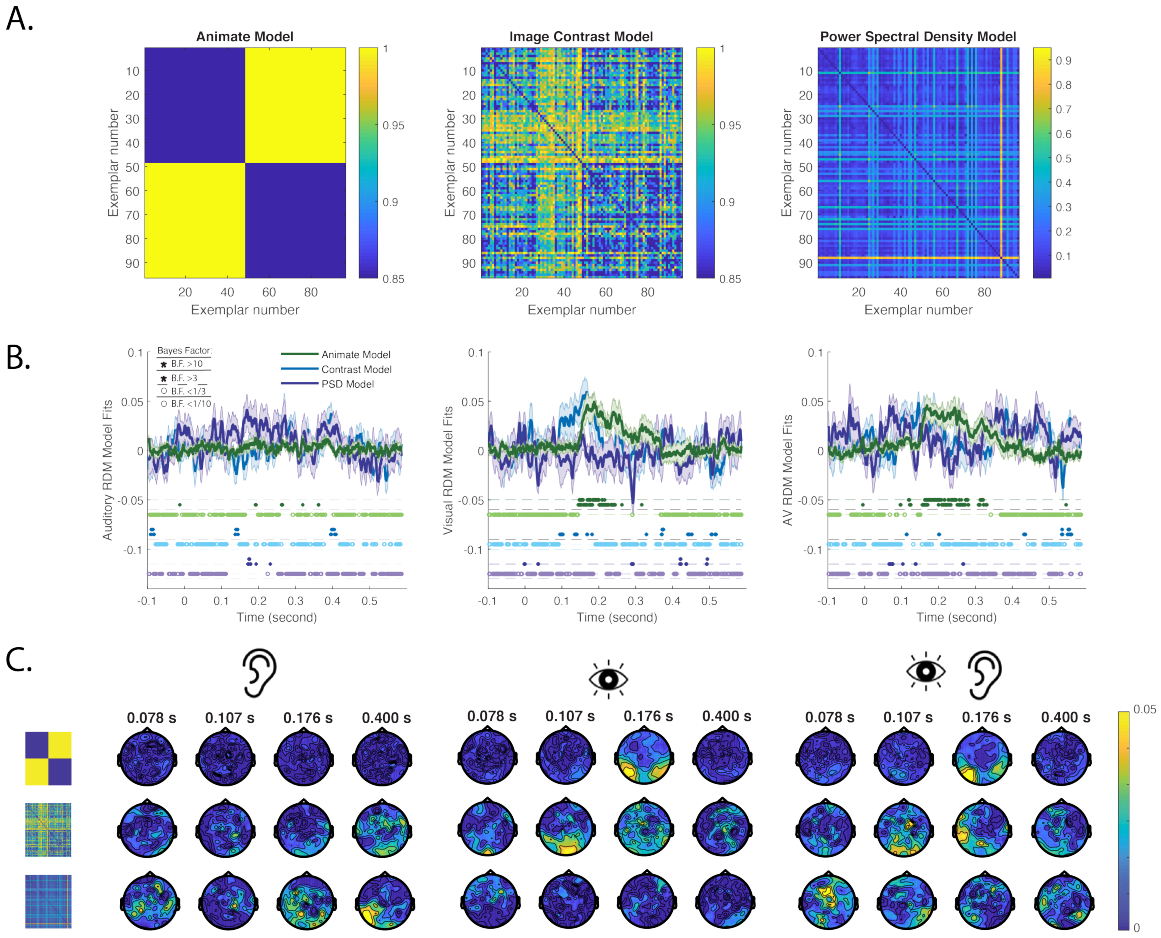

Figure 4.7: Model Testing: Abstract Category Models Predict Neural Activity Better than Low-Level Feature Models. (A) Category and low level visual and auditory feature models. The animacy category model was constructed using a "0" for within animacy category exemplars and "1" for between animacy category exemplars. For the image contrast RDM, since all images were black and white drawings, the contrast vector consisted of the intensity values of each image. The power spectral density RDM was built using a Welch's power spectral density estimate and converted to a single vector for each sound. Each RDM was then constructed by taking the Spearman distance of each respective pairwise stimulus comparison. (B) Each model RDM was then tested with the Auditory, Visual, and Audiovisual time resolved RDMs on a subject by subject basis. Shaded area around lines indicate standard error across subjects with asterisks indicating substantial and strong evidence for the alternative hypothesis (B.F. > 3 and >10) of a correlation above 0 and null hypothesis (B.F. < 1/3 and <1/10). (C) Model testing performed on electrode specific RDMs using a searchlight analysis.

4.5 Discussion

In this study, we leveraged the visual and auditory encoding bias that has been observed for animate objects over inanimate objects (Grootswagers, Ritchie, et al., 2017; Guerrero & Calvillo, 2016; Murray, 2006; Tzovara et al., 2012; Vogler & Titchener, 2011) to study how perceptual biases across object categories influence the multisensory enhancement of audiovisual objects. Using behavioral measures and neural decoding, we found additional support for previous findings showing visual and auditory perceptual advantages for animate objects over inanimate objects. However, and somewhat surprisingly, we found that the advantage for animacy was not evident when objects were presented as audiovisual objects. Using RSA, we show that the lack of an animacy bias in audiovisual objects is in the context of an overall expansion of representational space when compared to visual and auditory objects. Further analysis showed that audiovisual presentations preferentially enhanced neural decoding of inanimate objects. A searchlight analysis and representational connectivity analysis showed that the presentation of inanimate objects in an audiovisual context may improve their encoding through increased representational connectivity between brain areas. We finally linked neural decoding and behavioral performance by using a distance to bound analysis and found that improved neural decoding for visual and audiovisual objects was associated with faster reaction times in the animacy categorization task. Furthermore, the decoding differences between visual and audiovisual objects was also predictive of their reaction time differences. Taken together, the results of our study provide new insights into the encoding of unisensory and multisensory objects, establishes critical links between neural activity and behavior in the context of object categorization, as well as explores potential mechanistic differences in multisensory integration for weakly and strongly encoded objects.

Although stimulus features clearly contribute to the formation of object categories, including the distinction between animate and inanimate objects, there is ample evidence that the animate-inanimate distinction transcends stimulus features and can be thought of as an

abstract category distinction. The distinction is present for stimuli presented in both the visual and auditory modalities, suggesting that animacy is a general organizing principle. Furthermore, category-specific deficits in naming animate objects have been found in patients who have suffered brain damage (Capitani, Laiacona, Mahon, & Caramazza, 2003; Clarke et al., 2002; Kolinsky et al., 2002; Vignolo, 1982; Vignolo, 2004; Warrington & Mccarthy, 1987). The category distinction is preserved across species; being present in both monkey inferotemporal (IT) cortex and human IT cortex. Furthermore, the use of carefully controlled stimuli that account for stimulus features have reinforced the categorical nature of animacy (Bracci, Ritchie, & de Beeck, 2017; Ritchie & Op De Beeck, 2018). Similarly, auditory studies have also provided evidence for animacy as an abstract category distinction (De Lucia et al., 2010; Giordano et al., 2013; Murray, 2006). In the current study, we corroborate these findings by showing a significant correlation between an animacy model and neural response patterns, but a lack of consistent correlations between low-level stimulus features such as visual contrast and auditory power spectrum with neural response patterns.

Our study showed overall magnitude and temporal enhancement for audiovisual objects over visual and auditory objects consistent with recent findings (Brandman et al., 2019; Mercier & Cappe, 2019), and we additionally provide new insights into how audiovisual benefits selectively enhance the category of inanimate objects. Specifically, we found that the animacy bias for auditory and visual objects is absent in audiovisual objects. We hypothesized that the brain may be preferentially integrating the visual and auditory components of the more weakly encoded inanimate objects. Thus, greater multisensory integration for inanimate objects may serve to close the perceptual gap between animate and inanimate objects, consistent with the concept of inverse effectiveness (Stein & Meredith, 1993; Wallace et al., 2004). To test whether there were behavioral differences in multisensory integration across categories, we examined our behavioral data for a prediction made by maximum likelihood estimate models (Ernst & Banks, 2002): there is stronger multisensory benefit when the unisensory reliability or other measure of variability between senses

is closer (i.e. smaller differences between visual and auditory reaction time). In agreement, we found that smaller RT differences between visual and auditory objects led to faster multisensory reaction times for inanimate, but not for animate objects. In the same vein, the neural decoding bias for animate over inanimate objects was no longer present for audiovisual presentations. When we subtracted audiovisual decoding from visual decoding, we found that decoding was only enhanced for inanimate objects, lending further evidence that audiovisual presentations selectively improved encoding of inanimate objects.

To investigate the potential mechanism by which audiovisual presentations asymmetrically enhance the decoding of inanimate objects, we utilized representational connectivity analysis across all EEG sensors. Representational connectivity analysis has been previously used in a more limited way to assess representational similarity between two brain areas (Kriegeskorte et al., 2008). In our analysis, we used a moving searchlight consisting of each electrode and its immediately surrounding neighbors to measure the different patterns of activity for each given stimulus. By doing so, we are able to use RCA as a tool to acquire a data driven measure of how similar response patterns are topographically across the brain. We predicted that animate and inanimate exemplars might demonstrate differences in connectivity measures, as previous studies have shown increased connectivity for biologically plausible motion over mechanical motion (Hillebrandt et al., 2014). Note that in this analysis, neighboring electrodes will have shared signals simply due to proximity. Therefore, the importance of these connectivity measures is the relative difference between animate and inanimate categories. We found increased representational connectivity for animate objects when presented in vision and when compared with inanimate objects. However, much like for our RSA results, these connectivity differences were no longer present for when these objects were presented in an audiovisual context. Additionally, the connectivity increase for inanimate objects occurs within the 100-200 ms time epoch we have previously noted as the time period in which audiovisual presentations showed the greatest enhancement over visual presentations. One possible explanation for these results is that

there may be increased audiovisual integration for inanimate objects relative to animate objects, leading to greater spread of neural representation across brain areas. However, the current analysis cannot exclude the possibility that the increase in inanimate connectivity for audiovisual presentations may also be due a more localized spread within electrodes in close proximity.

Next, we directly linked the neural results to behavioral results at the exemplar level by using a distance to bound approach (Carlson et al., 2014; Grootswagers, Ritchie, et al., 2017; Ritchie et al., 2015). This approach is a data-driven way of determining the relationship between neural representational space and behavioral measures (i.e., reaction times). In this analysis, we found a significant relationship between visual and audiovisual decoding distances and reaction times during two distinct post-stimulus time epochs. One corresponded to peak decoding in our RSA analysis (i.e., 100-200 ms) and the other emerged approximately 150-200ms later. These intervals and the corresponding topographic analyses in Figure 4.6A correspond to periods and electrodes associated with sensory evidence accumulation and decision-making, respectively (Murray, Imber, Javitt, & Foxe, 2006; Tzovara et al., 2012). We next directly correlated multisensory neural decoding enhancements to reaction time improvements. Interestingly, we found that despite an overall neural enhancement for audiovisual presentations, some exemplars showed possible effects of audiovisual interference effects. In these cases, visual decoding distances were greater than audiovisual decoding distances. These effects were largely reflected in the reaction time differences between audiovisual and visual presentation, with an overall significant negative correlation between behavioral audiovisual enhancement and neural audiovisual enhancement. These results provide evidence that the added sensory information in audiovisual presentations did not just provide the classifier with more information, but in fact provide further value for the object categorization task (Grootswagers, Cichy, & Carlson, 2018). However, it does not eliminate the possibility that added neural information was also used for other aspects of the perceptual response not tapped in the current paradigm (e.g., response confidence).

In conclusion, our study introduces new insights into the brain's representation of sensory and multisensory information as it relates to object encoding. The greater neural encoding benefits for inanimate stimuli seen under audiovisual conditions compliments prior work, where sensory information was selectively removed from object stimuli, resulting in a selective contraction of the representational space of animate objects (Grootswagers, Ritchie, et al., 2017). Collectively, these findings show that neural representational space and the encoding of objects is impacted by both semantic congruence and stimulus modality (stimulus combinations) in a dynamic fashion. Future directions of our current work include approaches to investigate the interplay between parametrically reducing neural encoding by degrading visual stimuli while simultaneously using audiovisual presentations to enhance neural encoding. Understanding the computational framework the brain uses to maximize the sensory information it captures across sensory systems has broad implications for how stimuli perturbations and sensory integration affects object encoding.

4.6 References

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403:309–312.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3):317–335.

Berry, D. A. and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1-2):215–227.

Bracci, S., Ritchie, J. B., and de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105(June):153–164.

Braga, R. M., Hellyer, P. J., Wise, R. J., and Leech, R. (2017). Auditory and visual connectivity gradients in frontoparietal cortex. *Human Brain Mapping*, 38(1):255–270.

Brandman, T., Avancini, C., Leticevscaia, O., and Peelen, V. M. (2019). Auditory and semantic cues facilitate decoding of visual object category in meg. *Cerebral Cortex*, (June):1–10.

Capitani, E., Laiacona, M., Mahon, B., and Caramazza, A. (2003). *What are the facts of semantic category-specific deficits? A critical review of the clinical evidence*, volume 20.

Cappe, C., Thelen, A., Romei, V., Thut, G., and Murray, M. M. (2012). Looming signals reveal synergistic principles of multisensory integration. *Journal of Neuroscience*, 32(4):1171–1182.

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10):1–19.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., and Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, 26(1):132–142.

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–62.

Clarke, S., Bellmann Thiran, A., Maeder, P., Adriani, M., Vernet, O., Regli, L., Cuisenaire, O., and Thiran, J. P. (2002). What and where in human audition: Selective deficits following focal hemispheric lesions. *Experimental Brain Research*, 147(1):8–15.

De Lucia, M., Clarke, S., and Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *Journal of Neuroscience*, 30(33):11210–11221.

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211.

Gelman, A. and Tuerlinckx, F. (2000). Type s error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, 15:373–390.

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., and Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, 23(9):2025–2037.

Grootswagers, T., Cichy, R. M., and Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, 179(June):252–262.

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., and Carlson, T. A. (2017a). Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions. *Journal of cognitive neuroscience*, 29(12):1995–2010.

Grootswagers, T., Robinson, A. K., and Carlson, T. A. (2019). The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188(October 2018):668–679.

Grootswagers, T., Wardle, S. G., and Carlson, T. A. (2017b). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, 29(4):677–697.

Guerrero, G. and Calvillo, D. P. (2016). Animacy increases second target reporting in a rapid serial visual presentation task. *Psychonomic Bulletin and Review*, 23(6):1832–1838.

Hillebrandt, H., Friston, K. J., and Blakemore, S. J. (2014). Effective connectivity during animacy perception - dynamic causal modelling of human connectome project data. *Scientific Reports*, 4:1–9.

Jackson, R. E. and Calvillo, D. P. (2013). Evolutionary psychology. *Evolutionary Psychology*, 11(5):1011–1026.

Jarosz, A. F. and Wiley, J. (2014). What are the odds? a practical guide to computing and reporting bayes factors. *Journal of Problem Solving*, 7:2–9.

Jeffreys, H. (1998). *The Theory of Probability*. Oxford.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48):19313–19317.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(11):4302–11.

Kolinsky, R., Fery, P., Messina, D., Peretz, I., Evinck, S., Ventura, P., and Morais, J. (2002). The fur of the crocodile and the mooing sheep: A study of a patient with a category-specific impairment for biological things. *Cognitive Neuropsychology*, 19(4):301–342.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):4.

Kriegeskorte, N., Mur, M., Ruff, D., and Kiani, R. (2008b). Matching categorical object

representations in inferior temporal cortex of mand and monkey. *Neuron*, 60(6):1126–1141.

Laws, K. R. (2000). Category-specific naming errors in normal subjects: The influence of evolution and experience. *Brain and Language*, 75(1):123–133.

Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., and Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg- and meg-data. *Journal of Neuroscience Methods*, 164(1):177–190.

Mercier, M. R. and Cappe, C. (2019). The interplay between multisensory integration and perceptual decision making. pages 1–26.

Murray, M. M. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience*, 26(4):1293–1302.

Murray, M. M., Imber, M. L., Javitt, D. C., and Foxe, J. J. (2006). Boundary completion is automatic and dissociable from shape discrimination. *Journal of Neuroscience*, 26(46):12043–12054.

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti, M., Gors, J., Gobbini, M. I., and Haxby, V. J. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, pages 1–15.

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598–16603.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). Fieldtrip: Open source

software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011.

Oosterhof, N. N., Connolly, A. C., and Haxby, V. J. (2016). Cosmomvpa: Multi-modal multivariate pattern analysis of neuroimaging data in matlab/gnu octave. *Frontiers in Neuroinformatics*, 10(JUL):27.

Ritchie, J. B. and Carlson, T. A. (2016). Neural decoding and "inner" psychophysics: A distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in Neuroscience*, 10(APR):1–8.

Ritchie, J. B. and Op De Beeck, H. (2018). Using neural distance to predict reaction time for categorizing the animacy, shape, and abstract properties of objects. pages 1–19.

Ritchie, J. B., Tovar, D. A., and Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology*, 11(6).

Robinson, A. K., Grootswagers, T., and Carlson, T. A. (2019). The influence of image masking on object representations during rapid serial visual presentation. *NeuroImage*, 197(January):224–231.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, 16(2):225–237.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, 55(1):62–71.

Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures : Norms for name agreement , image agreement , familiarity , and visual complexity. *Journal of Experimental Psychology*, 6(2):174–215.

Stein, B. E. and Meredith, M. A. (1993). *The merging of the senses*. The MIT Press, Cambridge, MA.

Thelen, A., Cappe, C., and Murray, M. M. (2012). Electrical neuroimaging of memory discrimination based on single-trial multisensory learning. *NeuroImage*, 62(3):1478–1488.

Tzovara, A., Murray, M. M., Plomp, G., Herzog, M. H., Michel, C. M., and De Lucia, M. (2012). Decoding stimulus-related information from single-trial eeg responses based on voltage topographies. *Pattern Recognition*, 45(6):2109–2122.

van den Hurk, J., Van Baelen, M., and Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22):E4501–E4510.

VIgnolo, L. (1982). Auditory agnosia. *Philosophical Transactions of the Royal Society B*, 298:49–57.

Vignolo, L. A. (2004). Music agnosia and auditory agnosia. *Annals of the New York Academy of Sciences*, 999(1):50–57.

Vogler, J. N. and Titchener, K. (2011). Cross-modal conflicts in object recognition: Determining the influence of object category. *Experimental Brain Research*, 214(4):597–605.

Wallace, M. T., Ramachandran, R., and Stein, B. E. (2004). A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences*, 101(7):2167–2172.

Warrington, E. K. and Mccarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, 110(5):1273–1296.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E. J.

(2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3):291–298.

Yuval-Greenberg, S. and Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research*, 193(4):603–614.

Part 3

"What I cannot create I do not understand"

-Richard Feynman

Chapter 5

Getting the gist faster: Blurry images enhance the early temporal similarity between neural signals and convolutional neural networks

The contents of this chapter are adapted from

Tovar, D. A., Grootswagers, T., Jun, J., Cha, O., Blake, R., & Wallace, M. T. (In Prep). Getting the gist faster: Blurry images enhance the early temporal similarity between neural signals and convolutional neural networks.

5.1 Abstract

Humans are able to recognize objects under a variety of noisy conditions, so models of the human visual system must account for how this feat is accomplished. In this study, we investigated how image perturbations, specifically reducing images to their low spatial frequency (LSF) components, affected correspondence between convolutional neural networks (CNNs) and brain signals recorded using magnetoencephalography (MEG). Using the high temporal resolution of MEG, we found that CNN-Brain correspondence for deeper and more complex layers across CNN architectures emerged earlier for LSF images than for their unfiltered broadband counterparts. The early emergence of LSF components is consistent with the coarse-to-fine theoretical framework for visual image processing, but surprisingly shows that LSF signals from images are more prominent when high spatial frequencies are removed. In addition, we decomposed MEG signals into oscillatory components and found correspondence varied based on frequency bands, painting a full picture of how CNN-Brain correspondence varies with time, frequency, and MEG sensor locations. Finally, we varied image properties of CNN training sets, and found marked changes in CNN processing dynamics and correspondence to brain activity. In sum, we show that image perturbations affect CNN-Brain correspondence in unexpected ways, as well as provide a rich methodological framework for assessing CNN-Brain correspondence across space, time, and frequency.

5.2 Introduction

The human visual system has been characterized as a hierarchical system that begins with extraction of information about simple features (e.g., oriented contours) registered by neurons whose receptive fields are retinotopically organized, followed by increasingly refined analysis of more complex aspects of the visual scene via neurons with increasingly large receptive fields (Hubel & Wiesel, 1977; Poggio & Riesenhuber, 1999; Serre, Oliva, & Poggio, 2007; Vinckier et al., 2007). Generally, inspired by this biological organization, convolutional neural networks (CNNs) built for image classification have been similarly constructed such that early convolutional layers register simple features in small receptive fields, followed by pooling layers that progressively increase receptive field size, allowing subsequent convolutions to extract complex features that are then passed to fully connected layers for classification (Kietzmann, Mcclure, & Kriegeskorte, 2019; Lecun, Bengio, & Hinton, 2015; Richards et al., 2019). Although neural networks are biologically implausible in some ways, such as weight sharing and backpropagation, they are nevertheless increasingly recognized as useful models of neural processing (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). Still, recent studies question the generality of the correspondence between neural networks and neural activity, noting that the relationship between fMRI activation patterns and CNNs is considerably weakened when visual images are degraded or comprised of artificial objects (Xu & Vaziri-Pashkam, 2021). However, it remains possible that the poor temporal resolution of fMRI obscures the category structure that emerges as a function of the temporal dynamics of processing within the visual stream (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Cichy, Pantazis, & Oliva, 2014; Wardle & Baker, 2020). Thus, the degree of correspondence between CNN models and dynamic brain signals associated with degraded visual images remains an open question. In the current work, we address this question by measuring brain responses to degraded images using high temporal resolution magnetoencephalography (MEG) and comparing these to performance in a number of CNNs.

The form of visual image degradation we have focused on is motivated by the coarse-to-fine manner by which the brain is thought to optimize object recognition (Bar, 2003a; Bar, Kassam, Ghuman, Boshuan, et al., 2006; Petras, ten Oever, Jacobs, & Goffaux, 2019). This view posits that low spatial frequency information is processed by the faster magnocellular pathway (Kauffmann, Ramanoël, Guyader, Chauvin, & Peyrin, 2015; Tootell, Silverman, Hamilton, Switkes, & De Valois, 1988), which creates an initial coarse representation of the image/object. Called "scene gist", those initial representations or "hunches" are then refined as more detailed information emerges in the form of high spatial frequencies processed by the slower parvocellular pathway traveling through the ventral visual stream (Bar, 2003a; Bar, Kassam, Ghuman, Boshuan, et al., 2006; Bruner & Potter, 1964; Snodgrass & Hirshman, 1991; Tootell et al., 1988). Low frequency information was initially thought to enhance processing within the ventral visual stream through feedback signals originating in the orbitofrontal cortex (OFC) to category selective areas in inferotemporal (IT) cortex (Bar, 2003; Bar, Kassam, Ghuman, Boshuan, et al., 2006). However, recent evidence suggests that feedback processes are more diffuse along the ventral visual stream. For example, an fMRI occlusion paradigm that selectively manipulated the spatial frequency along different receptive fields found that low frequency information is conveyed through feedback signals throughout the ventral visual stream, including in primary visual cortex (Revina, Petro, & Muckli, 2018). Additionally, high spatial frequency processing domains are segregated from low spatial frequency processing domains as far upstream as V4, indicating that unique spectral information is preserved within feedforward processing (Lu et al., 2018). Collectively, it thus appears that coarse-to-fine processing comprises a combination of dynamic feedforward and feedback interactions. This implies that the extent to which the brain relies on low spatial frequencies to initiate top-down processes depends on the available spectral and contextual information present in an image.

Modeling the visual system requires capturing the dynamics of object recognition under a variety of task constraints, including degraded images that necessitates varying degrees

of feedback/top-down processing. The sluggish fMRI signal makes it difficult to differentiate between the dynamics of early feedforward and later feedback processes; these dynamics take on particular importance as we go beyond assessing CNN-Brain correspondence with natural images (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Güçlü & van Gerven, 2015, 2017; Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann, Spoerer, et al., 2019; Kong, Kaneshiro, Yamins, & Norcia, 2020; Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021; Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Schmidt, et al., 2018). Behavioral studies have shown that CNNs differ from human vision in terms of susceptibility to the impact of image distortion on object recognition. For example, distortions such as color remapping, low pass filtering and high pass filtering reduce CNN performance in object classification but have considerably less effects on human performance (Geirhos et al., 2018). Thus, the effect of image perturbations on CNN-Brain correspondence is best suited using brain signals measured with high temporal resolution techniques such as M/EEG.

Consequently, in the current work, we have studied the temporal correspondence between neural activity collected using MEG and a diverse set of CNN architectures for clear images as well as for degraded images containing only low spatial frequency components. The added temporal resolution in the MEG allows us to make inferences regarding how the correspondence between MEG signals and the CNN activations evolves throughout the stimulus presentation, and whether image perturbations change the timing of when the correspondence emerges. We predicted that for all images (clear and degraded) there would be a general temporal relationship between CNN layer depth and the time course of the MEG signal following stimulus presentation. In such a framework, shallow CNN layers (those close to the input layer) will correspond to earlier times in the MEG signal and deep CNN layers will correspond to later times in the MEG signal when participants have been allowed more time to fully process an object. However, for the low spatial frequency images, we hypothesized that an enhanced contribution of top-down feedback would result in

the more rapid emergence of correspondence between deep CNN layers and MEG signals.

5.3 Methods

5.3.1 Data Set

We used a data set originally published in Grootswagers, et al., 2017. The data comprised results from 20 participants (four men; mean age = 29.3 years) with normal or corrected-to-normal vision participating in an MEG experiment. Stimuli consisted of 48 grayscale images comprised of an even split of animate and inanimate objects on a phase-scrambled natural image background (Figure 1A). Importantly, the stimuli did not include humans and better accounted for shape and other confounds present in the stimuli in other datasets (Grootswagers & Robinson, 2021). The objects were presented in a clear condition and a degraded condition intermixed within eight blocks, resulting in 32 trials for each respective clear and degraded object. Degraded images were constructed by convolving a sombrero function over a Fourier transformed image and selecting varying radii of pixels from the image, resulting in different degrees of low spatial frequency blur (Figure 1A and Supplemental Figure 1). Given that different types of blurring can affect object recognition to different degrees (Kadar & Ben-shahar, 2012), each image was blurred based on the results from a separate online MTurk experiment with blur being set as the radii by which at least 25% of participants could name the object in a naming task. Stimuli were projected (at $9° \times 9°$ visual angle) on a black background for 500ms with a random inter-trial interval between 1000 and 1200 milliseconds. Participants categorized the stimulus as animate or inanimate as fast and accurately as possible. Motor responses were remapped between alternating blocks to avoid potential motor confounds. Prior to the MEG experiment, a familiarization task was used to make sure that all participants could categorize all clear and degraded stimuli as animate or inanimate with accuracy scores of at least 80%. Each MEG recording was done with a whole-head MEG system (model PQ1160R-N2; KIT, Kanazawa, Japan) while participants lay in a supine position inside a magnetically

shielded room. Trials were sliced into 700ms epochs spanning from 100ms prior to stimulus onset to 600ms post stimulus onset.

5.3.2 Decoding between clear and degraded images

To determine when differences emerged in time between clear and low spatial frequency degraded images (Figure 1B), we trained and tested a classifier using linear discriminant analysis (LDA) (Duda & Hart, 2001). In this procedure, we used a four-fold, leave one-fold out train to test split, iteratively changing which folds were trained and tested. We performed this analysis using all of the MEG sensors to compute an overall decoding classification performance for clear unfiltered images and for low spatial frequency degraded images. Statistical significance was computed by comparing the decoding performance to chance level decoding (50%), correcting for multiple comparisons using FDR correction. In addition, to obtain a topographic estimate of how clear and degraded images are distinguished in the brain, we performed a moving searchlight analysis (Etzel, Zacks, & Braver, 2013), iteratively decoding clear from degraded images at each sensor and its immediate surrounding neighboring sensors. This procedure produced a topographic heat map of decoding performance along 100ms intervals, spanning from 100ms prior to stimulus presentation to 600ms post stimulus presentation.

5.3.3 Neural RDMs

To capture the time resolved neural relationship between objects, we used representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008). For each exemplar, we performed pairwise decoding using LDA with four-fold leave one-fold out cross validation for all stimulus comparisons within the clear and low spatial frequency degraded images until we had decoding scores across all possible exemplar comparisons across all time points. Together, these formed time-resolved representational dissimilarity matrices (RDMs) for clear and degraded images respectively (Figure 1C).

5.3.4 CNN RDMs

Network RDMs were similarly constructed using RSA (Figure 1C). We chose a diverse set of six CNNs of varying depth as well as different types of connections, including skip connections (He, Zhang, Ren, & Sun, 2015), inception layers (Szegedy et al., 2014), and recurrence (Kubilius et al., 2018). Instead of using cross validation, we used the square Euclidian distance between layer activations for each exemplar comparison to build the RDMs. We chose this distance measurement to make the fewest necessary assumptions regarding the relationship between layer activations for each object. Note that in this process, each n $\times$ n layer activation is converted to 1 $\times$ n vectors preserving the relative relationship of activation within each layer. To measure network dynamics and correspondence with brain activity, we selected all of the convolutional and fully connected layers within each network. However, we also performed the analysis using all possible computations within each network, including pooling layers where convolutional features are pooled, ReLU activation functions that convert all negative values to zero, and normalization layers that scale and center the activations, finding qualitatively similar results.

5.3.5 Probing Neural and Network Dynamics Separately

To probe whether CNNs and brain activity exhibit similar dynamics when processing clear images and degraded images, we correlated RDM averaged across all participants for each time across all other RDMs in our stimulus window (-100ms to 600ms). This analysis was performed using participant averaged brain RDMs instead of individual RDMs in order to have more stable neural representations. The RDMs are consistently changing in time, so by doing a cross correlation across timepoints we are capturing the dynamics of how each participant processed the clear and degraded objects. We performed a similar procedure separately for CNNs, using layers instead of time (Figure 1D). Given that the neural time window included time before stimulus presentation and that additional time elapses for neural signals to travel from the retina to visual cortex, we chose to begin the cross cor-

relations 50 ms after stimulus. Additionally, since each of the different CNN architectures contains different depths and layer, we interpolated each of the network activations to fit the same dimensions as the brain RDMs (Figure 2A) using a nearest-neighbor interpolation. The nearest-neighbor interpolation duplicates individual pixel values to fit the brain RDM values. We performed this analysis for clear images and for degraded images, and then correlated the relative representational geometry between the clear and degraded images for the brain RDMs and the various CNN architectures separately.

5.3.6 Correspondence between brain and CNN RDMs

To relate brain RDMs to the CNN layer specific RDMs, we used a non-parametric Spearman correlation between the brain and CNN matrices across each time point and network layer to avoid making any assumptions of linearity for the Brain-CNN correspondence. We then measured the time in which each CNN layer was maximally correlated to brain data. In addition, we calculated the lower bound of the brain noise ceiling for clear image presentations and degraded image presentations separately. The lower bound of the noise ceiling was approximated by iteratively calculating across all participants the mean correlation between each individual participant with the grand mean RDM minus that participant (Nili et al., 2014) (Figure 3A).

5.3.7 Topographic Correspondence between Brain and CNN RDMs

To assess how CNN-Brain correspondence changed as a function of sensor location, we constructed sensor by sensor RDMs using an electrode and its immediate surrounding neighbor sensors. As mentioned in the previous sections, we assessed CNN-Brain correspondence using Spearman correlations for each individual participant and then averaged the correlations across participants (Figure 4A). These results were tested for significance against zero correlation and corrected for multiple comparisons using FDR. To highlight the difference between clear and degraded images, we performed a pairwise test between conditions, correcting for multiple comparisons. For this analysis, we chose CORnet-S as

it was found to be one of the most brain-like networks (Kubilius et al., 2018; Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Geiger, et al., 2018) consisting of only five layers (layers 1-5 are labeled V1, V2, V4, IT and Decoder) and ResNet-50 (included in the supplemental material) which was the largest net we tested.

5.3.8 Spectral Correspondence between Brain and CNN RDMs

To capture spectral information, MEG signals were passed through a series of band-pass bidirectional Butterworth filters from 5 Hz to 45 Hz. We used a sliding window including the frequency of interest and 2 Hz above that frequency, such that 5 Hz represents 5-7 Hz, and 6 Hz represents 6-8 Hz, and so on and so forth. From the band-passed signals, we constructed frequency specific RDMs and then for each one of these frequencies measured the correspondence with CORnet-S RDMs and ResNet-50 RDMs (Figure 5A) for the same reasons described for the topographic correspondence. For ResNet-50, the RDMs were limited to one shallow, middle and deep layer; for CORnet-S, we included all the layers.

5.3.9 Stylized images and CNN transfer learning

To test how CNN training, and specifically the features included within the images in the training set, affected CNN-Brain correspondence, we made use of a ResNet-50 architecture trained on a stylized ImageNet set (Geirhos et al., 2019), which we will refer to as "StyleNet". The stylized images are the various images from ImageNet but with style transfer (Huang & Belongie, 2017) of textures from a diverse set of paintings (Figure 6A). For this network, the training parameters were as follows: 60 epochs with stochastic gradient decent, momentum term of 0.9, learning rate of 0.1 multiplied by 0.1 after 20 and 40 epochs, and a batch size of 256. In addition, we performed transfer learning on an AlexNet architecture, applying to ImageNet the low spatial frequency degradation that was used in the MEG experiment. Here, we used a degradation radius of 8 pixels on the cylinder in the sombrero convolution and applied this across all images. During transfer learning, we used a randomized subset of 250 of the 1000 image categories in ImageNet. The transfer

learning parameters were as follows: 60 epochs with stochastic gradient decent, momentum term of 0.9, learning rate of 0.001, and batch size of 64. We then used these networks to measure the dynamics, CNN-Brain correspondence, and topographic CNN-Brain correspondence using the procedures described in the previous sections.

5.4 Results

5.4.1 Difference between degraded low spatial frequency images and clear (i.e., unfiltered) images lateralizes to the right hemisphere

To determine when brain signals begin to diverge for low spatial frequency and clear, unfiltered images, we trained a classifier to distinguish between the two image types regardless of the specific exemplar. We found significant decoding onset at 50ms (Figure 1B), defined as at least two consecutive time points of significant decoding (Carlson et al., 2013). Decoding remained above chance throughout the stimulus period (500ms), peaking at 100ms post stimulus onset and extended to the end of the decoding window (100ms after stimulus offset). Using a searchlight analysis, we also measured topographic variation in the information regarding whether the image was clear or degraded. In the topographic maps, we found evidence of lateralization to the right hemisphere beginning at about 200ms and becoming more lateralized in time until 400ms. The lateralization of low spatial frequency information to the right hemisphere has been noted in previous studies (Flevaris & Robertson, 2016; Kauffmann, Ramanoël, & Peyrin, 2014; Schyns & Oliva, 1999). However, given that the difference between the low spatial frequency and the clear unfiltered image is the high frequency content, these results were somewhat surprising; high frequency information has been shown to lateralize to the left hemisphere (Flevaris & Robertson, 2016; Kauffmann et al., 2014; Schyns & Oliva, 1999). Thus, these results seem to suggest that the primary neural difference between the low spatial frequency and the unfiltered images are attributable to neural processing of low spatial frequencies.

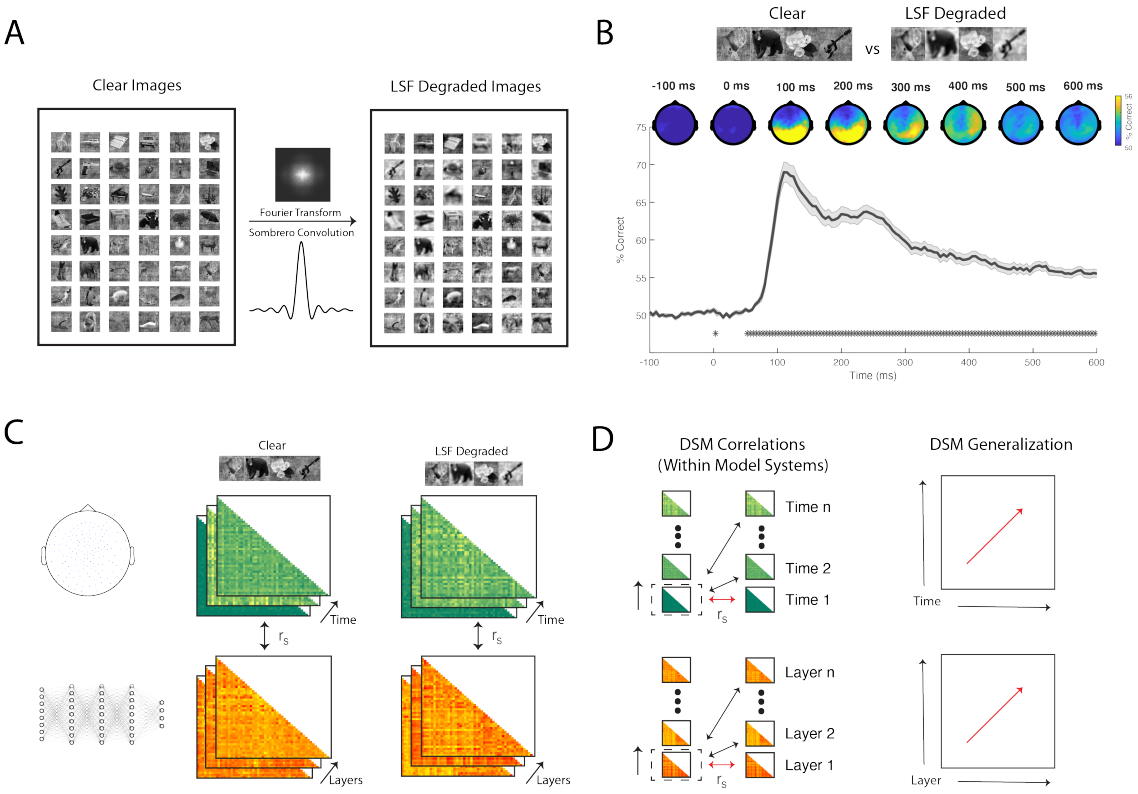

Figure 5.1: Study Design and Analysis Overview. A) Stimuli consisted of 48 achromatic visual objects that included 24 animate and 24 inanimate objects shown in prototypical viewpoints. No human faces were included in the data set. Images were placed on a phase scrambled background. Images were degraded using a Fourier transform and sombrero function to preserve the low spatial frequencies individually calibrated for each image to preserve recognition. (B) Time-resolved decoding plot between clear and degraded images for MEG signals. On the x-axis time in milliseconds; on the y-axis decoding performance. Significance is indicated with asterisks above the abscissa using Wilcoxon signed-rank test against chance decoding (50%), FDR corrected, $q < 0.025$. On the top of the plot, exploratory searchlight analysis shows the topographic distribution of the decoding performance in time. (C) Representational Dissimilarity matrices were calculated using LDA 4-fold cross validation MEG signals and across layers using squared Euclidean distance for each layer activation. RDMs in time and across layers were correlated between MEG and neural networks (D) The evolution of the signal was assessed by correlating RDMs iteratively across all timepoints for MEG signals and layers for neural network activation, creating a RDM generalization matrix.

5.4.2 Relative responses to unfiltered and low spatial frequency images are similar between neural networks and brains

We investigated the dynamics of how clear and degraded images were processed within brain signals and within CNNs by correlating RDMs for each respective model system (Figure 2A-B). In the correlation plots (Figure 2C-D), dark blue indicates low correlation between RDMs and bright yellow indicates higher correlations between RDMs. Qualitatively, we found similarities in the ways that both brain signals and CNNs process images (Figure 2C). While there were correlations in neighboring time points as well as between layers, there appeared to be a chain-like sequential processing of stimuli, such that the representational dynamics changed in time and across layers and no longer shared correlations to earlier times or layers. However, there were some notable differences from this general pattern. For example, CORnet-S had more shared similarity in shallow layers than deeper layers, a pattern that was in contrast to brain responses. For degraded images (Figure 2D), we found similar dynamics but found that there was relatively less correlation between neural signals and time as well as between CNN layers in architectures such as CORnet-S.

Of greatest relevance for our purposes, we computed the correlations between clear images and degraded images for both brain signals and CNN architectures (Figure 2E). Here, we found that degraded image information appears closely related to the information found in clear images at approximately 200ms. Moreover, the various CNN architectures embody this clear-degraded relationship to different degrees. To assess the similarity in the relative dynamics between CNN and brain signals for clear and degraded images, we calculated a similarity score by computing the squared Euclidean distance. We subtracted the total distance from one, such that higher scores indicate more similarity and lower scores indicate less similarity. The scores indicated that of all of the networks tested, VGG-19 had the most similar dynamic relationship between clear and degraded images to the brain. Overall, these results show that the dynamics and processing of clear and degraded images are similar between CNNs and brains.

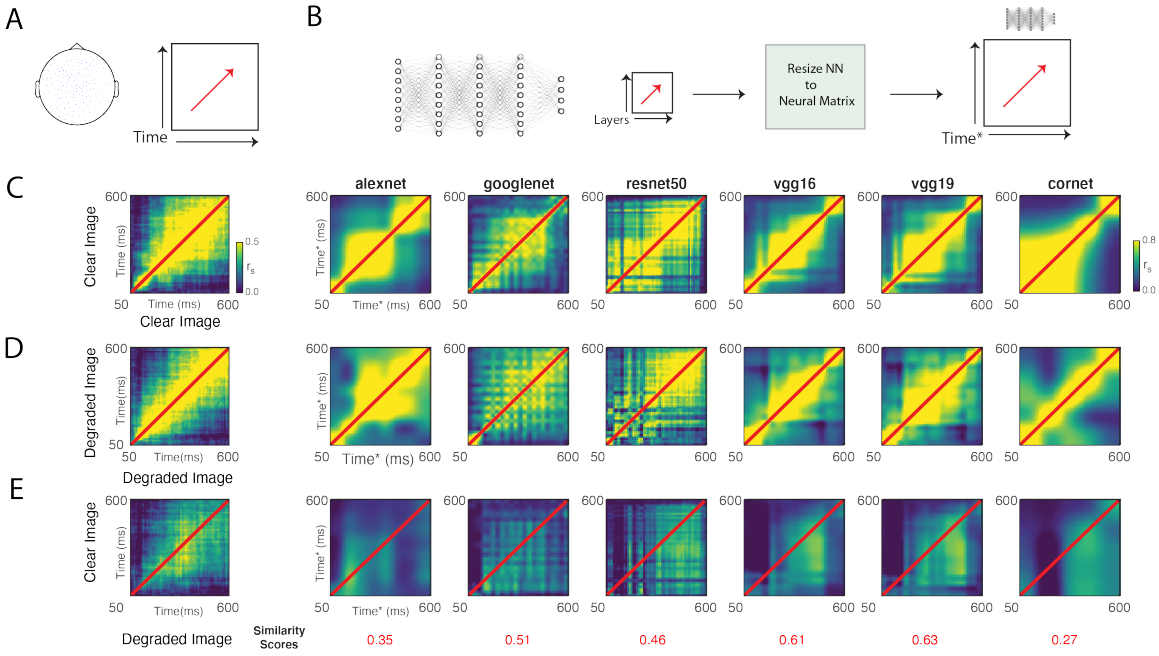

Figure 5.2: Image processing dynamics to clear and degraded images in brains and networks. (A) Correlations between neural RDMs (using all channels) for clear, degraded, and between conditions were calculated using a Spearman correlation. To compare the temporal evolution of the signals in time between neural RDMs and network RDMs, the neural RDM time interval was restricted to approximately when the signals first appear in V1. (B) Neural network RDMs across a wide assortment of networks that include shallow CNNs and deep CNNs, skip connections, as well as recurrence. To compare to the neural RDMs, the RDMs of the various networks were scaled to match the dimensions of the neural RDM using a nearest neighbor interpolation. (C-E) Resulting RDM correlation matrices with neural RDMs on the leftmost column and CNN RDMS to the right for clear images (C) degraded images (D), and the cross correspondence between clear and degraded images (E). For panel E, the similarity score was calculated as (1-squared Euclidean distance) between neural RDMs and CNN RDMs shown on the abscissa.

5.4.3 Degraded images lead to earlier brain CNN correspondence with deeper CNN layers

To directly assess correspondence between CNN activations and brain signals, we used a Spearman correlation to correlate the RDMs for each layer across CNNs in time. In general, we noted emergence of similar patterns across network architectures (Figure 3B and Supplemental Figure 3). For the clear images, there exist distinct peaks of correspondence for the shallow and deeper layers within the CNNs. In contrast, for the low spatial frequency degraded images, only a single peak was evident. We next quantified this observation by measuring the time at which the maximum correlation for each of the layers emerged (Figure 3C). First, we found that there was a positive correlation between layer depth and time across all architectures and across both types of image presentations with the exception of AlexNet with degraded images. Additionally, we found a steeper slope and higher degree of correlation for the clear images when compared with the degraded images across CNN architectures. We next measured the total amount of explained variance maximally achieved by each network architecture across each layer (Figure 3D). Using this approach, we found that the largest differences between clear and degraded images arose within the shallow layers.

In comparing across the various CNN architectures, we note some subtle differences between them. Deeper CNNs (i.e., those with more layers) as well as those that included recurrence showed sustained correlations later in the signal for deeper layers. For example, the last layer of CORnet-S had the highest explained variance (58.3%) at stimulus offset (500ms) for clear images. In comparison, the highest explained variance for ResNet-50 for clear images (51.0%) was seen 100 ms post stimulus offset (600ms). The highest explained variance regardless of layer depth or image presentation was found for ResNet-50 at 160ms in layer 17—a convolutional layer (res3c_branch2a). The high degree of explained variance (92%) was seen for degraded images. Collectively, the observed pattern of results imply that recurrence and deeper layers allow CNNs to be better models at higher stages of visual

processing, agreeing with previous studies (Kietzmann, Spoerer, et al., 2019). Overall, we found that restricting an image to low spatial frequencies led to earlier CNN-Brain correspondence for the deeper layers when compared with clear images.

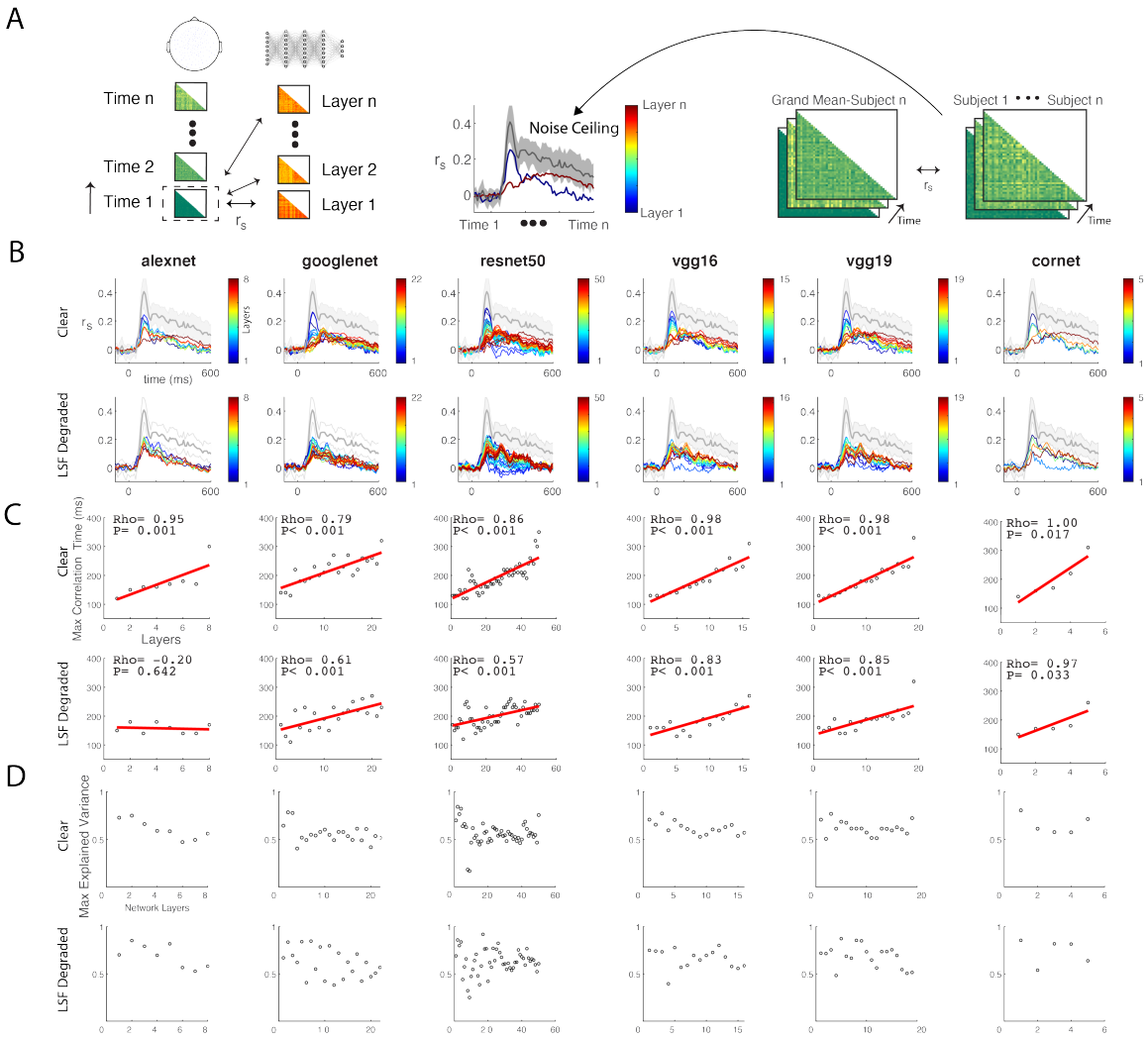

Figure 5.3: Temporal correspondence between MEG and Convolutional Neural Networks. (A) Schematic of the calculation to measure representational correspondence between MEG and CNNs. Spearman correlations were calculated iteratively in time between each participant's MEG RDMs in 10ms increments from -100 to 600ms and across CNN RDMs derived from layer activations. Lower bound of the noise ceiling was calculated by iteratively correlating individual RDMs to the group mean RDM, excluding the individual RDM. Standard deviation is shown as shading around noise ceiling. (B) Time-resolved neural-CNN correspondence with x-axis as time in milliseconds and y-axis as Spearman rho. Color indicates CNN layer depth with blue representing shallow layers and red representing deep layers. (C) Top and bottom row show the time of maximum correspondence for each of the network layers with layers on x-axis and time in milliseconds on the y-axis. (D) Maximum explained variance calculated by neural-CNN correspondence divided by the lower bound of the noise ceiling for each CNN layer.

5.4.4 Topographic CNN-Brain correspondence differs between clear and degraded images

Next, we quantified topographic correspondence between brain signals and CNNs by using a moving searchlight analysis to create electrode specific RDMs from the MEG signal. For this analysis, we limited the correlations to CORnet-S in our main analysis (Figure 4A) and ResNet-50 as a supplemental analysis. We chose CORnet-S due to the differences noted in the timings in the later layers in the previous section, its recurrence connections, and its relatively lower number of layers compared to other networks, allowing for easier visualization of the CNN-Brain correspondence. At 110ms, we find that CNN-Brain correspondence is primarily localized to the occipital MEG sensors across all CORnet-S layers. When we look at significant differences between the clear and degraded images (Figure 4D), we find that the correlation is significantly stronger for the clear images in layers V1 and V2 of CORnet-S. In comparison, the degraded images have stronger correspondence to frontal sensors, including sensors over orbitofrontal cortex. Over time, this pattern begins to change in such that CORnet-S layers V4 and IT show overall stronger correspondence with degraded images, including occipital sensors. Progressing forward in time, we find that the clear image correspondence stays fairly localized to visual cortex while the degraded image correspondence becomes more diffuse. This difference becomes most apparent at 210ms in nearly all layers except for layer V2. Progressing yet further in time, the CNN-Brain correspondence in later network layers is now lateralized to the right hemisphere and the differences between clear and degraded images become less apparent. However, topographic differences still exist in layer V1 with degraded images showing strong correspondence with frontal sensors and clear images showing some small localized increased correspondence in the right lateralized sensors. Together, these results show how the CNN-Brain correspondence in both time and across layers changes depending on whether participants and CNNs are processing clear images vs degraded images.

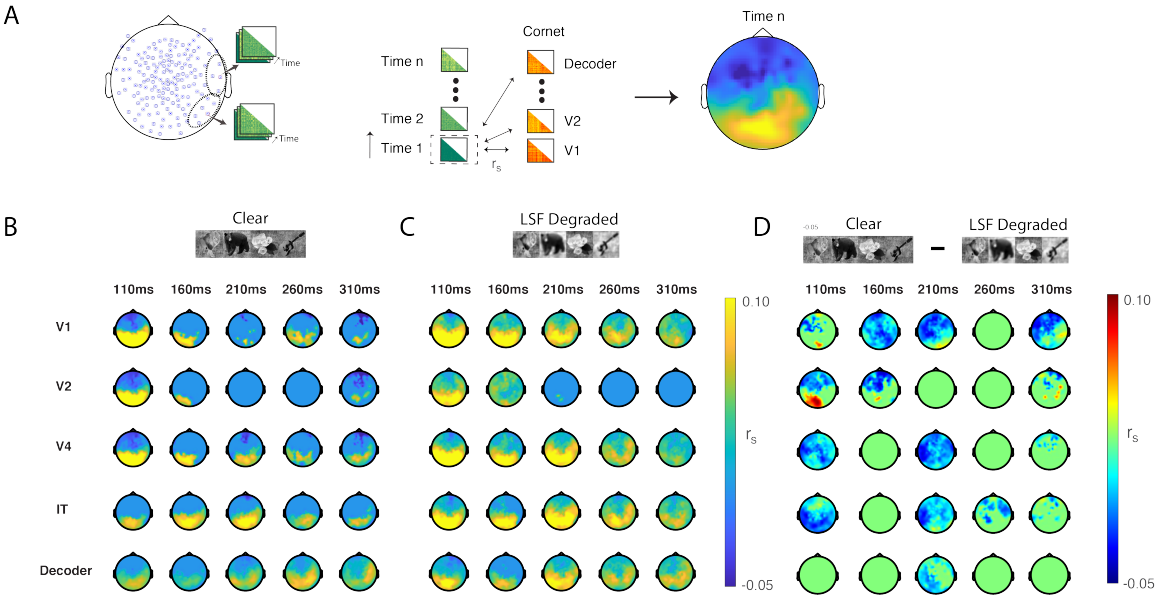

Figure 5.4: Topographic correspondence between MEG and CORnet-S. (A) Schematic of the searchlight procedure used to build electrode specific RDMs that can then be used to assess time resolved topographic correspondence between MEG and CORnet-S. (B-C) Topographic correspondence between all layers of CORnet-S at representative time periods to display how correspondence across MEG electrodes evolves in time. All correspondence is thresholded for significance using a Wilcoxon signed-rank test across participants against a null correlation, FDR corrected for multiple comparisons, $q<0.05$. (D) Differences in neural network correspondence between clear and degraded images at the same representative times, similarly thresholded for significance and corrected for multiple comparisons as in (B) and (C).

5.4.5 Spectral CNN-Brain correspondence differs between clear and degraded images at different CORnet-S layers

Given the known existence of functional differences in information processing between frequency bands, such as gamma being more associated with feedforward processing and alpha and beta being more associated with feedback processing (Bastos et al., 2015; Belitski et al., 2008; Van Kerkoerle et al., 2014), we tested how correlations between brain signals and CNNs varied as a function of frequency. We again chose CORnet-S as the CNN for the reasons cited earlier. To extract frequency specific data, we used bidirectional Butterworth filters (Maier, Aura, & Leopold, 2011), capturing 3 Hz bands over the frequency range spanning 5 Hz to 45 Hz. After the results were tested for significance against zero correlation using a Wilcoxon signed rank test with FDR correction for multiple comparisons (Figure 5B and 5C), we found a general pattern emerge. In this pattern, early CORnet-S layers sharing broadband correspondence to brain signals, especially during the transient response for both clear and degraded images. Following this transient, frequency bands below 30 Hz captured the most correspondence between signals. Progressing into deeper CORnet-S layers, the correspondence was primarily localized to the lower frequency bands (<15hz).

In this frequency analysis, the difference between clear and degraded image correspondence showed a dissociation between network layers. The V2 layer in CORnet-S had higher correspondence in low frequency bands (< 30 Hz) for clear images than for degraded images. However, in deeper layers, specifically layer IT, degraded images had higher correspondence extending into the gamma range (30 - 45 Hz) for the transient peak. This advantage for degraded images was observed at the final decoder layers at low frequency bands (<30hz) during the sustained response, especially in the lowest frequency bands tested (5 Hz). In general, these findings support the notion that early CNN layers are more closely tuned to features that are present in brain signals of clear images but are missing from degraded images. In contrast, the degraded images, which still contain the conceptual aspects

of the image, correspond more with the later layers of a neural network at low frequencies.

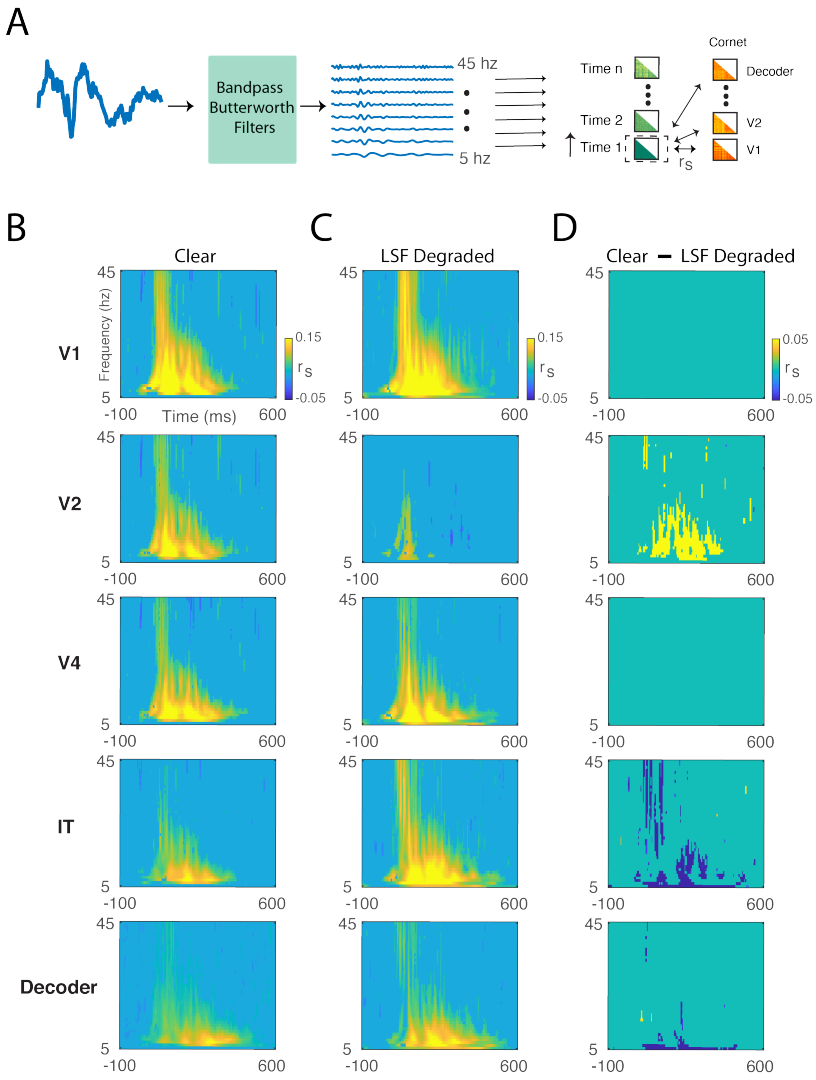

Figure 5.5: Time frequency correspondence between MEG and CORnet-S. (A) Schematic of time frequency analysis using 2nd order Butterworth filters to iteratively filter out frequency components. A 3 Hz sliding window, moving 1 Hz at a time until all frequency bands between 5-45 Hz were extracted. RDMs were then constructed at each frequency and correspondence between MEG frequency and CORnet-S was assessed (B-C) Time frequency correspondence between MEG signals and CORNet-S from shallow (top) to deep (bottom), thresholded for significance against 0 and corrected for multiple comparisons ($q<0.05$) for both clear (B) and degraded (C) images. (D) Significant difference between clear and degraded images for MEG-CORnet correspondence and corrected for multiple comparisons ($q<0.05$).

5.4.6 Training CNNs with stylized images disrupts CNN-brain correspondence

Previous studies have shown that training CNNs with images of varying levels of abstraction shifts the focus of the CNN (such as StyleNet) more to shape rather than texture (Geirhos et al., 2019). Here, we tested the CNN-Brain correspondence for StyleNet and BlurNet. StyleNet is a ResNet-50 architecture trained on a stylized ImageNet image composed a wide variety of artistic styles. BlurNet is an AlexNet architecture that was trained using the same low spatial frequency manipulation used in the current study. As shown in figure 6B, we find that the dynamics in StyleNet are different than those seen in the brain, with each layer having shared representations with other layers. This pattern is seen for clear images as well as degraded images. Furthermore, when looking at the relationship between clear and degraded images, we find that clear-degraded generalization pattern in CNNs is different than the clear-degraded generalization pattern in the brain signals. Specifically, the RDMs for clear images in deeper layers correlate with degraded images across all layers for the CNNs but not for the brain. BlurNet showed similar dynamics for clear images as StyleNet with widely shared representations following the initial layers. Interestingly with the degraded images, there was a more chain-like dynamic as observed in the CNNs in Figure 2. However, the clear-degraded generalization pattern was again different from what was observed in the brain signals.

When looking at the direct CNN-Brain correspondence, we see that all of the layers correspond to early times within the MEG signal (Figure 6C). This result most likely reflects the shared correlation between the layers shown in Figure 6B. For clear images, the explained variance at later times dropped from what was found in the ResNet-50 architecture with StyleNet explained variances at layer 50 of 0.5% and -8.0% at 500ms and 600ms. In comparison, the last layer in ImageNet-trained ResNet-50 yielded explained variance of 49.9% and 51.0%. For BlurNet, explained variance to clear images was 10.6% and 11.9% at 500ms and 600ms while AlexNet had explained variances of 36.3% and 21.9%.

For degraded images, the CNN-Brain correspondence decreased for StyleNet but im-

proved for BlurNet. StyleNet had explained variance of -15.5% and 12.5% at 500ms and 600ms while the comparable values for ResNet-50 at these times was 23.3% and 23.0%. The last layer of BlurNet had explained variance of 31.2% and 25.1% at 500ms and 600ms while the last layer of AlexNet had explained variance of 14.8% and -1.7% at those times. However, there was no longer a direct linear relationship in time and within layers for either StyleNet or BlurNet for degraded images. Lastly, the late layers for both clear and degraded images in StyleNet localized to occipital sensors (Figure 6D) across time points. Overall, these results show that training a neural network with stylized images leads to poor correspondence with brain responses, especially for signals in the later portions of the evoked MEG response to stimuli; the notable exception to this generalization are results from BlurNet on degraded images.

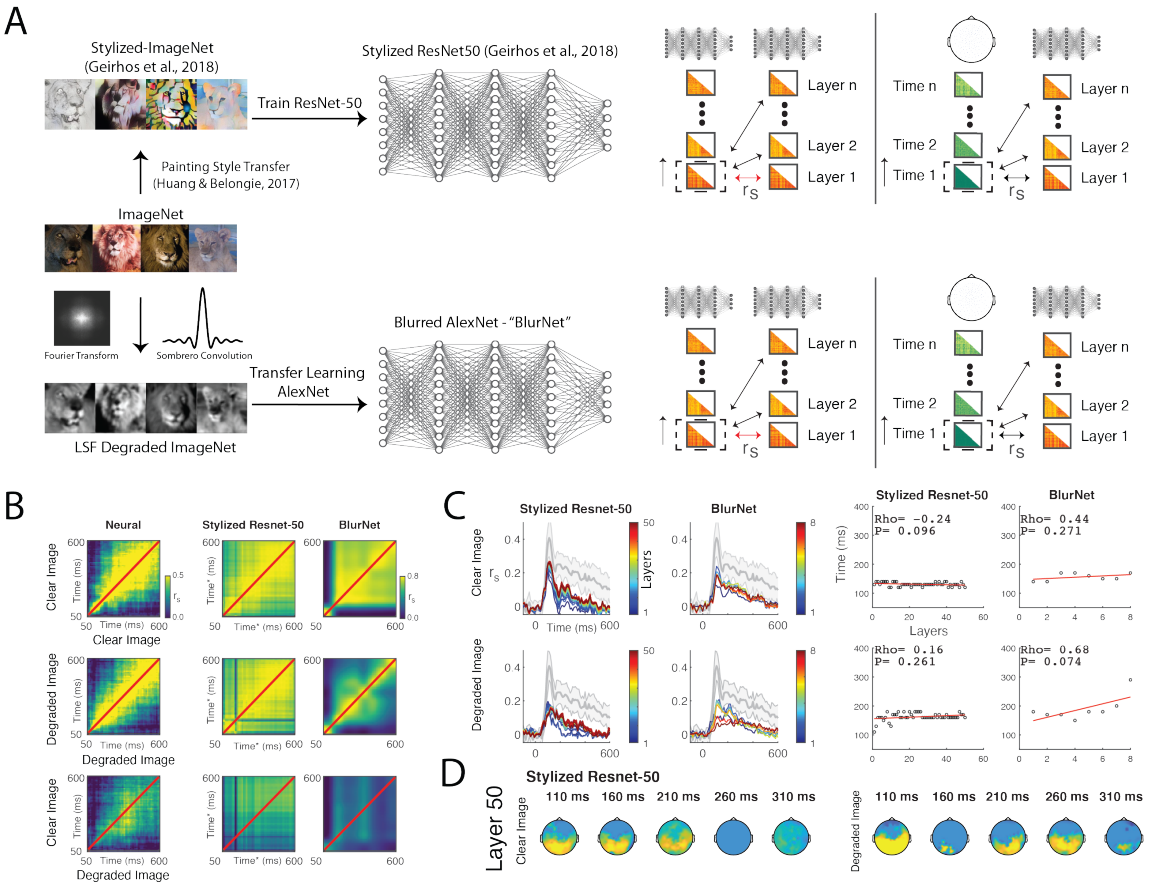

Figure 5.6: Assessing MEG-CNN correspondence with CNNs trained on stylized and low spatial frequency degraded images. (A) Schematic of procedures used to assess CNN trained on either a stylized version of ImageNet or on reduced ImageNet (250 categories) using LSF degraded image. Similar procedures as described in previous figures were then used to assess neural network correspondence. (B) Image processing dynamics as done in Figure 2 shown for StyleNet (Stylized ResNet-50) and BlurNet as well as previously shown neural dynamics for reference. (C) Temporal correspondence between modified CNNs and MEG (left panel) along with times of best correspondence for each layer. (D) Topographic correspondence for the last layers of StyleNet (Stylized-ResNet-50). See Supplemental Figure 4 for reference of ResNet-50 trained without image stylization.

5.5 Discussion

In this study, we investigated the effects of image perturbations, notably LSF blurring, on how well CNNs modeled dynamic brain signals by measuring layer-by-layer correspondence between CNNs and the time resolved MEG signal. The major finding of the study is that CNN-Brain correspondence emerged earlier in time when images were degraded than when they were clear. When comparing brain activity associated with viewing clear vs. degraded images, we found that decoding was lateralized to MEG sensors in the right hemisphere, the brain hemisphere that preferentially processes low spatial frequency visual information (Flevaris & Robertson, 2016; Kauffmann et al., 2014; Schyns & Oliva, 1999). These results suggest that the earlier CNN-Brain correspondence to degraded images is primarily driven by differences in how the brain processes low spatial frequencies. The absence of high spatial frequency content in the blurred images effectively boosted the impact of the low spatial frequency information on brain activity, perhaps through what Bar (2021) refers to as "initial guesses" about what one is viewing that is signaled via feedback from higher brain areas. The CNN-Brain topographic results further fit within a broader coarse-to-fine theoretical framework (Bar, 2003b, 2021; Goddard, Carlson, Dermody, & Woolgar, 2016; Kauffmann et al., 2015; Lu et al., 2018) in which we find correspondence between early visual sensory areas and shallow CNN layers early in time for clear unfiltered images while degraded low spatial frequency images have stronger correspondence to deeper CNN layers and MEG sensors in frontal areas soon after stimulus presentation.

Our findings are at odds with recent fMRI results pointing to shared Brain-CNN correspondence within low level visual areas but not high level visual areas, and particularly decreased correspondence with degraded images (Xu & Vaziri-Pashkam, 2021). We believe these apparent contradictions are attributable, at least in part, to the temporal fine structure that can be resolved in MEG signals but not in fMRI BOLD signal. For example, Xu and Vaziri-Pashkam (2021) found that ResNet-50 was one of the only CNNs that had shared correspondence with higher level visual areas. Similarly, we found that ResNet-50

accounted for the greatest variance in later times of the MEG evoked response for clear images (i.e.,100ms following stimulus offset). However, by using the time-resolved MEG signal, we also found that earlier correspondence for degraded images was localized in MEG parietal and frontal sensors. Thus, we conjecture that fMRI studies are unable to resolve this aspect of CNN-Brain correspondence owing to the sluggishness of the BOLD response. In turn, this suggests that the fMRI signal is likely to be unable to register signals associated with recurrent dynamics, signals that are best captured with recurrent CNNs. Indeed, our study showed that CORnet-S improved late brain-fMRI correspondence compared with other CNNs that did not have recurrent connections, in agreement with previous work (Kietzmann, Spoerer, et al., 2019).

Beyond demonstrating that aspects of CNN-Brain correspondence may be obscured within the sluggish BOLD signals measured using fMRI, MEG studies reveal a key temporal correspondence between brain signals and CNN layers: early brain signals correspond to shallow CNN layers and late brain signals correspond to deep CNN layers (Cichy et al., 2016; Greene & Hansen, 2018; Kietzmann, Spoerer, et al., 2019; Kong et al., 2020; Seeliger et al., 2018). Thus, information is lost if we do not account for the time varying signals that the brain uses (Carlson et al., 2013; Cichy et al., 2014) when measuring correspondence to object processing in the layers of a CNN. In our study, we found such temporal correspondence but further leveraged this relationship and specifically probed the dynamics in time and between CNN layers by generalizing RDMs in time as well as across layers. We found that not only were there shared dynamics in processing clear and degraded images, but also similarities in the way that CNNs and brains respond to image perturbations. By using dynamics to gauge for similarity in processing dynamics, we were able to learn another important lesson: when CNNs are trained using stylized image sets (Geirhos et al., 2019) or degraded image sets, they no longer share similar processing dynamics as the brain, despite explaining comparable variance during the peak of the MEG signal. From this, we put forth that when modifying training sets to build CNNs that can

serve as better models of the brain (Mehrer et al., 2021), measuring dynamics to image perturbations may serve as an effective metric to index CNN-brain correspondence.

The dynamic MEG signal also allows one to probe how correspondence between brains and CNNs may change as a function of brain oscillations. We found correspondence was strongest between the V2 layers of CORnet-S and MEG signals for clear images during the sustained response and predominated in the alpha/beta range to lower theta frequency bands. However, this pattern reversed with higher correspondence in deeper CORnet-S layers for degraded images in the gamma band during the transient and theta band during the sustained response. These findings are consistent with earlier work showing that low spatial frequency image information is preferentially carried in gamma bands while higher frequency image information is preferentially carried in alpha bands (Bar, Kassam, Ghuman, Boyshan, et al., 2006; Flevaris & Robertson, 2016; Fründ, Busch, Körner, Schadow, & Herrmann, 2007). Additionally, gamma band oscillations have also been linked with magnocellular and dorsal stream activity (Merigan & Maunsell, 1993; Tootell et al., 1988), which ostensibly carry the coarse information in the coarse-to-fine processing framework (Bar, Kassam, Ghuman, Boshuan, et al., 2006). The differences found between frequency bands in MEG signals provides motivation to further investigate the correspondence in laminar and direct local field potential (LFP) recordings, which have shown rich frequency specific LFP differences in feedforward and feedback processes within localized circuits (Bastos et al., 2012; Bastos et al., 2015; Maier et al., 2011; Mineault, Zanos, & Pack, 2013; Van Kerkoerle et al., 2014). For such studies, there are a number of potential targets including the distinct magno- and parvocellular layers in LGN (Poltoratski, Ling, McCormack, & Tong, 2017; Tootell et al., 1988), V1 layers where spatial frequency continues to be dissociated between layers 4Cb and 4Ca respectively (Tootell et al., 1988), as well as area V4 which contains separate low spatial frequency and high spatial frequency domains (Lu et al., 2018).

The general tendency we observed for deep CNN layers to show higher correspon-

dence with degraded images earlier in time may point to categorical commonalities between CNNs and brains that are largely missing at the exemplar level (Rajalingham et al., 2018). Since low spatial frequency images prompt the processing of visual images at the superordinate level, which defines category wide attributes (Ashtiani, Kheradpisheh, Masquelier, & Ganjtabesh, 2017), individual CNN-Brain correspondence may become higher as the exemplar become less distinguishable and the images are reduced to possible membership in broad categories. In addition, correspondence could be improved through modifications to CNNs that create more stable exemplar representations. For example, a recent study found that exemplar representations vary between network initializations (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020), and that averaging across several different initializations can improve CNN representations. Alternatively, CNNs trained on datasets that include object categories that are more relevant to humans rather than those comprising ImageNet, which includes an overemphasis on categories such as dog breeds, could also provide more brain-like exemplar representations (Mehrer et al., 2021). Finally, another potential avenue to explore are CNNs that have been trained on sets of low spatial frequency images with decreasing degrees of blur, thus simulating visual development in infants. CNNs trained in that way have shown better performance than CNNs trained on unblurred images from the outset, leading to the speculation that graded training makes the CNN more brain-like (Avbersek, Zeman, & Op de Beeck, 2021). Probing different modifications to CNN training paradigms will be essential in testing how the image statistics in trainings affect the Brain-CNN correspondence across a number of different image perturbations and differing levels of occlusion (Rajaei, Mohsenzadeh, Ebrahimpour, & Khaligh-Razavi, 2019; Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Geiger, et al., 2018).

5.5.1 Conclusion

In conclusion, we have provided evidence of earlier correspondence between brains and deep CNN layers in degraded images that support the coarse-to-fine conceptual framework of visual image processing. In addition, we have provided a rich methodological framework by introducing a number of analyses that can be used to assess the dynamics of CNNs and compare these with brain activity across the dimensions of space, time, and frequency. This framework can be extended to include a number of image perturbations as we test the limits of CNN-brain correspondence with CNNs that are purposefully created to be more brain-like (Kubilius et al., 2018) or those that inadvertently become so (Schrimpf, Kubilius, Hong, Majaj, Rajalingham, Issa, Kar, Bashivan, Prescott-Roy, Geiger, et al., 2018). Finally, there are a number of potentially revealing experimental manipulations that could enhance efforts to examine possible CNN-brain correspondence. Those include manipulations of stimulus duration (Grootswagers, Robinson, & Carlson, 2019), creation of visual stimuli comprising object textures devoid of explicit shapes (Grootswagers, Robinson, Shatek, & Carlson, 2019; Long, Yu, & Konkle, 2018), visual images that are accompanied by congruent or incongruent sounds (Tovar, Murray, & Wallace, 2020), and creation of hybrid stimuli consisting of conflicting low spatial frequency and high spatial frequency information (Schyns & Oliva, 1999). These kinds of manipulations, together with expanded CNN architectures and training sets, will push the boundaries of understanding of the potential correspondence between brains and CNNs.

5.6 References

Ashtiani, M. N., Kheradpisheh, S. R., Masquelier, T., and Ganjtabesh, M. (2017). Object categorization in finer levels relies more on higher spatial frequencies and takes longer. *Frontiers in Physiology*, 8(JUL):1261.

Avbersek, L. K., Zeman, A., and Op de Beeck, H. (2021). Training for object recognition with increasing spatial frequency : A comparison of deep learning with human vision . *bioRxiv*.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15(4):600–609.

Bar, M. (2021). From objects to unified minds. *Current Directions in Psychological Science*.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshuan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M., Marinkovic, K., Schacter, D. L., Rosen, B. R., and Halgren, E. (2006a). Top-down facilitation of visual recognition. *PNAS*, 103(2):449–454.

Bar, M., Kassam, K. S., Ghuman, A. S., Boyshan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M., Marinkovic, K., Schacter, D. L., Rosen, B. R., and Halgren, E. (2006b). A 'missing' family of classical orthogonal polynomials. *PNAS*, 103(2):449–454.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., DeWeerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401.

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2008). Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *Journal of Neuroscience*, 28(22):5696–5709.

Bruner, J. S. and Potter, M. C. (1964). Interference in visual recognition. *Science*, 144(3617):424–425.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10(12).

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10):1–19.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(June):1–13.

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462.

Duda, R. and Hart, P. (2001). *Pattern Classification*. 2nd edition.

Etzel, J. A., Zacks, J. M., and Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, 78.

Flevaris, V. A. and Robertson, L. C. (2016). Spatial frequency selection and integration of global and local information in visual processing: A selective review and tribute to shlomo bentin. *Neuropsychologia*, 83:192–200.

Fründ, I., Busch, N. A., Körner, U., Schadow, J., and Herrmann, C. S. (2007). Eeg oscillations in the gamma and alpha range respond differently to spatial frequency. *Vision Research*, 47(15):2086–2098.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture. *Iclr*, (c):1–20.

Geirhos, R., Schütt, H. H., Medina Temme, C. R., Bethge, M., Rauber, J., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS 2018):7538–7550.

Goddard, E., Carlson, T. A., Dermody, N., and Woolgar, A. (2016). Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage*, 128:385–397.

Greene, M. R. and Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Computational Biology*, 14(7):1–17.

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., and Carlson, T. A. (2017). Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions. *Journal of cognitive neuroscience*, 29(12):1995–2010.

Grootswagers, T. and Robinson, A. K. (2021). Overfitting the literature to one set of stimuli and data. *arXiv*, pages 3–8.

Grootswagers, T., Robinson, A. K., and Carlson, T. A. (2019a). The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188(October 2018):668–679.

Grootswagers, T., Robinson, A. K., Shatek, S. M., and Carlson, T. A. (2019b). Untangling featural and conceptual object representations. *NeuroImage*, 202(July):116083.

Güçlü, U. and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Güçlü, U. and van Gerven, M. A. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–17.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:1510–1519.

Hubel, D. H. and Wiesel, T. N. (1977). Ferrier lecture. functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London - Biological Sciences*, 190(1130):1–59.

Kadar, I. and Ben-shahar, O. (2012). A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of vision*, 12(13):1–17.

Kauffmann, L., Ramanoël, S., Guyader, N., Chauvin, A., and Peyrin, C. (2015). Spatial frequency processing in scene-selective cortical regions. *NeuroImage*, 112:86–95.

Kauffmann, L., Ramanoël, S., and Peyrin, C. (2014). The neural bases of spatial frequency processing during scene perception. *Frontiers in Integrative Neuroscience*, 8(MAY):1–14.

Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11).

Kietzmann, T. C., Mcclure, P., and Kriegeskorte, N. (2019a). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019b). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, 116(43):21854–21863.

Kong, N. C., Kaneshiro, B., Yamins, D. L., and Norcia, A. M. (2020). Time-resolved correspondences between deep neural network layers and eeg measurements in object processing. *Vision research*, 172(May):27–45.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):4.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, 12(4):1–26.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., and Dicarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, pages 1–9.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Long, B., Yu, C. P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, 115(38):E9015–E9024.

Lu, Y., Yin, J., Chen, Z., Gong, H., Liu, Y., Qian, L., Li, X., Liu, R., Andolina, I. M., and Wang, W. (2018). Revealing detail along the visual hierarchy: Neural clustering preserves acuity from v1 to v4. *Neuron*, 98(2):417–428.e3.

Maier, A., Aura, C. J., and Leopold, D. A. (2011). Infragranular sources of sustained local field potential responses in macaque primary visual cortex. *Journal of Neuroscience*, 31(6):1971–1980.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences of the United States of America*, 118(8):1–9.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, 11(1):1–12.

Merigan, W. H. and Maunsell, J. H. (1993). How parallel are the primate visual pathways? *Annual Review of Neuroscience*, 16:369–402.

Mineault, P. J., Zanos, T. P., and Pack, C. C. (2013). Local field potentials reflect multiple spatial scales in v4. *Frontiers in Computational Neuroscience*, 7(MAR):1–15.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4).

Petras, K., ten Oever, t. S., Jacobs, C., and Goffaux, V. (2019). Coarse-to-fine information integration in human vision. *NeuroImage*, 186(October 2018):103–112.

Poggio, T. and Riesenhuber, M. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Poltoratski, S., Ling, S., McCormack, D., and Tong, F. (2017). Characterizing the effects of feature salience and top-down attention in the early visual system. *Journal of Neurophysiology*, page jn.00924.2016.

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., and Khaligh-Razavi, S. M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Computational Biology*, 15(5):1–30.

Revina, Y., Petro, L. S., and Muckli, L. (2018). Cortical feedback signals generalise across different spatial frequencies of feedforward inputs. *NeuroImage*, 180(March 2017):280–290.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C.,

Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007.

Schyns, P. G. and Oliva, A. (1999). Dr. angry and mr. smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, 69(3):243–265.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J. M., Bosch, S. E., and van Gerven, M. A. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180(July 2017):253–266.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–6429.

Snodgrass, J. G. and Hirshman, E. (1991). Theoretical explorations of the bruner-potter (1964) interference effect. *Journal of Memory and Language*, 30(10):273–293.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., Hill, C., and Arbor, A. (2014). Going deeper with convolutions. pages 1–9.

Tootell, R. B., Silverman, M. S., Hamilton, S. L., Switkes, E., and De Valois, R. L. (1988). Functional anatomy of macaque striate cortex. v. spatial frequency. *Journal of Neuroscience*, 8(5):1610–1624.

Tovar, D., Murray, M., and Wallace, M. (2020). Selective enhancement of object representations through multisensory integration. *Journal of Neuroscience*, 40(29):5604–5615.

Van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M. A., Poort, J., Van Der Togt, C., and Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(40):14332–14341.

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., and Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, 55(1):143–156.

Wardle, S. G. and Baker, C. (2020). Recent advances in understanding object recognition in the human brain: Deep neural networks, temporal dynamics, and context. *F1000Research*, 9:1–14.

Xu, Y. and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 12(2065):1–16.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

5.7 Supplemental Figures

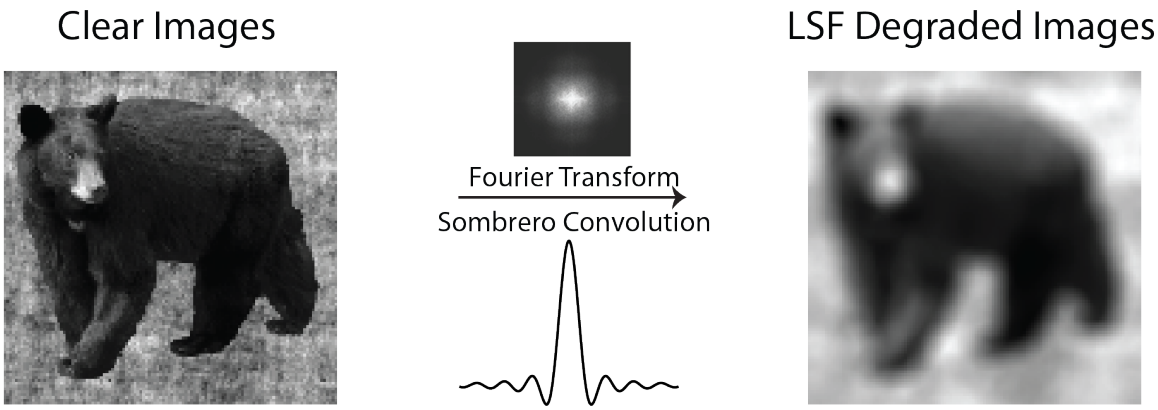

Figure 5.7: Supplemental for Figure 5.1. (Example LSF degradation of one of the exemplars.

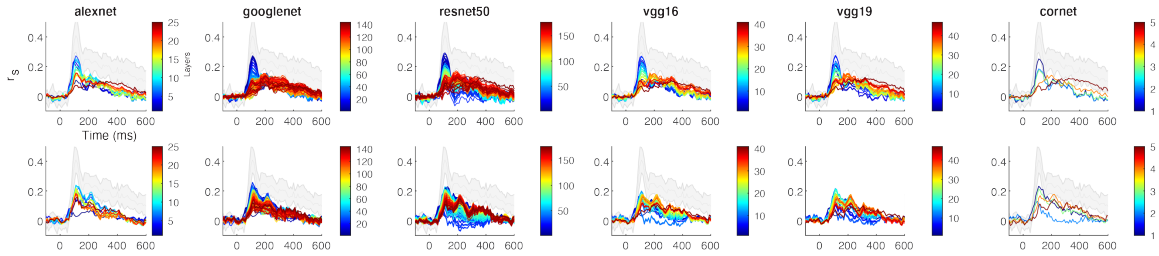

Figure 5.8: Supplemental for Figure 5.3. Correspondence between MEG and CNNs using all operations including pooling, ReLU, and normalization.

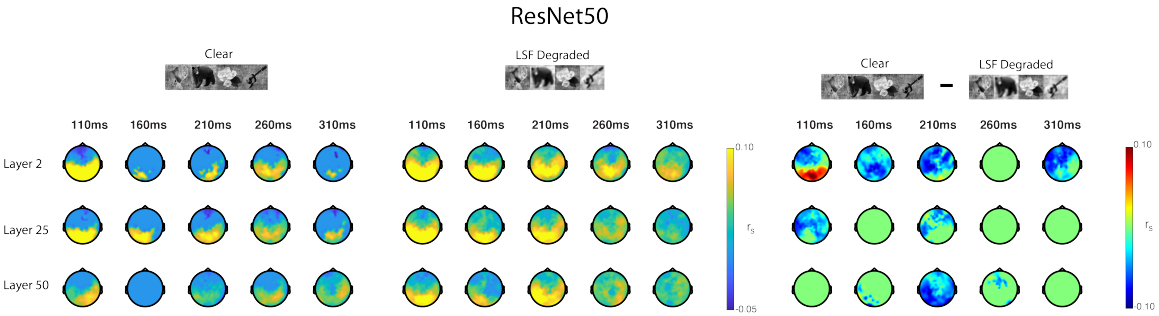

Figure 5.9: Supplemental for Figure 5.4. (Supplemental topographic correspondence between MEG and ResNet50 that largely complement the findings found using CORnet.

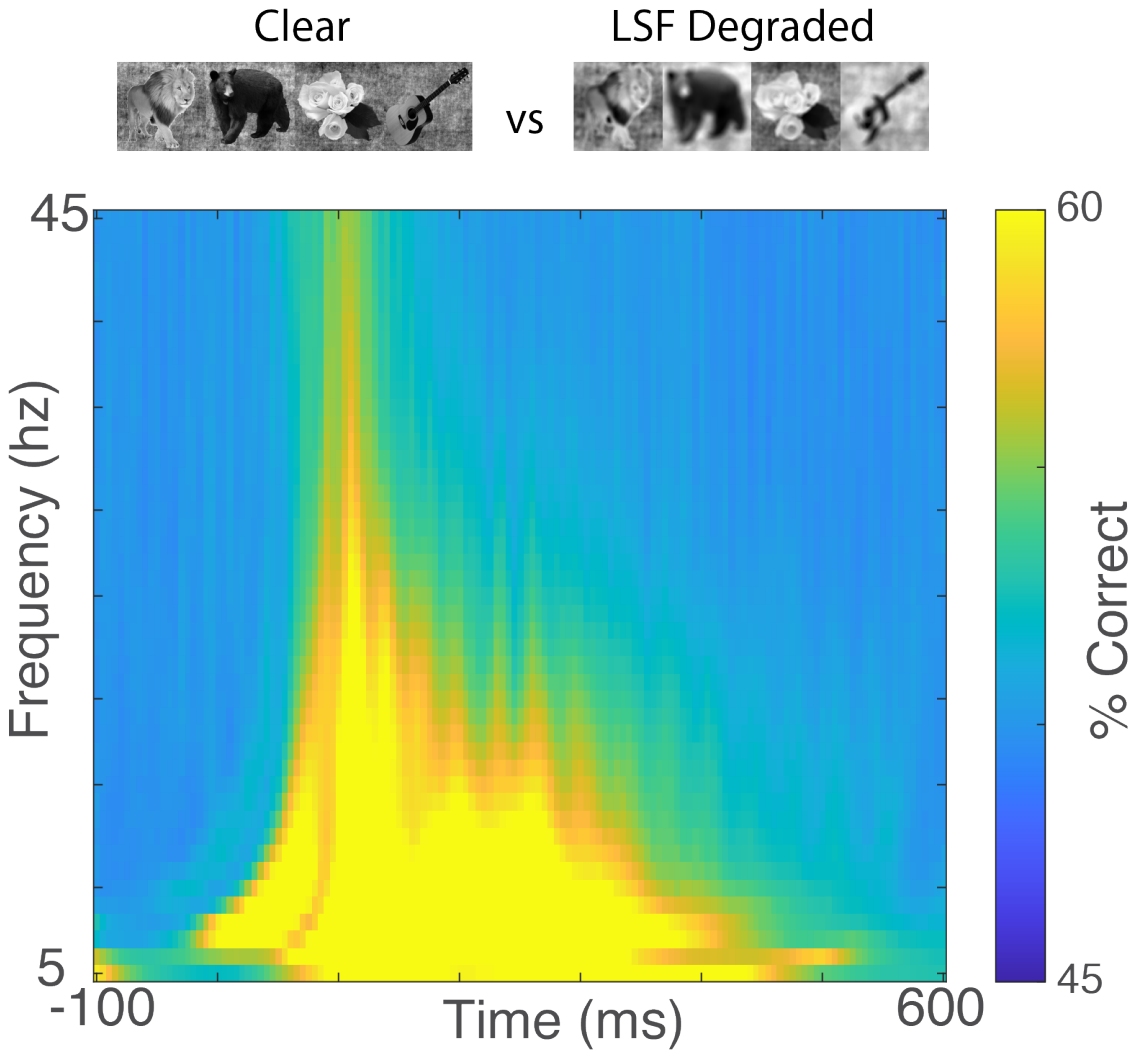

Figure 5.10: Supplemental 1 for Figure 5.5. Time frequency decoding between clear and degraded images, showing the spectrotemporal profile differentiating between the images.

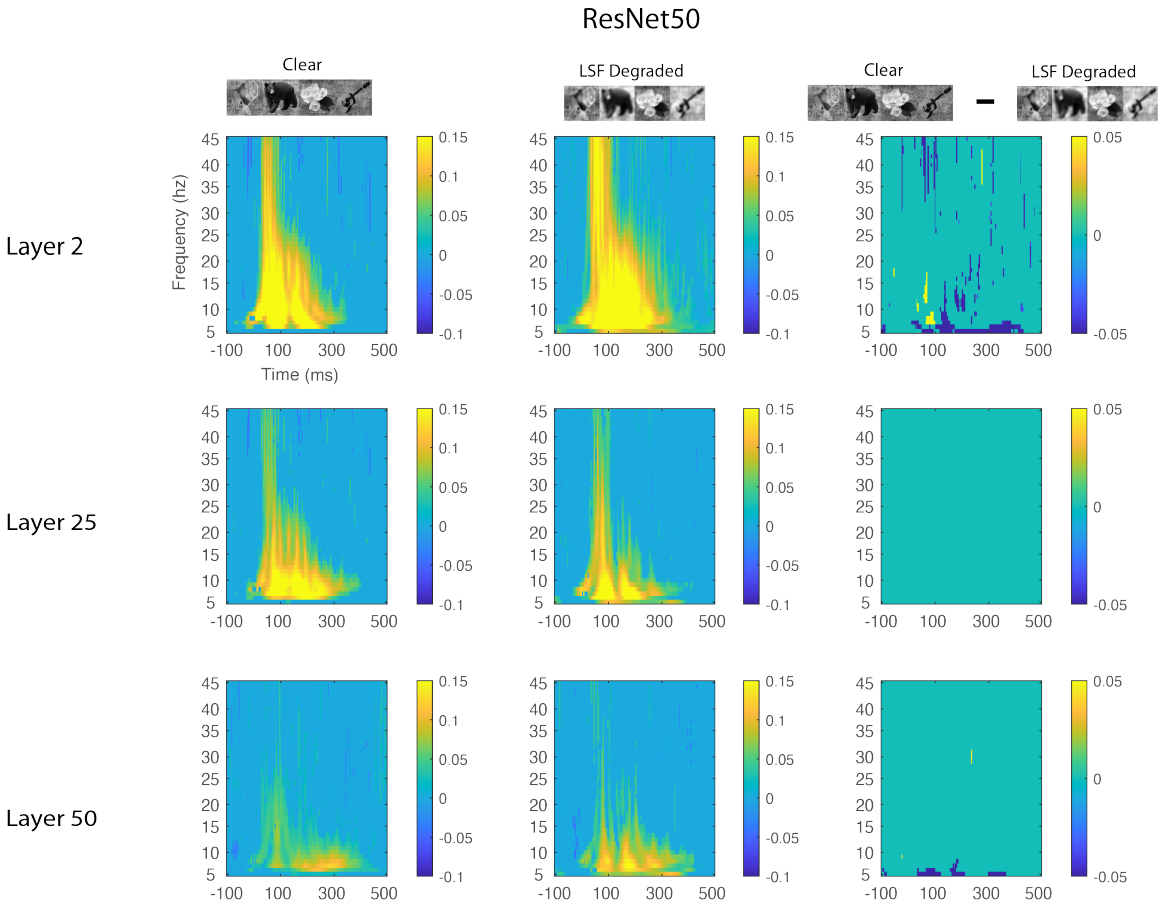

Figure 5.11: Supplemental spectrotemporal correspondence between MEG and ResNet50 that shows some differences between CORnet-S and ResNet50. Namely, that there is higher correspondence between degraded images and ResNet50. However, note that these are not all of the layers in ResNet50, just representative ealy, middle, and layers.

“They fell into a silence. They looked at one another, amazed. This thing they had never really believed in was coming true.”

-John Steinbeck, Of Mice and Men

Chapter 6

General Discussion

By studying perceptual phenomena at three broad levels, one potentially risks having a disjointed group of projects. Indeed, at first glance, my work looking at stimulus feature specific information within the V1 microcircuit seems to be far removed from looking at the effects of image perturbations on brain correspondence with convolutional networks. However, there is a unifying thread between chapters in that I use machine learning methods throughout each chapter to extract information from neural signals. I will elaborate on this theme and provide commentary on why it is a useful analytical method. Furthermore, after summarizing the main findings from each part of the dissertation and contextualizing them within the extant literature, I will then discuss how the findings from each of the studies connects to the other levels. Finally, I will conclude with a broad vision of how I see neuroscience progressing symbiotically with artificial intelligence.

6.1 Multivariate pattern analysis as a method of extracting neural information

One of the ways that neuroscientists have historically sought to understand the brain is by defining discrete areas that contain different functions—such as with Brodmann areas, which divide the brain based on its cytoarchitecture (Judaš, Cepanec, & Sedmak, 2012), to more recent fMRI studies that define areas based on their responses to different objects, such as faces or places (Epstein, Harris, Stanley, & Kanwisher, 1999; Kanwisher, McDermott, & Chun, 1997). This type of modular approach has lent itself well to univariate analyses that average responses over a number of trials and defines areas based on the object or feature that induces the most vigorous overall response. However, these approaches have lacked the granularity to investigate distributed codes that the brain might use to have the necessary computational flexibility to accurately code for a diverse set of stimuli (DiCarlo & Cox, 2007; Haxby et al., 2001). For example, with objects, any given object can have multiple viewpoints, levels of occlusion, lighting, and size. With a distributed code, similar

objects can be teased apart from each other in a variety of viewing conditions (Olshausen & Field, 2004; Quiroga, Kreiman, Koch, & Fried, 2008). Thus, in order to understand the brain with the granularity needed to differentiate distributed codes, we require analyses that do not average over these codes, but rather account for different activity patterns.

In order to be able to account for distributed codes, in this dissertation, I have used multivoxel (multivariate) pattern analysis (MVPA) to extract information from neural signals. These methods have shown that even when modular areas such as the fusiform face area (FFA) and parahippocampal place area (PPA) are removed from fMRI analysis, it is still possible to determine the categories of face and place exemplars based on the surrounding brain areas (Grill-Spector, 2003; Haxby et al., 2001). One of the dangers in using decoding methods as a means of extracting neural information is that it is possible that the codes extracted by the classifier are epiphenomenal in that they can be used for classification but are not necessarily used in their current state by the brain. Ostensibly, all of the neural information from a visual object is hypothetically present beginning at the level of the retina and with a sufficiently complex non-linear classifier, it would be possible to decode this information and falsely attribute a host of functions to the retina. To address these concerns, I have a made a number of analysis choices. First, I have used linear classifiers with the assumption that the brain reduces information or representations such that it can apply a linear read out process (Grootswagers, Cichy, & Carlson, 2018; Tong & Pratte, 2012). Then, where appropriate, I have linked these linear readouts to behavior (Carlson, Ritchie, Kriegeskorte, Durvasula, & Ma, 2014; Grootswagers, Ritchie, Wardle, Heathcote, & Carlson, 2017; Ritchie, Tovar, & Carlson, 2015), such as in Chapter 4 (Tovar, Murray, & Wallace, 2020). Furthermore, I have grounded multivariate results where possible to previous univariate results, as in Chapter 2 (Tovar et al., 2020), providing added confidence to the analytical extensions that could only be found in multivariate analyses.

Thus, I generally view MVPA as an analytical tool that can supplement univariate analyses in probing the distributed code the brain uses to encode information. Furthermore,

beyond providing added granularity, it makes it possible to generalize specific information across a wide assortment of stimulus features (Chapter 2 and 3), object categories (Chapter 4) and relationship between objects (Chapter 5) across modalities, time, frequency, and model systems. MVPA has been widely used in fMRI studies (for review see Tong & Pratte, 2012), with more recent M/EEG applications (for review see Wardle & Baker, 2020) and with the emergence of multichannel arrays with increasingly higher number of electrode contacts, will see expanded use in neurophysiology studies. In the next sections, I will summarize the take home points and lessons we learned using MVPA, particularly in neurophysiology studies (Part 1) and M/EEG (Part 2 and Part 3).

6.2 Part 1

In part 1, we found that stimulus features embedded in spiking data are distributed with unique spatiotemporal profiles within the V1 microcircuit. For example, we found that eye-of-origin information was significantly decreased in the infragranular layers, agreeing with previous studies (Blake & Cormack, 1979; Dougherty, Cox, Westerberg, & Maier, 2019; Hubel & Wiesel, 1977). Information regarding stimulus history was found primarily in the supra- and infragranular layers, consistent with previous reports (Van Kerkoerle et al., 2014; Westerberg, Cox, Dougherty, & Maier, 2019). The agreement with the literature validated the use of MVPA and a moving searchlight analysis as a framework to extract stimulus features from spiking activation sequences. Furthermore, adopting a multivariate pattern framework allowed us to apply time generalization techniques to relate spike patterns in time, something we would have otherwise been unable to do using traditional univariate analysis.

Next, we quantified spatiotemporal profiles for LFP data, specifically focusing on the role of volume conduction on LFP signals. The study added to previous literature that studied volume conduction (Kajikawa & Schroeder, 2011; Katzner et al., 2009; Kreiman et al., 2006; Xing, Yeh, Burns, & Shapley, 2012) by going beyond grand mean responses,

but specifically quantifying which stimulus features were most volume conducted. The need to characterize the features contained in the volume conducted signal was apparent given earlier findings that signal correlation and coherence could affect the amount of volume conduction by an order of magnitude (Leski, Lindén, Tetzlaff, Pettersen, & Einevoll, 2013; Lindén et al., 2011; Rosenbaum, Smith, Kohn, Rubin, & Doiron, 2017). Additionally, given reports that volume conduction differed as a function of frequency (Leski et al., 2013) we used a butterworth bandpass filter to decompose a CSD derived LFP signal to quantify volume conduction. In the process, we further contributed to previous literature that characterized the relationship between LFP frequency bands (Bastos et al., 2015a; Belitski et al., 2008; Van Kerkoerle et al., 2014). Together, these findings provide an analytical framework regarding which specific stimulus features are processed locally.

Comparing the spatiotemporal profile features contained within spiking data with those found to the LFP data in chapters 2 and 3, it is apparent that the spatiotemporal profiles differed, especially in higher frequency bands. These findings agree with a recent study that showed that high gamma colocalizes with spikes in supragranular layers, but not elsewhere (Leszczynski et al., 2020). Planned future work will make the relationship between spikes and LFP signals more explicit. Specifically, I plan to use confusion matrices to make input (LFPcal) and output (spike) inferences for both LFPs and spikes.

One of the biggest limitations to my work is that all the findings are all contained within one laminar probe. However, the analytical framework we have developed can be readily adopted to track how feature information flows from one cortical area to another, as well as within lamina. Additionally, multiple probes could also help in identifying the directionality of signals as done in previous work (Bastos et al., 2015a). Another limitation is that the stimulus conditions were too few in the studies of part 1 and thus limited our ability to make use of tools such as representational similarity analysis (RSA). Using RSA, I could adapt Granger analysis with RDM (Contini, Wardle, & Carlson, 2017) to make further directionality inferences both within and across laminar probes but doing so using

feature specific information rather than just grand mean responses. The findings in part one can thus be succinctly summarized in three broad statements. We: 1) established a MVPA framework to extract stimulus features in time and space from activation sequences, 2) characterized the information found in volume conducted signals and 3) noted differences between spiking and LFP/CSD signals.

6.3 Part 2

In part 2, we learned that inherent noise attributed to an object's membership to a category influenced how much benefit these objects incur in a multisensory context. In this study we specifically used the animate and inanimate distinction, one of the fundamental organizing principles the brain uses to process sensory information (Carlson, Tovar, Alink, & Kriegeskorte, 2013; Grootswagers, Ritchie, Wardle, Heathcote, & Carlson, 2017; Kriegeskorte, Mur, Ruff, & Kiani, 2008; Lindh, Sligte, Assecondi, Shapiro, & Charest, 2019; Ritchie, Tovar, & Carlson, 2015). We made use of previous studies that found that animate objects are processed preferentially over inanimate objects (Jackson & Calvillo, 2013; New, Cosmides, & Tooby, 2007; Vogler & Titchener, 2011). The results were consistent with what would be expected from maximum likelihood estimate models (Ernst & Banks, 2002) in that multi- sensory benefits are most evident when the reliability of the dominant signal is lower. We showed that this reliability weighting was found at the category level apart from low level visual attributes. By using black and white drawings as opposed to realistic images, we avoided many of the critiques surrounding the contribution of texture to the animate/inanimate distinction (Grootswagers, Robinson, Shatek, & Carlson, 2019; Long, Yu, & Konkle, 2018). There are several experimental design and analytical extensions that I would like to add to this study. The first would be to parametrically manipulate the visual and auditory sensory streams by introducing varying levels of noise to assess the interaction between low level and high-level sensory properties. The other would be to manipulate the causal structure (Körding et al., 2007) of the stimuli by

manipulating temporal and spatial synchrony of the visual and auditory component. Along these lines, it would also be beneficial to have video and audio streams and compare these to the static and dynamic coupling of vision and sound in this study. In terms of analytical approaches, I would like to use Granger causality using RDMs (Contini et al., 2017) in order to better quantify whether the multisensory effects were primarily feedforward or feedback.

6.4 Part 3

In Part 3, we investigated how image perturbations, specifically reducing images to their low special frequency components, affected the correspondence between CNNs and MEG signals. By doing this in a time-resolved manner, we were able to find that CNN-Brain correspondence emerged earlier when images were degraded than when they were clear. This finding fits within the broader coarse-to-fine theoretical framework (Bar, 2003, 2021; Goddard, Carlson, Dermody, & Woolgar, 2016; Kauffmann, Ramanoël, Guyader, Chauvin, & Peyrin, 2015; Lu et al., 2018) and additionally shows that low spatial frequency information is emphasized more in the brain for degraded low spatial frequency images than in their clear image counterparts. While other studies have similarly looked at how image degradation affected the correspondence between CNNs and the visual system, they were done with fMRI and as a result missed out on possible temporal effects (Xu, Vaziri, & Pashkam, 2021). In addition, I made use of the temporal structure of the MEG signal in order to compare how visual representations generalize across time, and how these then compare with how representations generalize within network layers. I introduced a number of analyses exploring components of the MEG signal that had previously been untapped. These analyses showed that correspondence varied based on frequency bands and the sensor locations of the MEG signal. While this study is an advance in terms of quantifying correspondence between brains and CNNs, I am just barely scratching the surface of the possible stimuli and analytical manipulations that can be done.

I have immediate plans to quantify the effects of several different analytical steps as previewed in the introduction section 1.3.6. In brief, these include the distance measurements used to build the RDM matrices, the effects of decoding or distance measurements used to create the neural RDMs, task demands on the subjects (i.e. active categorization vs passive viewing) as well as the layer selections from the CNN used to quantify the brain-CNN correspondence. In addition, I would like to measure the opposite of the stimulus manipulation we used in this study, the low spatial frequency blur. But rather than investigate high special frequencies, I will use textforms that are devoid of form, in order to specifically investigate the much-beleaguered role of texture in CNNs (Grootswagers et al., 2019; Long et al., 2018) and its resulting correspondence between brains and CNNs.

6.5 Connecting Part 1 to Part 2

The lessons learned in part 1 can serve as a framework for how we interpret some of the research regarding multisensory integration studied in part 2. Specifically, multisensory integration relies on similar stimulus features being temporally and spatially close in order to be integrated (Meredith, Nemitz, & Stein, 1987; Meredith & Stein, 1986). Much of the early multisensory integration research investigated this at the level of individual neurons (Meredith et al., 1987; Wallace, Meredith, & Stein, 1998). However, we know that these effects are present at the population level (Ma, Beck, Latham, & Pouget, 2006). Thus, the spatiotemporal framework for spikes as well as LFPs in part 1 will be helpful in terms of indexing measures of multisensory integration at the circuit level. Furthermore, using decoding to extract feature information allows us to avoid the fallacy that more activation in an area translates to more information (Harrison & Tong, 2009; Jehee, Brady, & Tong, 2011; Kok, Jehee, & de Lange, 2012; Laurienti, Perrault, Stanford, Wallace, & Stein, 2005).

In addition, the volume conduction results in part 1 serve as a caution to multisensory researchers from relying on LFP signals for source localization. Otherwise, researchers

run the risk of falsely identifying areas as sites of multisensory integration when they are in fact volume conducted signals (Kajikawa, Smiley & Schroeder, 2017). It is for this reason, that we were cautious in our interpretation of representational connectivity in figure 4.5. If direct electrical probes are used instead of EEG, it is much easier to use current source density or a current source density derived signal in order to make localization claims.

6.6 Connecting Part 1 to Part 3

While many of the same caveats found between part 1 and part 2 apply to localizing CNN-Brain correspondence in part 3, the link between CNNs and brain signals carry yet another dimension. While there are a considerable number of studies investigating how neural spikes relate back to layer activations (Kar DiCarlo, 2020; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Bashivan, Kar, & DiCarlo, 2019), the links between CNNs and CSD or even LFP have been largely ignored. Given that is possible to create localized LFP signals like we did in part 1 and decompose the LFP signal into frequency bands, these are missed opportunities. This is especially true given that there are distinct types of information embedded within different frequency bands (Bastos et al., 2015b; Belitski et al., 2008; Van Kerkoerle et al., 2014). Add in the possibility of also doing this analysis with laminar recordings with distinct compartments associated with feedforward and feedback (Van Kerkoerle et al., 2014; Westerberg et al., 2019), and the possibilities are seemingly limitless. Furthermore, there are powerful predictions that can be made between feedforward and recurrent neural networks activations and laminar compartments. For example, one would expect that the granular layer would have the most correspondence with feedforward networks, but a recurrent network might also show correspondence with supragranular and infragranular layers.

6.7 Connecting Part 2 to Part 3

The connections between 2 and part 3 of the dissertation are amongst the most exciting. As CNNs continue to become more sophisticated at modeling auditory brain responses

(Kell et al., 2018; Millet & King, 2021), there is rich opportunity to begin to build audio-visual neural networks for object recognition. One of the current challenges in doing this is finding sufficiently large training data of common audio and visual components. For the time being, much of the audiovisual training data sets are limited to examples of audiovisual speech.

However, there are a group of networks, named capsule networks, that theoretically will not require as much data as traditional CNNs (Doerig, Schmittwilken, Sayim, Manassi, & Herzog, 2020; Sabour, Frosst, & Hinton, 2017). While they are still in their infancy and have issues scaling up to image recognition datasets, they are promising architectures that contain units within the networks that behave somewhat similarly to multisensory neurons. Instead of using pooling as CNNs do, which either averages or takes the max value from a previous convolution, capsule networks use routing by agreement, hierarchically sending votes from the initial feature and therefore preserving more of the original input. Thus, in this way, they behave similarly to a multisensory neuron that is receptive to both visual and auditory input, instead of decidedly having to be one or the other.

6.8 Broad Implications for Neuroscience and AI

The title of this dissertation "Of Machines and Men" implies a symbiotic relationship between the brain and artificial neural networks/machine learning. Given my initial research interests and the fact that this is a neuroscience dissertation, my telling of the story has been predictably one sided. The chapters I have written are example cases of the ways we can use machine learning and artificial neural networks to improve our understanding of the brain. However, if we are to use the rich contributions of simple and complex cells in the genesis of convolutional neural networks (Lecun, Bengio, & Hinton, 2015), as a guide, there are still lessons that artificial systems can learn from the brain. Thus, the discussion here will focus on the symbiotic relationship between machines and men, briefly providing a short description of the areas of synergy listed in figure 6.1. These will include some

reference to my work in this dissertation but is not limited by them.

6.8.1 Using artificial neural networks to advance neuroscience

The following list is not exhaustive and instead include my personal biases of where I foresee machine learning assisting our understanding of the brain. I will begin with the topics that were touched upon in projects covered in this dissertation. Namely, those belong to the areas where we used CNNs as models of the visual stream. We were able to use a neural network model to capture the dynamics of the MEG signal in part 3. Different CNN models had different components such as skip connections, inception layers, and recurrence and like previous work (Kietzmann et al., 2019) we found that recurrence was especially useful at capturing variance later in time epochs. While we did not use the CNN models to discover new computations, it is quite possible to do so and an avenue for future work. Other groups have been able to modify neural networks in order to synthesize stimuli that maximally activate different regions in the brain (Bashivan, Kar, & DiCarlo, 2019; Ponce et al., 2019). This process in effect uses CNN architectures to reveal what different brain areas might be processing. In addition, by selectively adding and removing connections, and assessing CNN correspondence to brain activity, future work can provide additional key insights into how different algorithms are implemented across brain areas (Kar & DiCarlo, 2020; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019). Lastly, exploring how AI and brain activity can directly interface are appealing avenues of continued symbiosis. The efforts in brain machine interface have captured the imagination of many since Miguel Nicoleilis' lab (Carmena et al., 2003) was able to use a macaque's brain signals in order to control a robotic arm. In more recent times, Elon Musk's Neuralink implants (Musk, 2019) have pushed the boundaries in engineering with recording devices being housed in self contained bluetooth Neuralink implants. As recording technology continues to advance and allow for the simultaneous recording from thousands of channels, the importance of using machine learning and AI to improve our understanding of the brain will only increase as we attempt

makes sense of the added dimensionality in our data sets.

6.8.2 Using neuroscience to advance artificial networks

The fingerprints of neuroscience are evident in many facets of current CNNs. The most obvious being the initial idea behind using convolutional layers in a hierarchical fashion. However, many will argue that CNNs have notably now diverged from biology, rely heavily on non-biological backpropogation (Marblestone, Wayne, & Kording, 2016) and as a result there is not much that AI can learn from the brain. I however am of the opinion that by continuing to explore how the brain functions, we may stumble upon several valuable lessons that will create more efficient, flexible, and ultimately better performing artificial networks across a number of tasks. Even with backpropagation, we are finding biological evidence for mechanisms for the ultimate goal of backpropogation which is credit assignment-the process of assigning weights for optimization throughout a network. Evidence of credit assignment has been found in processes occurring between apical and basilar dendrite (Richards et al., 2019). By investigating these processes further, the brain may still yet offer learning algorithms that supplement or replace backpropogation. Furthermore, while hierarchical organization was an initial contribution from the brain to CNN, there are continued ways in which the brain might influence artificial network architectures as is evident in current efforts with capsule networks (Sabour et al., 2017) which I have discussed previously. When looking at learning in general, AI can further take lessons from how the brain performs meta learning, the human ability to learn how to learn something. Even currently, years of research of the reward system within the brain has formed the basis of reinforcement learning in AI research (Vu et al., 2018). In looking at the training sets we use for AI systems, we can also use developmental neuroscience research. The applications of such research are evident in recent efforts in modifying neural network training to include training images that are more representative of how humans learn object categories through development. These studies have shown that using training images that simulate

human development are successful in improving a networks' ability to generalize between tasks (Bambach, Crandall, Smith, & Yu, 2018; Kiar, Zeman, & Op de Beeck, 2021). Thus, by manipulating the training data, these studies represent promising inroads for improving the ability of networks to learn with fewer examples (i.e. one-shot learning), making for more efficient artificial networks. Lastly, the brain provides a wonderful constraint in terms of the limited energy that can be used to power neural computations. Thus, we can study the brain as a means of finding algorithms that will limit energetics and computational costs required to train CNNs as AI research continue to look to expand (Thompson, Greenewald, Lee, & Manso, 2020).

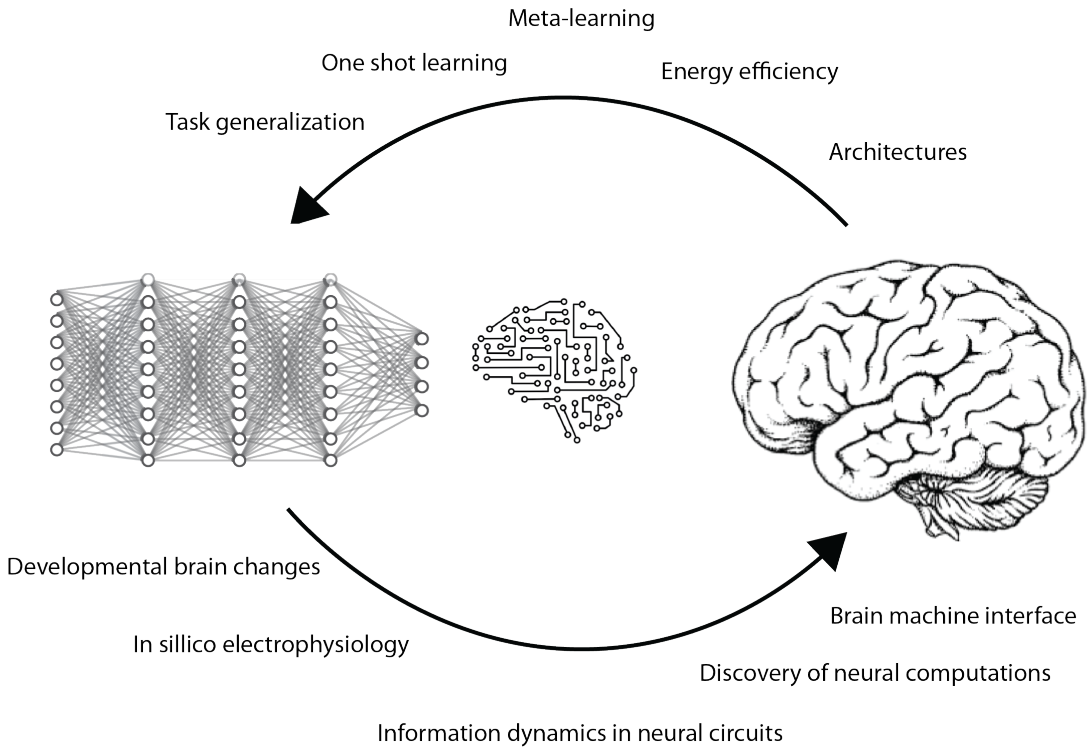

Figure 6.1: Opportunities to improve AI and neuroscience

6.9 References

Bambach, S., Crandall, D. J., Smith, L. B., and Yu, C. (2018). Toddler-inspired visual object learning. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):1201–1210.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15(4):600–609.

Bar, M. (2021). From objects to unified minds. *Current Directions in Psychological Science*.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439).

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., DeWeerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401.

Belitski, A., Gretton, A., Magri, C., Murayama, Y., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2008). Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *Journal of Neuroscience*, 28(22):5696–5709.

Blake, R. and Cormack, R. H. (1979). On utrocular discrimination. *Perception Psychophysics*, 26(1):53–68.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., and Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, 26(1):132–142.

Carlson, T. A., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13:1–19.

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., Patil, P. G., Henriquez, C. S., and Nicolelis, M. A. (2003). Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biology*, 1(2):193–208.

Contini, E. W., Wardle, S. G., and Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., and Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology*, 16(7):1–19.

Dougherty, K., Cox, M. A., Westerberg, J. A., and Maier, A. (2019). Binocular modulation of monocular v1 neurons. *Current Biology*, 29(3):381–391.e4.

Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, 23(1):115–125.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(24):429–433.

Goddard, E., Carlson, T. A., Dermody, N., and Woolgar, A. (2016). Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage*, 128:385–397.

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology*, 13:159–166.

Grootswagers, T., Cichy, R. M., and Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, 179(June):252–262.

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., and Carlson, T. A. (2017). Asymmetric compression of representational space for object animacy categorization under degraded viewing conditions. *Journal of cognitive neuroscience*, 29(12):1995–2010.

Grootswagers, T., Robinson, A. K., Shatek, S. M., and Carlson, T. A. (2019). Untangling featural and conceptual object representations. *NeuroImage*, 202(July):116083.

Harrison, S. A. and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635.

Haxby, V. J., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–30.

Hubel, D. H. and Wiesel, T. N. (1977). Ferrier lecture. functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London - Biological Sciences*, 190(1130):1–59.

Jackson, R. E. and Calvillo, D. P. (2013). Evolutionary psychology. *Evolutionary Psychology*, 11(5):1011–1026.

Jehee, J. F. M., Brady, D. K., and Tong, F. (2011). Attention improves encoding of task-relevant features in the human visual cortex. *Journal of Neuroscience*, 31(22):8210–8219.

Judaš, M., Cepanec, M., and Sedmak, G. (2012). Brodmann's map of the human cerebral cortex - or brodmann's maps? *Translational Neuroscience*, 3(1):67–74.

Kajikawa, Y. and Schroeder, C. E. (2011). How local is the local field potential? *Neuron*, 72(5):847–858.

Kajikawa, Y., Smiley, J. F., and Schroeder, C. E. (2017). Primary generators of visually evoked field potentials recorded in the macaque auditory cortex. *Journal of Neuroscience*, 37(42):10139–10153.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(11):4302–11.

Kar, K. and DiCarlo, J. J. (2020). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, pages 1–13.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983.

Katzner, S., Nauhaus, I., Benucci, A., Bonin, V., Ringach, D. L., and Carandini, M. (2009). Local origin of field potentials in visual cortex. *Neuron*, 61(1):35–41.

Kauffmann, L., Ramanoël, S., Guyader, N., Chauvin, A., and Peyrin, C. (2015). Spatial frequency processing in scene-selective cortical regions. *NeuroImage*, 112:86–95.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-haignere, V. S., Mcdermott, J. H., Kell, A. J. E., Yamins, D. L. K., Shook, E. N., and Norman-haignere, V. S. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):1–15.

Kiar, L., Zeman, A., and Op de Beeck, H. (2021). Training for object recognition with increasing spatial frequency : A comparison of deep learning with human vision . *bioRxiv*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, 116(43):21854–21863.

Kok, P., Jehee, J. F. M., and de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2):265–270.

Kreiman, G., Hung, C. P., Kraskov, A., Quiroga, R. Q., Poggio, T., and DiCarlo, J. J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron*, 49(3):433–445.

Kriegeskorte, N., Mur, M., Ruff, D., and Kiani, R. (2008). Matching categorical object representations in inferior temporal cortex of mand and monkey. *Neuron*, 60(6):1126–1141.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9).

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., and Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, 166(3-4):289–297.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leski, S., Lindén, H., Tetzlaff, T., Pettersen, K. H., and Einevoll, G. T. (2013). Frequency dependence of signal power and spatial reach of the local field potential. *PLoS Computational Biology*, 9(7).

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A. Y., Tal, I., Haegens, S., Melloni, L., Knight, R. T., and Schroeder, C. E. (2020). Dissociation of broadband high-

frequency activity and neuronal firing in the neocortex. *Science advances*, (August):1–13.

Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., and Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features.

Lindén, H., Tetzlaff, T., Potjans, T. C., Pettersen, K. H., Grün, S., Diesmann, M., and Einevoll, G. T. (2011). Modeling the spatial reach of the lfp. *Neuron*, 72(5):859–872.

Long, B., Yu, C. P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, 115(38):E9015–E9024.

Lu, Y., Yin, J., Chen, Z., Gong, H., Liu, Y., Qian, L., Li, X., Liu, R., Andolina, I. M., and Wang, W. (2018). Revealing detail along the visual hierarchy: Neural clustering preserves acuity from v1 to v4. *Neuron*, 98(2):417–428.e3.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Marblestone, A., Wayne, G., and Kording, K. (2016). Towards an integration of deep learning and neuroscience. 10(September):1–41.

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 7(10):3215–3229.

Millet, J. and King, J.-R. (2021). Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech.

Musk, E. (2019). An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fmri. *NeuroImage*, 56(2):400–410.

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598–16603.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009.e10.

Quiroga, R. Q., Kreiman, G., Koch, C., and Fried, I. (2008). Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, 12(3):87–91.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Ritchie, J. B., Tovar, D. A., and Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology*, 11(6).

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nature Neuroscience*, 20(1):107–114.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):3857–3867.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. (2020). The computational limits of deep learning. *arXiv*.

Tong, F. and Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, 63:483–509.

Tovar, D., Murray, M., and Wallace, M. (2020a). Selective enhancement of object representations through multisensory integration. *Journal of Neuroscience*, 40(29):5604–5615.

Tovar, D. A., Westerberg, J. A., Cox, M. A., Dougherty, K., Carlson, T. A., Wallace, M. T., and Maier, A. (2020b). Stimulus feature-specific information flow along the columnar cortical microcircuit revealed by multivariate laminar spiking analysis. *Frontiers in Systems Neuroscience*, 14(November):1–14.

Van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M. A., Poort, J., Van Der Togt, C., and Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(40):14332–14341.

Vogler, J. N. and Titchener, K. (2011). Cross-modal conflicts in object recognition: Determining the influence of object category. *Experimental Brain Research*, 214(4):597–605.

Vu, M.-A. T., Adali, T., Ba, D., Buzsaki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V., Widge, A. S., Mayberg, H. S., Sapiro, G., and Dzirasa, K. (2018). A shared vision for machine learning in neuroscience. *The Journal of Neuroscience*, 38(7):0508–17.

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology*, 80(2):1006–1010.

Wardle, S. G. and Baker, C. (2020). Recent advances in understanding object recognition in the human brain: Deep neural networks, temporal dynamics, and context. *F1000Research*, 9:1–14.

Westerberg, J. A., Cox, M. A., Dougherty, K., and Maier, A. (2019). V1 microcircuit dynamics: altered signal propagation suggests intracortical origins for adaptation in response to visual repetition. *Journal of neurophysiology*, 121(5):1938–1952.

Xing, D., Yeh, C. I., Burns, S., and Shapley, R. M. (2012). Laminar analysis of visually evoked activity in the primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 109(34):13871–13876.

Xu, Y. and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, 12(2065):1–16.

Appendix A

Appendix Ch A: The Neural Computations for Stimulus Presence and Modal Identity Diverge Along a Shared Circuit

The contents of this chapter are adapted from

Tovar, D. A., Noel, J.-P., Ishizawa, Y., Patel, S. R., Eskandar, E. N., Wallace, M. T. (2020). The neural computations for stimulus presence and modal identity diverge along a shared circuit. BioRxiv, 2020.01.09.900563. https://doi.org/10.1101/2020.01.09.900563

1.1 Abstract

The brain is comprised of neural circuits that are able to flexibly represent the complexity of the external world. In accomplishing this feat, one of the first attributes the brain must code for is whether a stimulus is present and subsequently what sensory information that stimulus contains. One of the core characteristics of that information is which sensory modality(ies) are being represented. How information regarding both the presence and modal identity of a given stimulus is represented and transformed within the brain remains poorly understood. In this study, we investigated how the brain represents the presence and modal identity of a given stimulus while tactile, audio, and audio-tactile stimuli were passively presented to non-human primates. We recorded spiking activity from primary somatosensory (S1) and ventral pre-motor (PMv) cortices, two areas known to be instrumental in transforming sensory information into motor commands for action. Using multivariate analyses to decode stimulus presence and identity, we found that information regarding stimulus presence and modal identity were found in both S1 and PMv and extended beyond the duration of significant evoked spiking activity, and that this information followed different time-courses in these two areas. Further, we combined time-generalization decoding with cross-area decoding to demonstrate that while signaling the presence of a stimulus involves a feedforward-feedback coupling between S1-PMv, the processing of modal iden-

tity is largely restricted to S1. Together, these results highlight the differing spatiotemporal dynamics of information flow regarding stimulus presence and modal identity in two nodes of an important cortical sensorimotor circuit.

1.2 Significance Statement

It is unclear how the structure and function of the brain support differing sensory functions, such as detecting the presence of a stimulus in the environment vs. identifying it. Here, we used multivariate decoding methods on monkey neuronal data to track how information regarding stimulus presence and modal identity flow within a sensorimotor circuit. Results demonstrate that while neural patterns in both primary somatosensory (S1) and ventral pre-motor (PMv) cortices can be used to detect and discriminate between stimuli, they follow different time-courses. Importantly, findings suggest that while information regarding the presence of a stimulus flows reciprocally between S1 and PMv, information regarding stimulus identity is largely contained in S1.

1.3 Introduction

Single-unit neurophysiological recordings demonstrate that neural activity within the primary somatosensory area (S1) is monotonically related to stimulus amplitude (Mountcastle et al., 1969). This suggests that a rate code is used to signal the probability of a somatosensory stimulus being present in the environment (Ahissar et al., 2000). Beyond this first cortical area, however, neurons show a variety of response patterns to different stimulus features. For example, some neurons show increasing spiking activity with increasing stimulus frequency, whereas others show the opposite relationship (Salinas et al., 2000). Furthermore, non-linear computations may effectively help filter which information is propagated forward in the cortical hierarchy to solve discrimination problems (Romo de Lafuente, 2012). Thus, the computational principles that appear best suited for stimulus detection are unlikely to be those best suited for stimulus discrimination. It is currently unclear how brain circuits support these various aspects of processing a sensory stimulus,

and how the same brain regions differ in this regard.

Arguably, understanding the mechanistic bases of how the brain signals the presence and the identity of a stimulus has been challenging partly due to the widespread use of univariate techniques and the heavy focus on characterizing the responses of single neurons. However, it is increasingly common to record multiple neurons concurrently across areas, and using multivariate frameworks, uncover neural codes (i.e., response patterns) that are present at the population level (Jonas Kording, 2017). In addition to understanding the basic characteristics of neural activity of specific neurons and within specific areas, multivariate analyses are able to further probe the manner by which distinct modules communicate with one another, and thus how information is propagated and transformed within the brain (Kumar et al., 2010l Stringer et al., 2019). With large-scale simultaneous multi-area recordings becoming commonplace (Jun et al., 2017; Steinmetz et al., 2018), these analyses are becoming increasingly important tools (Buzsaki, 2004; Stevenson Kording, 2011).

In the current study, we sought to track information flow relating to the presence and modal identity of a stimulus by examining global neural patterns using multivariate pattern analysis. We simultaneously recorded neuronal activity from two intermediate stages along the hierarchy from sensory input to motor output – primary somatosensory (S1) and ventral pre-motor (PMv) cortex. These areas are two key nodes in a well-established circuit for tactile detection and discrimination (Romo et al., 2004; de Lafuente Romo, 2005, 2006). In addition to its role in somatosensory function, the PMv cortex is known to be important in auditory discrimination (Lemus et al., 2009) and also possesses multisensory audio-tactile neurons (Graziano et al., 1997). Hence, recording simultaneously from these two areas provides the opportunity to not only examine how information flows between S1 and PMv to support tactile stimulus detection, but also to examine information encoding and flow in the context of determining stimulus modal identity (i.e., auditory, tactile, audio-tactile).

To address this question, tactile, auditory, and audio-tactile stimulation was passively

delivered to rhesus monkeys, and neural signals related to the presence and/or modal identity of the stimulus were decoded using multivariate methods (Edelman et al., 1998; Haxby et al., 2001; Kriegeskorte Kievit, 2013; Goddard et al., 2017). In addition to training and testing within neural areas and at similar time-points, we dissociate these time-periods (time-generalization technique; King Dehaene, 2014), as well as train and test neural decoders across brain regions. The novel joint application of the time-generalization technique and cross-area decoding allows the tracking of information transfer between S1 and PMv. This combination highlights strikingly different spatiotemporal dynamics in the transfer of information related to the presence vs. modal identity of the stimulus.

1.4 Methods

1.4.1 Animal Model

Two adult male monkeys (Macaca mulatta, 10 –12 kg; Monkey E and Monkey H) were used. Animals were handled according to the institutional standards of the National Institutes of Health (NIH) and protocols were approved by the institutional animal care and use committee at Massachusetts General Hospital.

1.4.2 Surgical Procedures

A titanium head post and a vascular access port in the internal jugular vein (Model CP6; Access Technologies) were surgically implanted on each of the two animals. Once the animals learned the behavioral task (see below), a craniotomy was performed and extracellular microelectrode arrays (Floating Microelectrode Arrays; MicroProbes) were implanted into S1 and PMv by following landmarks on the cortical surface and stereotaxic coordinates (Fig 1A). Each array (1.95x2.50 mm) contained 16 platinum–iridium recording microelectrodes (0.5 M, 1.5– 4.5 mm staggered length) separated by 400 μm. Monkey E had two arrays in S1 and another two in PMv (total of 32 electrodes in each area, all in the left hemisphere). Implantation for Monkey H was identical to that of Monkey E, with the exception that all electrodes were implanted in the right hemisphere. The recording experiments were

performed after 2 weeks of recovery following the array surgery. All experiments were conducted in a radio frequency-shielded recording enclosure.

1.4.3 Materials and Apparatus

Three different types of sensory stimulation were given: audio-alone, tactile-alone, and a combined audio-tactile multisensory conditions. The tactile stimuli were air puffs of 250ms duration delivered at 12 psi to the lower part of the face contralateral to the recording hemisphere. This tactile stimulus was delivered via a computer-controlled regulator with a solenoid valve (AirStim; San Diego Instruments). The eye area was avoided from the puff stimulation. Auditory stimuli were pure tones (4000 Hz at 80 dB SPL) lasting 250ms. These tones were generated by a computer and delivered using two speakers 40 cm from the animal. White noise (50 dB SPL) was applied throughout the trial to mask the air puff and mechanical noises. Audio-tactile stimulation was the synchronous administration of the auditory and tactile stimuli described above. All of the stimulus sets were presented randomly to the animal throughout the recording session.

1.4.4 Experimental Procedure

After a start tone (1000 Hz, 100 ms), the animals were required to initiate each trial by holding the button located in front of the primate chair using the hand ipsilateral to the recording hemisphere. Animals were required to hold the button until the end of a trial, which was indicated by a liquid reward 3 seconds after stimuli onset (Fig 1B). The monkeys were trained to perform a correct response in >90% of the trials consistently for longer than 1.5 h. One of the three sensory stimulus sets (audio, tactile, or audio-tactile), or a catch trial with no sensory stimulation, was delivered to the animal during the trial at a random delay. Each condition was equally likely to be presented.

1.4.5 Single-Unit Activity, Recording and Preprocessing

Neural activity was recorded continuously and simultaneously from S1 and PMv. Analog data was amplified, band-pass filtered between 0.5 and 8 kHz, and sampled at 40 kHz (OmniPlex; Plexon). Spiking activity was obtained by high-pass filtering at 300kHz and applying a minimum threshold of 3 standard deviations in order to exclude background noise from the raw voltage traces on each channel. Subsequently, action potentials were sorted using waveform principal component analysis (Offline Sorter; Plexon) and binned into 1 ms bins, effectively rendering the sampling rate of 1 kHz. Spike time-stamps were convolved with a 100ms long box-car window and moved in 1 ms steps (Fig 1C). Time-courses were then baseline-corrected by subtracting their pre-stimulus activity (-200 ms to 0 ms post-stimulus onset). This dataset has been previously reported in Ishizawa et al., 2016, and Noel et al., 2019.

1.4.6 Multivariate Pattern Analysis

Our aim here was to track message passing and information transformation within the cortex and hence focus on multivariate decoding techniques. Following data preprocessing, we used CoSMoMVPA (Oosterhof, et al., 2016) to decode stimulus presence vs. absence, as well as the sensory modality of the stimuli presented. Linear discriminant analysis (LDA; Duda et al., 2001) classifiers were trained and tested in 1ms increments using 4-fold cross validation. In this procedure, trials are randomly assigned to one of four subsets. Three of the four subsets (75% of the data) are pooled together to train the classifier and then decoding accuracy is tested on the remaining subset (25% of the data). This procedure is repeated a total of four times, such that each of the subsets is tested once. Decoding results are reported in percent correct of classifications at each time point in the time series ranging from -100ms to 1000ms relative to stimulus onset. This analysis was conducted independently for each recording session (n=18), distinction of interest (stimulus presence and stimulus modality), as well as within and across brain areas (S1 and PMv). Mean and

standard error were then calculated across recording sessions at each time point (Fig 1D).

Regarding statistical analyses, each time point was tested for the null and alternative hypotheses using Bayes' factors. The null hypothesis indicates that there is no information regarding the presence or absence of stimuli for stimulus detection and no information regarding the type of modality for modality discrimination. Thus, the null hypothesis would be the decoder guessing at chance, which would be 50.0% decoding accuracy for stimulus presence, and 33.3% decoding accuracy for modality discrimination. We then calculated the probability of the alternative hypothesis in relation to the null hypothesis. A Bayes' factor greater than 3 indicates substantial evidence for the alternative hypothesis, anything between 3 and 1/3 indicates insufficient evidence, and values less than 1/3 indicate evidence for the null hypothesis (Jeffreys, 1961; Wetzels et al., 2010). Substantial evidence for the alternative hypothesis indicates that the brain state contains meaningful information that the classifier can utilize to identify the correct trial condition for stimulus presence (stimulus present or absent) or stimulus identity (audio, tactile, or audio-tactile). Furthermore, Bayes' factors provide an added advantage over Frequentist inference: in addition to rejecting the null hypothesis, this framework can also provide support for either the null hypothesis or to determine that the data is insensitive. For both stimulus presence and modality discrimination, trials were balanced across conditions, as imbalance among class types can have the unwanted effect of biasing the classifier toward the class with more trials (Grootswagers et al, 2017).

1.4.7 Time Generalization Within Areas and Across Areas

To probe the dynamics of the available information used by the classifier to decode presence or absence of stimuli regardless of modality, as well as the modality of the stimuli presented, we used a time generalization decoding technique (Carlson et al., 2011; King Dehaene, 2014; Fig 1D). In this analysis, the classifier was trained on the same decoding distinctions as before (presence and identity of sensory stimuli). However, to investigate

how well neural data from one timepoint generalizes to all others, the classifier is trained on a particular timepoint within the time series (i.e., -100 to 1000 ms post-stimuli onset) and then tested with data from every timepoint in the time series. This procedure was repeated for every timepoint and concatenated to create 1100 x 1100 matrix containing every possible combination of training and testing timepoints. The diagonal along the matrix represents times in which training and testing were performed within the same timepoint. Lastly, we performed a similar time generalization analysis across areas in order to investigate how well different timepoints in one area can decode information at different timepoints in the other area – putatively indicating the flow of information from one area at one time-point, to another area at a different time-point. We trained across all timepoints in S1 and tested on PMv and then performed training on PMv and tested on S1. Since PMv and S1 had an unequal number of single units captured (S1 = 9.6 +/- 4.5, PMv = 5.7 +/- 2.3) we randomly subsampled from the area with more single units isolated. To eliminate potential sampling bias, we performed the cross-area time generalization analyses ten times, with different randomly subsample single units. The mean decoding results across the ten iterations was then computed for each recording session.

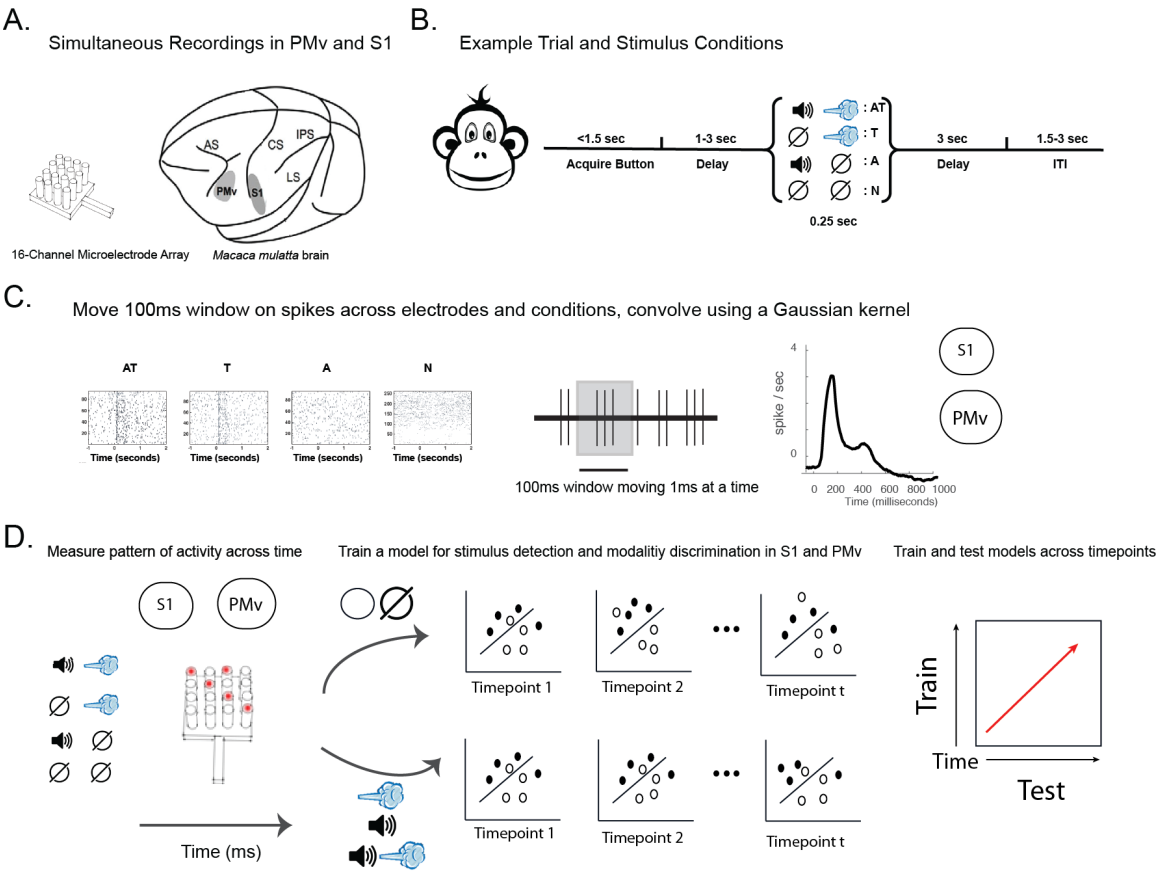

Figure 1.1: Experiment Schematic. (A) Neural recordings were effectuated via 16-electrodes platinum–iridium arrays implanted in S1 and PMv. (B) Animals were trained to initiate trials via button press, which following a delay would evoke one of three sensory stimulus sets (audio, tactile, or audio-tactile), or a catch trial with no sensory stimulation delivered. (C) Raster plots of an example session in S1 and the average S1 response to tactile stimulation after convolving spike trains with a box-car 100ms in length and moving in 1ms steps. (D) Multivariate classifiers (Linear Discriminant Analysis, LDA) were trained on each time-point to differentiate either between the absence and presence of sensory stimuli (regardless of the nature of the stimuli; detection), or to discriminate between sensory modalities (audio, tactile, or audio-tactile; discrimination).

1.5 Results

1.5.1 Multivariate Decoding Allows Tracking Stimuli Presence and Identity Over Long Periods

Both S1 and PMv showed evoked responses during the presentation of sensory stimuli. We used Bayes factors at each timepoint to assess whether the evoked responses diverged significantly from baseline activity. For the univariate analysis, when averaging responses over modalities, S1 showed a strong evoked response, showing substantial evidence for the alternative hypothesis (defined as Bayes factor [BF] >3) at two-time periods, from 19-184 ms and from 309-388 ms post-stimulus onset. PMv showed a later response, from 141-422 ms post-stimulus onset. When looking at evoked responses to specific modalities, S1 responds to tactile stimulation for the period between 36-184 ms post stimulus onset, to auditory stimuli from 21-102 ms post-stimulus onset, and to audiotactile stimulation from 19-197 ms and 328-414 ms post-stimulus onset. Responses of PMv to sensory stimuli are not as robust, but there is clear evidence for evoked responses to tactile stimuli from 131-157 ms and from 212m-448 ms post-stimulus onset, to auditory stimuli from 182-405 ms post-stimulus onset, and to combined audiotactile stimuli from 166-381 ms post-stimulus onset.

Using time-resolved LDA, we were able to decode the presence (vs. absence) of stimuli in S1 and PMv (Fig 2C). Onset decoding latencies, defined as the first timepoint of at least 20 ms of sustained significant decoding above chance (see Carlson et al., 2013) were found for S1 beginning 36 ms post-stimulus onset (Fig 2C, purple) and for PMv beginning at 58ms post-stimulus onset (Fig 2C, green). Maximum decoding performance was reached at 183 ms post-stimulus onset for S1 and at 222 ms post-stimulus onset for PMv. For both S1 and PMv, decoding remains significantly above chance for periods extending beyond 1000 ms post stimulus onset. This observation highlights the utility of indexing not only the activity of single neurons via traditional univariate approaches, but also in examining the responses of neuronal populations via multivariate decoding. For example, the average firing rate

produces a strong transient response to tactile stimulation followed by a sustained response in S1, which return to baseline within approximately 500ms. In contrast, it was possible to decode the presence of a stimulus in S1 for a period at least twice as long (1000ms) using multivariate approaches.

We next used Bayes' factors to look at the time-resolved differences in the decoding of stimulus presence between S1 and PMv. Results demonstrate significant evidence supporting the alternative hypothesis (BF>3), suggesting a differential time-course during which stimulus presence information is available in S1 and PMv (Fig 2C, black curve). Beginning at 40 ms and extending up until 186 ms, decoding was better in S1 than PMv, consistent with the earlier decoding onset found in the primary sensory area. Following 186 ms, evidence is stronger for the null hypothesis (BF<1/3) up until 651 ms post-stimulus onset. Following 651 ms, evidence for the alternative hypothesis is once again supported, but this time in PMv. These findings suggest that information regarding stimulus presence may be transferred between S1 (first) and PMv (later).

We then applied the same approach to determining when the modality (i.e., A, T, AT) of the stimulus could be decoded from the neural signals in S1 and PMv. Results suggested above chance decoding (i.e., >33.3%) starting 37 ms post-stimulus onset for S1 (Fig 2D, purple), and starting 70 ms post-stimulus onset for PMv (Fig. 2D, green). A maximum modality decoding performance of 50.0% was reached at 125 ms post-stimulus onset for S1 and a maximum modality decoding performance of 41.0% was reached at 213 ms post-stimulus onset for PMv. As shown by the difference in decoding performance within S1 and PMv (Fig 2D, black curve), decoding accuracy was significantly higher for S1 relative to PMv for two sustained periods - between 40-74ms post-stimulus onset, as well as between 348-470ms post-stimulus onset. Collectively, these results suggest that the computations underlying the detection of a stimulus and the identification of stimulus modality evolve over differing temporal epochs in S1 and PMv. More specifically, while information regarding detection appears later in PMv as compared to S1, and thus leading to a single

time-period where stimulus presence is more readily decoded in S1 than PMv, information regarding stimulus modality is more readily decoded in S1 over PMv over both an early and late temporal epoch.

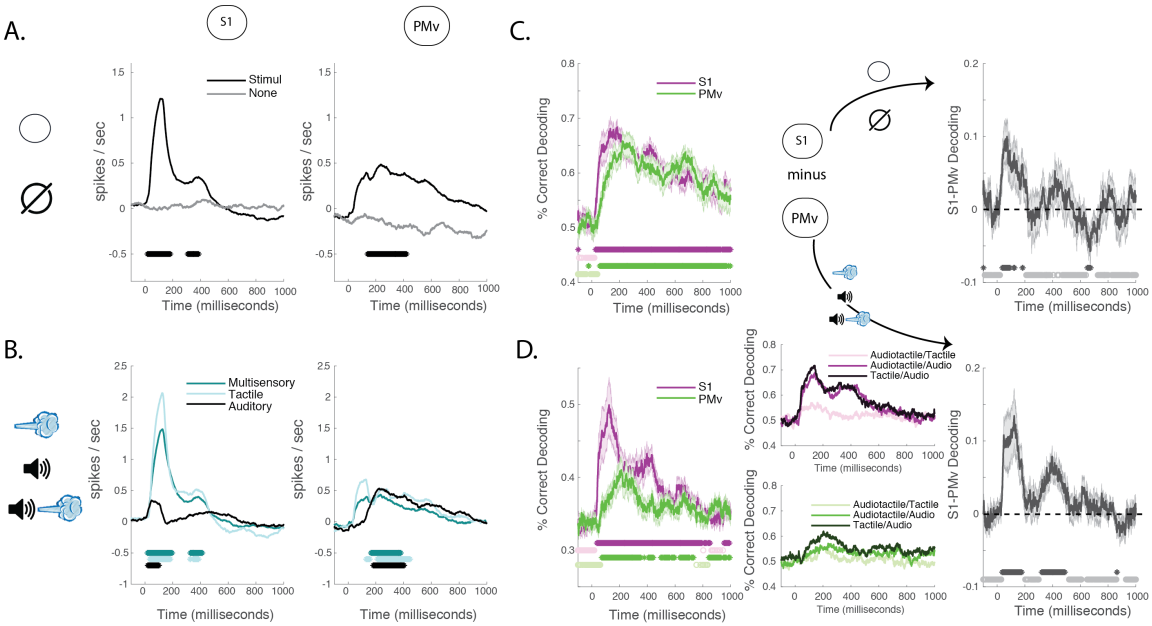

Figure 1.2: Univariate and Multivariate Responses to Sensory Stimulation. (A) Both S1 and PMv show evoked responses during the presentation of sensory stimuli. (B) S1 responds to tactile stimulation, while also responding to audio-tactile stimulation, but less to auditory stimuli alone. PMv does not show as clear evoked responses to sensory stimuli as primary somatosensory area does but shows less disparity in evoked responses across stimuli types. (C) LDA classified above chance either the presence or absence of sensory stimulation starting 36ms and 58ms for S1 and PMv respectively post-stimuli onset, and lasting 1s, well beyond the time-period where univariate responses are apparent. As illustrated by the difference in correct decoding between S1 and PMv, information regarding stimulus detection was present first in S1, then was present in both S1 and PMv, and finally was stronger in PMv than S1. (D) Discrimination of sensory modalities was also correctly decoded by LDA, with modal identity being clearer in S1 than PMv, particularly at early latency post-stimulus onset, and between approximately 200 and 400ms post-stimulus onset. Asterisks indicate significant decoding above chance, using Bayes' factors (Bayes' Threshold >3).

1.5.2 Information Regarding the Presence and Modal identity of Stimuli Follow Different Dynamics in S1 and PMv

To further explore how information dynamics regarding the encoding of the presence or modal identity of a stimulus varies across S1 and PMv, we used a time generalization approach (Carlson et al., 2011; King Dehaene, 2014) where a classifier is trained at one timepoint and then tested across the remaining timepoints. Specifically, it probes how information at a given timepoint generalizes to information throughout the time series to understand whether the information is increasing, decreasing, or re-emerging at later times. In the present study we leveraged the fact that decoding performance is better when training on a low signal-to-noise ratio (SNR) and testing on a high SNR (Fig 3A, van den Hurk Op de Beeck, 2019) to quantify the degree to which information at a particular time is changing (Fig 3B). Given that training timepoints are plotted along the y-dimension and testing times are plotted along the x-dimension in a time generalization matrix, if information at a particular timepoint increases during the time-course, this will appear as an off-diagonal shift in the horizontal (rightward) direction. Conversely if information at a timepoint decreases during the time course, this will appear as an off-diagonal shift in the vertical (upward) direction. To calculate the overall direction of information change (horizontal or vertical off-diagonal) we subtracted the vertical off-diagonal from the horizontal off-diagonal.

Regarding the decoding of stimulus presence, we calculated all of the times where there was very strong evidence (BF>30) for the alternative hypothesis to capture the most prominent information states (Fig 3C). The plots showed that decoding onset and off-diagonal spread across training-testing time periods varied between S1 and PMv for information regarding stimulus presence (see Fig 3C, first and third columns). This difference becomes apparent in the horizontal-vertical off diagonal difference histograms (see Fig 3C, second and fourth columns). In S1, the information states strengthen from 39-207ms and then again from 239-555 ms, after which information weakens until 638 ms and then ends with a final wave of strengthening. On the other hand, PMv shows the opposite pattern, with

information initially weakening beginning at 93 ms, then briefly strengthening at 261 ms, weakening again at 351 ms, and showing a final wave of strengthening beginning at 497 ms. Overall, these results highlight a general pattern in which these are opposing temporal dynamics to information strengthening and weakening for S1 and PMv (see Supplemental Fig 1), potentially implying that information is flowing back and forth between these two areas. Furthermore, whereas information regarding stimulus presence demonstrated a greater strengthening pattern in the initial response epoch for S1, PMv demonstrated a greater strengthening pattern in the late time period (after 500ms).

In contrast to the patterns seen for information regarding stimulus presence, the dynamics for information about modal identity in S1 and PMv show consistent strengthening and weakening, respectively. In S1, information regarding modal identity begins to increase at 65 ms and continues to increases until 387 ms. In contrast, in PMv information regarding modal identity shows a weakening pattern beginning at 90 ms and continues to weaken until 299 ms. In sum, the difference in dynamics regarding information pertaining to stimulus presence and modal identity strongly suggest differences in how this information is processed and shared between S1 and PMv. An additional finding that is illustrated by the on-diagonal analyses is that overall information regarding modal identity is short-lived as compared to information regarding stimulus presence.

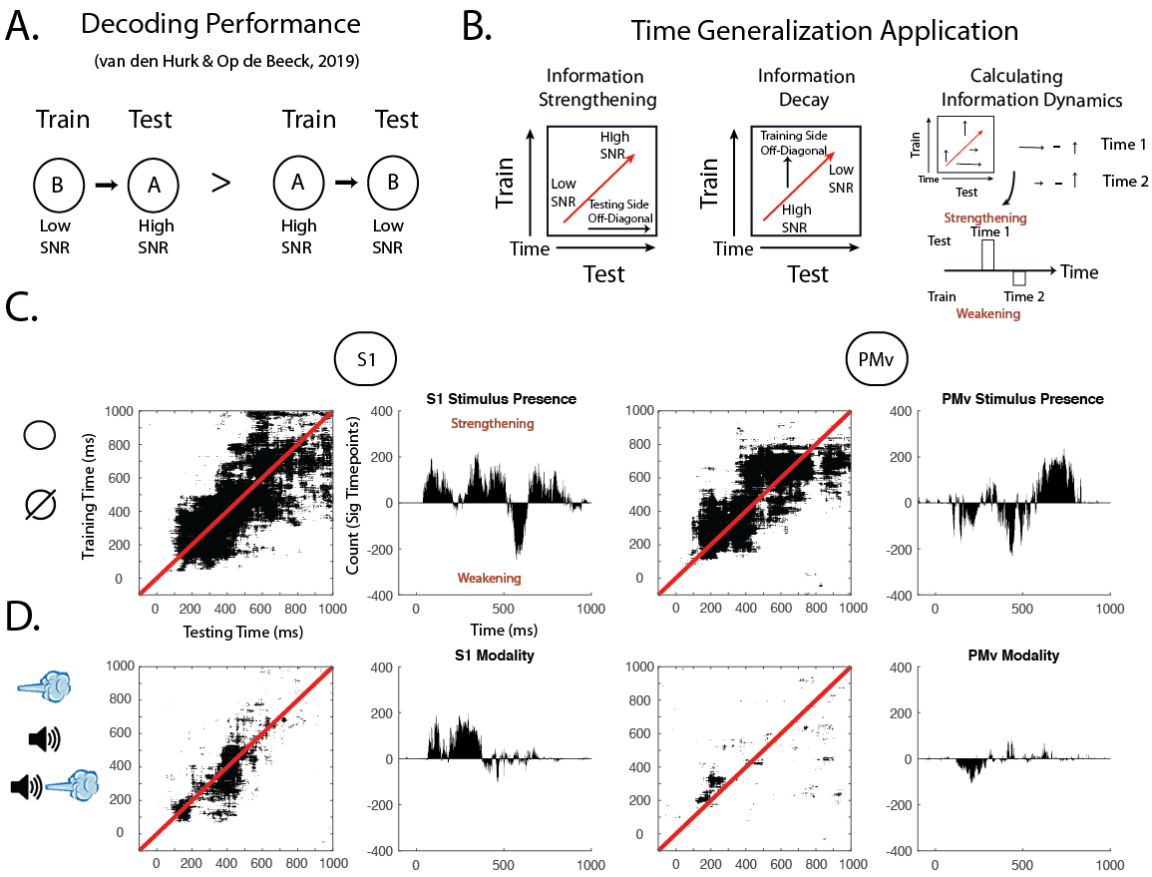

Figure 1.3: Time generalization results. (A) Decoding accuracy is better when training on low signal-to-noise (SNR) and testing on higher SNR, than when training on high SNR and testing on low SNR. (B) In conjunction with the time-generalization technique, this observation can be leveraged to estimate whether a particular information state is strengthening (i.e., becoming more discriminant with time) or weakening within a cortical area. (C) Time generalization plots for decoding stimulus presence in S1 (left) and PMv (right). Second and fourth columns show the difference in significant time-points over which decoding generalizes along the training and testing axis – positive counts indicate information strengthening, while negative counts indicate information weakening. (D) Follows the same convention as (C), for modal identity as opposed to stimulus presence.

1.5.3 Cross-Region Decoding Reveals Feedforward Presence Information and Feedback Identity Information

To more directly track shared information between S1 and PMv, we trained classifiers on neural data collected from one region and tested on another, while also performing time-generalization (Carlson et al., 2011; King Dehaene, 2014). For example, and as illustrated in Fig 4A, we can use this analysis to train on S1 and test on PMv to examine for potential significant horizontal off-diagonals (i.e., in the future along the testing dimension). Such a result would suggest that S1 shares common information that is present at a later time in PMv (see Fig 4A for other examples).

In decoding stimulus presence we found that training and testing across S1 and PMv along the same time points did not yield any periods of time with sustained significant decoding accuracy using a criterion of substantial evidence (BF>3) (Fig 4B on-diagonal). Such a result suggests that S1 and PMv do not contain common information regarding the presence of a stimulus during the same time period, although both do contain information regarding stimulus presence (Fig 2C). The fact that the within-area decoding is successful, but across area decoding is not, suggests that the codes for stimulus presence within S1 and PMv are likely of different format.

To better explore whether the lack of simultaneous shared information was due to a transformation of information from one area to another, we inspected the off-diagonals in the time generalization matrices. As shown in Fig 4B off-diagonal, results indicate that beginning at approximately 38ms and extending to 100ms post-stimulus onset, information regarding stimulus presence in S1 significantly generalizes to PMv for the time period spanning between 100-442 ms post-stimulus onset This result shows that information pertaining to stimulus presence in PMv at this later interval is similar to that seen earlier (38ms to 100ms) in S1. Training on PMv and attempting to decode within S1 across different time periods also yielded significant vertical off-diagonals (i.e., along the training dimension), beginning at 15 ms post-stimulus onset and extending forward to 97-287 ms. (Fig 4B).

Thus, whether training in S1 or PMv, information regarding stimulus presence generalized in the direction from S1 to PMv. The fact that training in PMv and decoding in S1 yielded a more restricted time-period of generalization than within area decoding in S1 and PMv (Fig 2C) could imply that while information regarding stimulus presence in PMv was initially of the same format as that in S1, it is subsequently transformed in such a way that the new format could not be generalized back to S1.

Regarding the discrimination of modal identity, just as for the decoding of stimulus presence, we found that training and testing along the same time points did not yield any time periods with significant and sustained decoding accuracy (Fig 4C). Extending sensory modality classifiers trained in S1 to PMv along the time generalization matrices did not demonstrate any time periods of successful classification (Fig 4C). On the other hand, when we trained in PMv and tested on S1, at 10 ms post-stimulus onset there was a higher than chance decoding accuracy in S1 along an array of time-points in the future. Thus, very early patterns of activity supporting the classification of modal identity in PMv are later found in S1. Thus, unlike stimulus presence, it appears that modal identity information generalizes in the direction of PMv to S1.

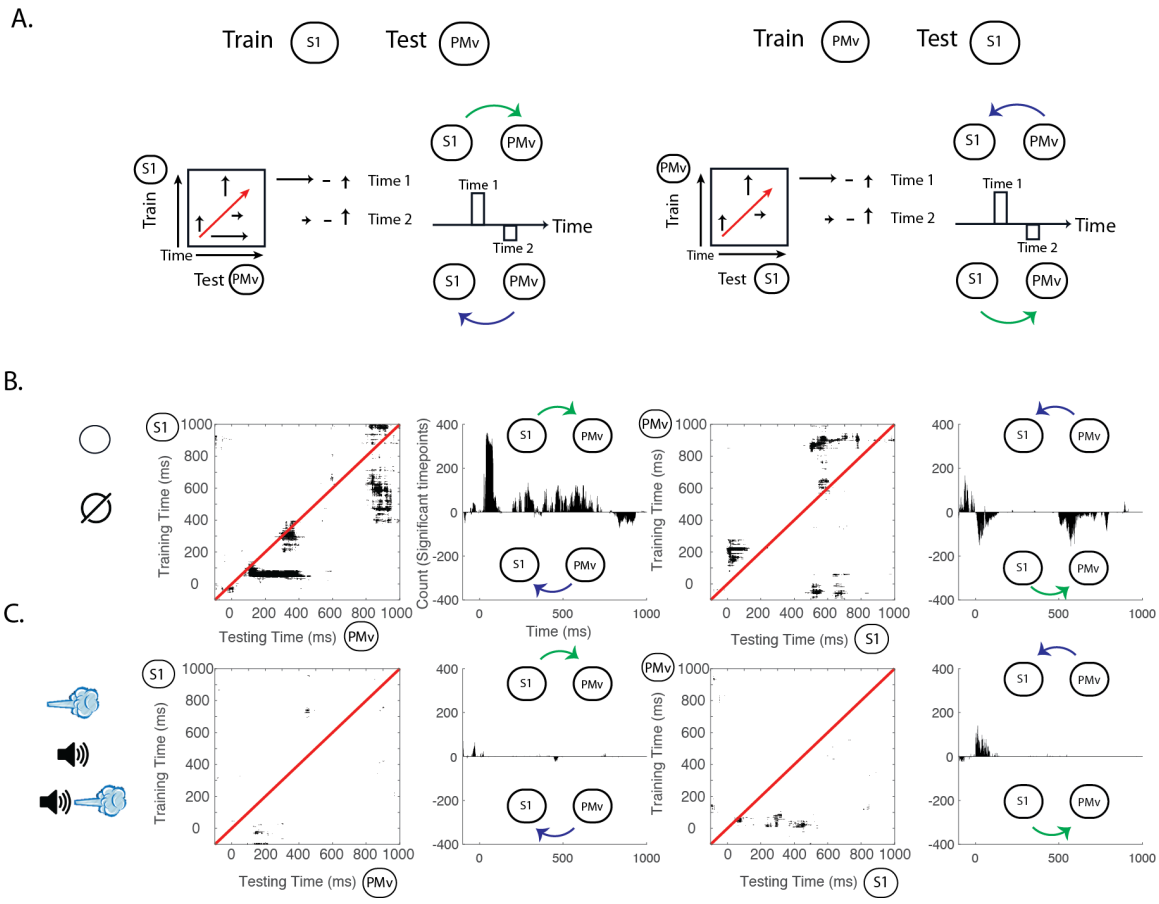

Figure 1.4: Cross-area time series decoding and time generalization. (A) Schematic showing two possible direction in which information can generalize. For training on S1 and testing on PMv. If information generalizes from S1 to PMv, a horizontal off-diagonal will be seen when training. On the other hand, an off-diagonal is vertical (i.e., later training periods can decode earlier ones), information is generalized in the direction of PMv to S1. (B) Cross-area off-diagonal examination for stimuli presence decoding in S1 and PMv. Results show a clear horizontal off diagonal when training in S1 and decoding in PMv. Training on PMv and decoding in S1 demonstrates a vertical off-diagonal (C) Cross area decoding of the identity of sensory stimuli. Training LDA in S1 does not afford the possibility of decoding sensory modality in PMv. Contrarily, training on PMv shows off-diagonal decoding along the testing dimension, suggesting information generalizes from PMv to S1.

1.6 Discussion

Simultaneous recordings of spiking activity across distinct nodes of a canonical sensorimotor circuit allowed us to study how information is shared and transformed between these areas. We recorded from arrays of electrodes placed in S1 and PMv – two areas known to be instrumental in transforming sensory information from different modalities (tactile, as well as auditory) into motor commands for action (Romo et al., 2004; de Lafuente Romo, 2005, 2006; Graziano et al., 1997; Noel et al., 2019). Specifically, we were interested in how information regarding stimulus presence and modal identity flowed and was altered between S1 and PMv and used different multivariate analyses to examine this question. The principal findings of the study are: 1) for decoding the presence of a stimulus, decoder performance fluctuated in a reciprocal manner between S1 and PMv for the interval up to 1 second after stimulus presentation, while decoding of modal identity was consistently higher in S1 than in PMv, 2) using time generalization, information regarding stimulus presence showed oscillatory strengthening and weakening dynamics in both S1 and PMv, while information regarding modal identity exhibited steady strengthening in S1 and weakening in PMv, 3) using cross-area time generalization, information regarding stimulus presence generalized between S1 and PMv, offset in time in the direction of S1 to PMv, while modal identity information only generalized weakly from PMv to S1. Together, these results highlight the different dynamics for the flow of information regarding stimulus presence and modal identity in two nodes of an important cortical sensorimotor circuit.

The findings fit within a larger and longstanding debate in neuroscience regarding whether sensory modality information is preserved as it ascends the processing hierarchy, or if that information ultimately transitions into an amodal format (Machery, 2016). Evidence for modality-specific information being preserved at high levels of representation comes from mental imagery, priming, and dreaming studies which show recruitment of sensory specific areas in the brain that are similar to their respective perceptual counter-

parts (Caramazza Mahon, 2003; Horikawa, Tamaki, Miyawaki, Kamitani, 2013; Ishai, Haxby, Ungerleider, 2002). On the other hand, evidence for amodal representations include task-specific recruitment of common brain areas for representations such as magnitude and numerosity regardless of sensory modality (Piazza, Mechelli, Price, Butterworth, 2006). Additionally, blind patients who hear sounds corresponding to objects viewed by sighted individuals shows similar brain activations (van den Hurk, Van Baelen, Op de Beeck, 2017). Our cross-area time generalization results provide evidence, that at least in the context of the passive delivery of stimuli studied here, as this perceptual information is hierarchically processed in the brain and transferred from sensory regions (S1) to regions closer to the motor circuitry (PMv), the representations become more amodal. Specifically, we found that information regarding stimulus presence in S1 generalized to PMv, but that information regarding modal identity only weakly generalized in the opposite direction from PMv to S1. However, it is important to note that our recordings were limited to S1 and PMv, and thus we cannot claim that modal identity is not preserved in other parts of the sensorimotor (or beyond sensorimotor) hierarchy. Ostensibly, the modal identity information transfer from PMv to S1 may represent the contribution of other nodes to modal identity that PMv is propagating backwards to S1.

In addition to what sensory information is transferred between brain areas, an equally important question is how sensory information is transformed as it ascends the sensory hierarchy. One important manner in which information can be transformed is through recurrent feedback (O'Connell, Dockree, Kelly, 2012). Notably, visual studies have found feedforward responses predominate during the first 200 ms of visual processing (Dehaene Changeux, 2011; Thorpe, Fize, Marlot, 1996) and thereafter recurrent process derived from temporal, parietal, and prefrontal cortices shape and ultimately transform the nature of these visual signals (Gold Shadlen, 2007; Gwilliams King, 2019; Kar, Kubilius, Schmidt, Issa, DiCarlo, 2019; Lamme Roelfsema, 2000). In our study, we found evidence of a recurrent process for encoding information regarding the presence of a stimulus in S1 and

PMv. Notably, when we compared results from the univariate and multivariate analyses, we found the decoding results to reveal that information pertaining to the presence of a stimulus was sustained for up to 1000 ms in both S1 and PMv, well beyond what averaged univariate responses revealed. This difference potentially reflects a change in information format from a standard rate code visible to univariate analyses to a code more reliant on sparse spatio-temporal patterns across the population that is only revealed through the application of multivariate methods. Further, information regarding stimulus presence was found to oscillate between strengthening and weakening in S1 and PMv up until 500 ms, after which it shows a steady strengthening (Fig. 3C). This oscillation coincides with an initial information transfer in the first 500 ms between S1 and PMv noted in the cross-area time-generalization results (Fig 4B). Thus, our results suggest that initial information regarding stimulus presence for auditory, tactile, and audiotactile stimuli is first transferred between S1 and PMv and finally transformed for subsequent decision making.

One caveat of the current results lies in the passive nature of the task, in which the monkey was not required to detect or discriminate between stimuli, but rather was only required to acquire the button after a start tone and release once it received a reward (thus maintaining vigilance). In many respects, this makes the results both surprising and compelling, in that there was no behavioral need to make use of the presented sensory information. Prior work in S1 has shown that responses to passive stimuli are depressed and have much more variability than when the animal is participating in an active process (Crochet Petersen, 2006; Schroeder, Wilson, Radman, Scharfman, Lakatos, 2010), and work in the visual system has shown that active tasks have longer sustained decoding than passive tasks when viewing identical stimuli (Ritchie, Tovar, Carlson, 2015). Collectively, this points to the current work representing an important foundation for future studies, as it illustrates the ability of decoding approaches to reveal differences in the dynamics of information flow in a classic sensorimotor cortical circuit – even when the stimuli are not used in the execution of an action. Future work should require animals to detect and/or discriminate between

sensory stimuli, in order to examine whether task demands potentially lead to longer periods of information transfer, and if information regarding modal identity is transferred in a discrimination task dependent upon stimulus identity. Moreover, an active task would allow the establishment of direct links between decoding performance and behavior through the use of distance from the decoding boundary in order to predict metrics such as accuracy and reaction time (Carlson, Ritchie, Kriegeskorte, Durvasula, Ma, 2014; Ritchie Carlson, 2016; Ritchie et al., 2015).

In conclusion, we have leveraged the ability to generalize neural activity across both space and time using multivariate techniques in order to garner insights into how information flows and is transformed from low level sensory areas to premotor areas in the brain, where it can be utilized for action. Specifically, we are able to provide empirical support that sensory information from S1 to PMv transitions to an amodal representation in the passive task that was employed. Importantly, our work provides a framework for which future work can explore how sensory information is transferred and transformed beyond just the two brain areas explored in this study and will allow examination of how task demands affect information flow.

1.7 References

Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: The evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, 13(10):1–19.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., and Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, 26(1):132–142.

Crochet, S. and Petersen, C. C. (2006). Correlating whisker behavior with membrane potential in barrel cortex of awake mice. *Nature Neuroscience*, 9(5):608–610.

Dehaene, S. and Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30(1):535–574.

Gwilliams, L. and King, J.-R. (2019). Recurrent processes emulate a cascade of hierarchical decisions: Evidence from spatio-temporal decoding of human brain activity. *bioRxiv*, page 840074.

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, 340(6132):339–346.

Ishai, A., Haxby, V. J., and Ungerleider, L. G. (2002). Visual imagery of famous faces: Effects of memory and attention revealed by fmri. *NeuroImage*, 17(4):1729–1741.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983.

Lamme, V. A. F. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23:571–579.

Machery, E. (2016). The amodal brain and the offloading hypothesis. *Psychonomic Bulletin and Review*, 23(4):1090–1095.

O'Connell, R. G., Dockree, P. M., and Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, 15(12):1729–1735.

Oosterhof, N. N., Connolly, A. C., and Haxby, V. J. (2016). Cosmomvpa: Multi-modal multivariate pattern analysis of neuroimaging data in matlab/gnu octave. *Frontiers in Neuroinformatics*, 10(JUL):27.

Piazza, M., Mechelli, A., Price, C. J., and Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fmri study. *Brain Research*, 1106(1):177–188.

Ritchie, J. B. and Carlson, T. A. (2016). Neural decoding and "inner" psychophysics: A distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in Neuroscience*, 10(APR):1–8.

Ritchie, J. B., Tovar, D. A., and Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology*, 11(6).

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., and Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology*, 20(2):172–176.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

van den Hurk, J., Van Baelen, M., and Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences*, 114(22):E4501–E4510.