BITS FROM BEHAVIORS: UNDERSTANDING FUNCTION USING INFORMATION IN EMBEDDED, EMBODIED, AND DYNAMICAL NEURAL NETWORKS

Madhavun Candadai

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the Cognitive Science Program
Indiana University

May 2020

Accepted by the Graduate Faculty, Indiana University,
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Doctoral Committee

Chair Eduardo J. Izquierdo, Ph.D.

John M. Beggs, Ph.D.

Joshua W. Brown, Ph.D.

David J. Crandall, Ph.D.

April 06, 2020

Acknowledgments

I have been lucky to have been mentored by Dr. Eduardo J. Izquierdo during my doctoral degree. He has been instrumental in shaping me as a scientist and I am grateful for the perspectives that I have developed regarding science, and research through his guidance.

I would like to thank Dr. John M. Beggs who has been a constant source of support, both scientifically as well as strategically, teaching me the nuances of navigating an academic environment.

I would also like to thank Dr. Martha White, Dr. Adam White, Dr. Josh Brown and Dr. David Crandall who have been on committees that have given valuable feedback in improving my work. Also, I am thankful for the very helpful feedback that Dr. Randall Beer and Dr. Ehren Newman provided that ultimately significantly improved my work.

I also greatly appreciate the support received from fellow researchers and friends. It has been invaluable both in my research as well as personally during my time here in Bloomington.

Of course, none of this would have been possible if not for the unconditional love and support from my parents, sister and my family. My heartfelt thanks go out to them and I shall be ever grateful for them.

Preface

Chapter 2. Parts of this chapter and small sections from other chapters have been published as a journal article (Candadai and Izquierdo, 2020). This work was performed in collaboration with Dr. Eduardo J. Izquierdo and this author is lead author on the manuscript and made all figures therein.

Chapter 3. A version of this chapter has been submitted to a journal and a pre-print has been made available (Candadai and Izquierdo, 2019b). This work was performed in collaboration with Dr. Eduardo J. Izquierdo and this author is lead author on the manuscript and made all figures therein.

Chapter 4. The material presented in this chapter summarizes the work of two conference publications (Vasu and Izquierdo, 2017a; Candadai and Izquierdo, 2018). Both works were performed in collaboration with Dr. Eduardo J. Izquierdo. This author is lead author on both works and made all figures therein.

Chapter 5. A version of this chapter was published as a journal article in collaboration with Matthew Setzler, Dr. Eduardo J. Izquierdo and Dr. Tom Froese (Candadai et al., 2019). This author is lead author on the publication and made all figures therein.

Madhavun Candadai

BITS FROM BEHAVIORS: UNDERSTANDING FUNCTION USING INFORMATION IN EMBEDDED, EMBODIED, AND DYNAMICAL NEURAL NETWORKS

One of the greatest scientific challenges of this century is to understand how the brain produces behavior. In this regard, information theory has emerged as one of the main theoretical frameworks that allow us to study neural information processing at all spatial and temporal scales. However, most work in this domain involves studying the brain in isolation without taking into consideration the body, the environment and the interaction between them. The goal of this dissertation is to advance our understanding of the neural basis of behavior by incorporating the environment into the information-theoretic analyses. Specifically, using computational models, I demonstrate that including the environment and agent-environment interaction in our study helps better understand three prominent phenomena of adaptive behavior, namely predictive coding, multifunctionality, and multiagent interaction. Under predictive coding, besides demonstrating that the environment is a source of predictive information, I present a framework based on Partial Information Decomposition to disentangle the source of information encoded in a neural network: extrinsically provided by the environment versus intrinsically generated in the neural network. Under multifunctionality, besides demonstrating that environmental feedback alone can drive multiple behaviors in the same sensorimotor system, I provide proof-of-concept about the extent to which neural resources can be reused across tasks; the same neural activity can produce multiple behaviors in an embodied agent. Finally, under multiagent interaction, I show that mutual interaction with other agents can result in neural dynamics that are qualitatively different from neural dynamics achievable in one-way interaction or isolation. Altogether, I show that informational structure either provided by the environment or acquired via agent-environment interaction shapes neural information encoding as well as ongoing neural dynamics in a way that can only be understood by expanding our unit of analysis to go beyond the brain and include the environment as well.

Chair Eduardo J. Izquierdo, Ph.D.

John M. Beggs, Ph.D.

Joshua W. Brown, Ph.D.

David J. Crandall, Ph.D.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

One of the greatest scientific challenges of this century is to understand how the brain produces behavior. In this regard, information theory has emerged as one of the main theoretical frameworks that allows us to study neural information processing at all spatial and temporal scales. However, most work in this domain involves studying the brain in isolation without taking into consideration the body, the environment and the interaction between them. The goal of this dissertation is to advance our understanding of the neural basis of behavior by incorporating the environment into the information theoretic analyses. More specifically, we demonstrate that three prominent phenomena of adaptive behavior, namely predictive coding, multifunctionality and multiagent interaction cannot be fully understood unless the environment and agent-environment interaction are included in our study. In each of these contexts we analyze how informational structure, either provided by the environment or acquired via interaction with the environment, shapes neural information encoding as well as ongoing neural dynamics.

The research approach adopted in this dissertation follows a long line of work that employs computational brain-body-environment models to advance our understanding of the adaptive behavior (Beer and Gallagher, 1992; Harvey et al., 2005). This involves building idealized computational models of dynamical neural networks, optimizing them to perform tasks specifically designed to study each concept, and then analyzing an ensemble of successful models to better understand how the model's neural dynamics reflect those concepts. The models were designed to be minimal in-

tentionally, to make in-depth analysis tractable. The tasks were designed such that the ground-truth knowledge about the relevant variables is available. Further, an evolutionary algorithm was used to optimize the parameters of the neural network controlling the behavior of the agent. Evolving generic neural networks instead of hand-designing specific architectures allowed us to make minimal prior assumptions about how various behaviors must be implemented in the neural network thereby maximizing the potential to reveal counter-intuitive solutions to the production of behavior (Harvey et al., 1996; Harvey, 1997, 2000). Finally, tools from information theory (Shannon, 1948), including its recent extensions for multivariate data (Williams and Beer, 2010b), were utilized to understand information processing in the optimized models. This is the general approach adopted to study predictive coding, multifunctionality and social cognition.

Under predictive information, we demonstrate that the environment is a source of predictive information and environmental regularities shape predictive information dynamics in neural networks (Candadai and Izquierdo, 2019b). One of the prevailing theories of adaptive behavior is that organisms are constantly making predictions about their future environmental stimuli. In explaining the role of predictive information in adaptive behavior, two prominent theories propose different sources from which it is acquired by an organism: predictions are generated from an agent's internal model of the world or predictions are extracted directly from the environmental stimulus. In this work we demonstrate that predictive information, measured using mutual information between current neural activity and a future stimulus, cannot distinguish between information from these two sources. To understand the role of predictive information in adaptive behavior, we need to be able to identify its source: environment or neural network. To do this, we propose a framework where we decompose information transfer across the different components of the organism-environment system and track the flow of information in the system over time. We validate this framework on a set of computational models of idealized agent-environment systems. Analysis of the model systems revealed three key insights. First, the source of predictive information can change dynamically during the course of a behavior. Second, predictive information provided by the environment is encoded in any agent irrespective of its ability to perform a task. Third, the magnitude of predictive

information in a system can be different for the same task if the environmental structure changes.

In our study of multifunctional neural networks, we demonstrate that environmental feedback enables reuse of neural resources in the same neural network across different tasks down to the time-scale of ongoing neural activity, even in the absence of neuromodulation or plasticity (Vasu and Izquierdo, 2017a; Candadai and Izquierdo, 2018). One of the most interesting properties of living organisms is their ability to perform multiple behaviors across different environments. It is increasingly accepted that the same neural networks can be responsible for multiple behaviors. However, it is often assumed that these networks undergo dynamical reconfiguration due to neuromodulation or neural plasticity to perform the different tasks. In this work, we demonstrate that neural networks can perform multiple tasks (object categorization and pole-balancing) with no changes to parameters or a task-identifying signal. In order to understand this further, it is important to understand how the network operates under the different task conditions. Analysis of these multifunctional neural networks yielded three key insights. First, the effective networks estimated from neural activity during different trials of each task were more similar to each other within trials of the same task than to effective networks from trials across tasks. Second, the effective networks for each task, while distinct, were not disjoint and reused the same neurons or synapses. Finally, to understand the reuse of neurons and synapses further, we performed a dynamical systems theoretic analysis which revealed that neural resources were reused down to the level of ongoing neural activity during the behaviors. In other words, the same neural activity can drive multiple behaviors. Only including the state of the environment, allows us to distinguish the behaviors, and merely looking at the neural activity alone does not.

Our work on multiagent interaction demonstrates that mutual interaction with a dynamical environment significantly impacts neural activity to the extent that studying animals in isolation may not reveal the true neural complexity they exhibit in the real world (Candadai et al., 2019). Living organisms and their environments are both dynamical systems that are constantly influencing each other. However, in most studies, any stimuli provided to the brain are pre-designed and do not dynamically respond to the agent. Is environmental feedback merely a source of complex inputs to

an individual or does it enhance the agent's intrinsic richness of neural activity? In this work, we study whether the complexity (Shannon entropy) of neural activity varies in a mutually interacting environment as opposed to in a static environment. More specifically, we empirically compared the maximum complexity of neural dynamics achievable under three conditions - no stimulus, passively receiving complex stimulus and actively receiving complex stimulus via interaction with another agent. This study revealed that neural networks of the same size present significantly different levels of neural complexity in each condition, with most complexity exhibited during active interaction with another agent. These results show that environmental interaction can produce qualitatively different neural dynamics and therefore needs to be taken into account in the study of neural basis of behavior.

Altogether, this dissertation demonstrates the influence of the environment on neural network operation through the lens of information theory at several levels: as a source of information, as enabling degenerate information encoding across tasks, and finally as a source of enhancing richness of neural activity. In addition to the theoretical insights, we provide open-source Python tools to perform similar information theoretic analyses on experimental data (Candadai and Izquierdo, 2019a). Ultimately, this dissertation underscores the need to consider characteristics of the environments and its dynamics over time to develop a comprehensive understanding of the principles that underlie neural information processing in adaptive behavior.

1.2 Thesis organization

The thesis is organized as follows. In the next (second) chapter, I discusses the methodological frameworks that has been adopted in the work presented in this dissertation. Following that chapter, the main contributions of the dissertation have been organized into three chapters titled "Predictive Coding", "Multifunctionality", and "Multiagent Interaction". In each chapter, I present the motivating reasons for each of those projects, survey relevant literature in those domains, explain the specific methods adopted in that project and finally present and discuss the results. Finally, in a concluding chapter I present an overall summary and discuss the implications of the results from

this dissertation to the study of neural basis of behaviors. All work presented here was performed with the guidance of my supervisor, Eduardo J. Izquierdo.

In Chapter 2, I discuss and justify the computational modeling approach that has been adopted in this work. I present examples of past work that have utilized computational models to advance our understanding of neural network operation. Following that, the different neural network models are presented and contrasted against each other, and similarly optimization methodologies that are capable of training neural network models to perform different tasks are discussed. Following these model building and optimization sections, the main analysis tools that have been adopted in this dissertation are explained. Specifically, I outline information theoretic tools such as mutual information, partial information decomposition and transfer entropy. Next, past work in Neuroscience that have used these tools are discussed, followed by a section that presents all available tools that facilitate performing such analyses. Finally, I provide a description of our approach: building computational models of neural networks, optimizing them to perform specific tasks using evolutionary algorithms, and analyzing them using information theory.

In Chapter 3, the first main contribution of this dissertation is presented: a theoretical framework to detect the source of predictive information in dynamical neural networks. The first section of this chapter motivates this work in the context of two existing streams of research that attributes different sources to predictive information in neural networks. Following that, a more in-depth survey of these two research streams is provided along with my ideas for the potential for their convergence. Next, the methodological details of our model is presented, following which the results are presented and their implications are discussed.

In Chapter 4, the second main contribution of this dissertation is presented: neural resources can be reused across multiple behaviors down to the level of on-going neural dynamics. This chapter first motivates the study based on the pervasiveness of the ubiquitous ability of living organisms to perform multiple behaviors often using the same neural resources. Next, a survey of existing literature on how multifunctionality has been studied along with different mechanisms that facilitate neural reuse that have been proposed and observed are presented. Following this, the

model and analysis methods used in our work is outlined. Finally, the results from the analysis of a multifunctional agent are presented, followed by a discussion of the implications of our results.

In Chapter 5, the third main contribution of this dissertation is presented: neural dynamics in the presence of two-way multiagent interaction is qualitatively different from what can be achieved in one-way interaction or isolation. The first section of this chapter motivates this work in the context of brain-in-a-vat approaches to understand cognition and the contrasting enactive approach. Following that, existing work that emphasizes the role of social interaction in enabling behaviors that are not possible in its absence is discussed. Next, the modeling approach taken in the presented work is explained, followed by the results and a discussion of its implications.

In the final chapter, I present a summary of the contributions of this dissertation along with suggestions for extensions of the work presented here. Finally, I discuss the implications of our results for the study of neural basis of behavior.

Chapter 2

Background

In this chapter I discuss and justify the approach to research taken in this dissertation in light of previous work that have taken a similar approach. First, I discuss the benefits of employing computational models in science, specifically in Neuroscience with examples of previous work that demonstrate how they have helped advance our understanding of neural network operation. Second, I discuss the different components of building computational models such as choosing a neural network model and optimization algorithm. Third, I explain how the tools of information theory have been used to understand neural information processing. Finally, I present how our approach to research combines these above mentioned methodologies to study the neural basis of behaviors.

2.1 Computational models in Neuroscience

The number of questions that we as scientists ask about natural systems, and the number of hours it would take to perform the experiments required to answer those questions far outnumber the number of hours we collectively have. This is true just for the questions that we know to ask at this moment in time. A hallmark of seminal work in any scientific discipline is its ability to open up the possibilities for new questions to be asked. Furthermore, the solutions that evolution found to problems that natural systems face, significantly dwarfs our imagination and problem-solving capabilities. Thus, the quest for understanding natural systems is a challenging feat that scientists take on fully aware that their best-case-scenario would be to make a tiny dent in the otherwise vast

ever-expanding envelope of scientific progress. Rest assured, this bleak opening statement about the limited capacities of the human intellect, while humbling, is a segue into the benefits of taking a computational modeling approach to science.

Thought-experiments have long been drivers of scientific progress by identifying areas for experimentalists to focus on, based on careful observation of known facts of a phenomena. Simulation models can serve as a mathematically grounded replacement to thought-experiments (Di Paolo et al., 2000; Nersessian et al., 2012), with the potential to go beyond the biases that we as humans bring into thought experiments. Models of natural systems are often considered to be built as a tool to generate predictions about the system under study, like a model of thunderstorms used to predict its movement trajectory. However, the objective of building models goes beyond making predictions (Epstein, 2008). Models serve to: provide an explanatory account of observed phenomena, guide experimental design and data collection based on hypothesis generated from analyzing the model, suggest analogies by demonstrating parallels between different natural systems, provide existence-proofs for mechanisms underlying behavior of natural systems thus raising new scientific questions, explore the hypothesis space at a comparatively low cost compared to experimental settings, and finally, act as idealized test beds for novel analytical methods since the ground-truth knowledge is available about these systems. Importantly, the scientific approach to modeling involves building “opaque” models constrained only by known details, only to then be analyzed later to reveal the internal workings. As a result of this, modelers naturally adopt the "embracing multiple hypotheses" approach, an approach that arguably leads to better and less-biased science (Chamberlin, 1890). When integrated into a modeling-experiment cycle, computational models can continuously inform future experimental design while getting updated based on results from past experiments (Alexander and Brown, 2015). This virtuous cycle of experiments and models informing each other has been proven to provide a structured way to explore the intractable hypothesis space for understanding natural systems (Izquierdo, 2019).

Computational models (or models of any kind) are tractable versions of the complex natural system that we aim to understand. Justifiable simplifications in combination with powerful com-

puting systems have enabled the construction of a myriad of computational models in Neuroscience ranging from models of individual neurons to integrated brain-body-environment models. Some of the apparent practical benefits of computational models are as follows: first, unlike experimental in settings computational models are fully observable - all variables are completely accessible at all times of the simulation, and beyond being observable, all variables in the system are manipulable and enable ablation as well as stimulation tests in any part of the system. However, like the statistician George E. P. Box eloquently stated, "All models are wrong but some are useful". Computational models are idealized abstractions of the system of interest and therefore necessarily make simplifying assumptions in their design. Consequently, insights obtained from models unlike those from experiments do not serve as definitive proof but instead provide hypotheses and existence proofs that would then need to be tested.

Cohen (2004) delineated six types of questions that are asked by scientists: How did it begin? (Origins), How is it built? (Structures), What is it for? (Functions), How does it work? (Mechanisms), What goes wrong? (Pathologies), and How is it fixed? (Repairs). Computational models are useful in Neuroscience to the extent that they help answer at least one of these questions. In this section, I outline examples of computational models being applied to each of these questions within the domain of Neuroscience. I only point out examples to provide a proof-of-existence; this is by no means an exhaustive list.

2.1.1 How did it begin? (Origins)

New phenomena originate in natural systems at several time-scales, from the first time it appears in any species such as the first occurrence of a neuron in evolutionary time (Villegas et al., 2000); to the first occurrence in an individual over developmental time scales such as the origin of cortical interneurons hypothesized to be in cortical subventricular zone (Wonders and Anderson, 2006); to the origin of specialized neural dynamics during on going behavior such as what causes different types of oscillations to begin in the same neural circuit of the crab stomatogastric ganglion (Marder and Bucher, 2007). An example of how computational models have been used to answer questions

of origin is the work by Krishnan et al. (2018), that provided a hypothesis for the origin of spontaneous resting state dynamics observed in fMRI, EEG and local field potential recordings. Using a computational model of intra- and extracellular K^+ and Na^+ ion flow across the brain network based on the CoCoMac connectivity data, they were able to reproduce the infra-slow fluctuations in brain activity. Thus, a computational model has helped hypothesize that the origin of spontaneous neural dynamics could be due to dynamics of ion flow mediated by neuronal and glial activity.

2.1.2 How is it built? (Structures)

Sometimes, there is only partial knowledge of the different components of a natural system and a complete understanding of its function requires that the gaps are filled. For instance, even in *C. elegans* which has been nearly completely genetically, developmentally and neurologically mapped, while the entire connectome of the 302 neuron network is known, the polarity of the connections (inhibitory vs excitatory) are not known. This is particularly true in larger animals where the micro-connectome is not fully known. An example of where computational models can be utilized to advance our understanding of such systems is the work by Real et al. (2017), where they optimized a model to match observed functional characteristics constrained by known structural properties to make predictions about the unknown structural attributes of the vertebrate retinal ganglion. Specifically, they started with an existing model made with known anatomical details and built four models that incorporated different anatomical characteristics in sequence: first, the addition of a non-linear pooling layer; second, a feedback loop around the non-linear pooling layer; third, another feedback loop around but this time around a different downstream layer that feedback of that matched firing rate of ganglion cells; and finally, a delay in processing before the pooling layer. These four modifications were chosen based on the expectation of matching increasingly more functional characteristics that were observed in the retinal ganglion. They showed that these four modifications when applied in a cascaded fashion performed progressively better in explaining the variance in activity of the ganglion cells. Importantly, the model was re-discovered known

attributes of ganglion cells such as center-surround receptive fields at appropriate scales in each layer in accordance with observed experimental data. Furthermore, they conducted experiments to test if the nature of the receptive fields formed in the model matched with what can be observed biologically and found that they indeed matched. Thus, this study is an example for how models can help answer the question of how a neural network is built to perform a specific function.

2.1.3 What is it for? (Functions)

A sub-problem in developing a thorough understanding of a natural system is understanding the role played by its different component systems. Experimental results often provide an incomplete account of the role of a component because experiments are often geared towards testing one specific hypothesis. As a result, different experiments might provide seemingly conflicting accounts that will require reconciliation. The computational modelling example for answering questions of function comes from the work of Alexander and Brown (2010) regarding the role of Anterior Cingulate Cortex (ACC) in reinforcement learning. Their Predicted Response-Outcome model involved representing a vector-valued prediction of a sequence of future states of the prefrontal cortex given the current state and action. Further, in accordance with existing experimental data they reinterpreted the polarity of the prediction error to mean positive surprise (events that were not predicted but occurred) and negative surprise (events that were predicted but did not occur). Crucially, according to this model, it does not matter if a predicted state is desirable or undesirable, but it only matters whether the outcome was predicted or not. This generalized model of the role of ACC as providing "valence-neutral prediction error" provided a unified account of seemingly disparate experimental observations that was thought to be impossible to reconcile. Additionally, the model makes predictions for action selection based on a combination of learned state-response associations and response-outcome associations. Thus, computational approaches have enabled identification of the functional role of the ACC thereby helping answer what is it for.

2.1.4 How does it work? (Mechanisms)

In addition to understanding the role of individual components, it is crucial to understand how the different components interact to produce the observable phenomenon. This is another context where computational models could help generate testable hypotheses. Such an example is the work of Olivares et al. (2018) which provided existence-proof of purely central pattern generator driven locomotion in *C. elegans*. They built a neural network model of the repeating neural unit of the segmented worm constrained by the known connectome, and optimized the unknown parameters such that the model generated neural activity that matched dynamics expected during locomotion. Analyzing an ensemble of models that performed similarly revealed multiple ventral nerve cord architectures that could result in oscillatory dynamics suitable for locomotion. The most frequently found solution in the ensemble involved a dorsal oscillatory circuit that drove the ventral out-of-phase oscillations thereby producing the requisite alternating dynamics for locomotion. Following this, experimental studies have provided support for the possibility of intrinsic oscillations in the ventral nerve cord (Xu et al., 2018; Fouad et al., 2018; Gao et al., 2018). Thus, computational models can help explore the potential space of mechanisms that can produce a behavior.

2.1.5 What goes wrong? (Pathologies)

The first step in treating any pathology is understanding its cause. To this end, several hypotheses might be proposed that would result in the observed phenotypic abnormality. For instance, schizophrenia was theorized to be caused by disruptions in the interactions between dopamine and the prefrontal cortex thus disabling the individual by not being able to hold or update information about environmental context. Braver et al. (1999) built a computational model that incorporated a noisy dopamine signal and the subsequent improper modulation of information processing in the prefrontal cortex. This model replicated the expected behavioral aberrations in a continuous performance test that has been shown to capture critical aspects of cognitive control that are affected in schizophrenic individuals. In fact, such approaches have led to development of the sub-field devoted to applying computational approaches to mental disorders: computational psychiatry (Wang

and Krystal, 2014). Thus, a computational model has provided further support for a theory by concretizing it in a mathematical framework.

2.1.6 How is it fixed? (Repairs)

Besides being subject to pathological interruptions, natural systems are capable of recovering and self-repair in a variety of ways. Computational models can not only model the pathology in neural networks but also the process that following a pathological interference. One such example is the work of Naeem et al. (2015) who built a computational model of the self-repair process in neural circuits mediated by astrocytes. They proposed a new learning rule that mediated interaction between astrocytes and neurons to reproduce the self-repair process. While the astrocytes were coupled to neurons they also implemented communication between astrocytes to achieve network level repair. They demonstrated that their proposed learning rule reestablished the firing rates of neurons post failure of synapses in a neural network. Thus, computational models enable study of neural processes that facilitate repair following trauma.

2.1.7 Validating analytical methods

In addition to the scientific questions listed above, the fact that the ground-truth about operations in computational model is known has been used to demonstrate the validity of analytical methods. For instance, Ito et al. (2011a) proposed an extension to transfer entropy that will allow the detection of transfers at different delays, and demonstrated its capability by testing it on spiking neural network model with a known connectivity pattern and range of synaptic delays. While the existing approach that only accounted for delays of one time step captured only 36% of the true connections in the network, their proposed enhancement captured as much as 73% of them. Thus, computational models serve as an ideal test bed for evaluating and validating analytical methods before applying them to experimental data.

2.1.8 Replicating natural systems

Inspired by the adaptability and robustness of natural systems, the entire field of Artificial Intelligence aims to replicate natural intelligence in digital systems. To this end, neural network models of various kinds have been developed to mimic biological neural processes. Tremendous progress has been made in the last few decades with the advent of optimization methods such as backpropagation that enable training multi-layer neural networks. With human- and super-human-level performance as the goal, such artificial systems have become masters of strategy games such as chess (Silver et al., 2018), Go (Silver et al., 2017), as well as Arcade games (Mnih et al., 2013) and multiplayer video games (Berner et al., 2019). From a practical standpoint, such models have been adopted for aiding cancer diagnosis (Levine et al., 2019), aggregating information about global events from public crowd-sourced images (Wang et al., 2013) and so on. While such systems have a long way to go in order to achieve the robustness of natural systems, rapid progress made in the last decade shows promise for the future.

Altogether, a computational modeling approach spans all levels of scientific inquiry from origin to destruction. The examples provided here are one of several in each of those domains. As mentioned previously, computational models provide unique benefits that enable in-depth analysis and manipulability that cannot be achieved in experimental settings. Computational models have a long history of aiding experimentalists, and will continue to develop along side experiments as a resource for generating hypothesis, testing feasibility of theories and validating analytical approaches.

2.2 Neuron and neural network models

Computational models of neurons and neural networks have been built at many different levels of abstraction, on a spectrum from biophysically realistic models to simplified linear models. These models are also on a complementary spectrum of computational efficiency spectrum where the most biophysically realistic models are the most computationally expensive to linear models that are least computationally expensive. Several models exist on this spectrum and models can be appropriately

chosen that is a trade-off between computational efficiency and biophysical realism. Some of the key characteristics that decide where a model would lie on these spectra are: spiking versus non-spiking models, with or without internal state, and recurrent versus feed-forward connectivity. Within neuron models that have an internal state, there is a distinction between continuous- versus discrete-time systems that are defined by differential and difference equations respectively. While this is a theoretical distinction, the distinction is not significant in simulation since continuous-time systems are also simulated in digital computers using discrete approximations such as Euler integration. In this section, I describe the primary neuron models adopted in the work presented in this dissertation. I explain where they lie on the spectrum of biophysical realism as well as computational efficiency in relation to other models.

At the very end of the biophysically realistic spectrum lies the Hodgkin-Huxley model Hodgkin and Huxley (1952). It was published in 1952 by Alan Hodgkin and Andrew Huxley based on perhaps the most extensive studies of biological neurons, specifically the squid's giant axon. A single biological neuron is modeled as an electrical circuit where parameters of ion flow across the cell membrane is modeled using electrical analogues such as resistors and capacitors. This non-linear continuous-time model exhibits spiking activity extremely identical to to biological neurons when injected with external current. As such, it is a spiking neuron model defined by continuous-time differential equations, and is on the far end of the biophysically realistic models side of the spectrum. Consequently, simulating a single neuron per this model is quite expensive computationally. Since the model is that of a single neuron, it does not place an emphasis on how neurons are connected to one another and therefore theoretically allows any kind of connectivity. Following the Hodgkin-Huxley model, a simplified version of it was proposed in 1961 by Richard FitzHugh (FitzHugh, 1961), and later augmented with a circuit model by Nagumo et al. (1962). With the main motivation of developing a model of the injection and propagation of current through a neuron, this simplified version abstracted away from independently modeling individual ion flows. For weak external excitation, the model produces activity that resembles subthreshold membrane potential dynamics. With a sufficiently high external current, the model produces

spikes. Importantly, the model does not exhibit all-or-none spiking activity: for intermediate external excitation, the model produces spikes of smaller magnitudes. Finally, with consistently high external stimulation, the model produces oscillatory spiking activity. With only 2 state-variables as opposed to Hodgkin-Huxley's 4, this model enables a visualization of the entire phase-space of the neural dynamics and its phase-space has been analyzed extensively (Izhikevich and Moehlis, 2008). Since the model has abstracted away from modeling ion-channels in detail, it is computationally more tractable than the Hodgkin-Huxley model. It is still biophysically realistic in its ability to produce spiking and refractory dynamics similar to biological neurons. While these two models form the foundation for biophysical modes of neuraons, several others have been proposed at different levels of simplification: Izhikevich spiking model Izhikevich (2003), Integrate-and-fire family of models (Abbott, 1999), Continuous-Time Recurrent Neural Network (CTRNN) Models (Beer, 1995b), Long-Short Term Memory discrete time models (Hochreiter and Schmidhuber, 1997), and the discrete time state-less Perceptron model (Rosenblatt, 1957).

From this non-exhaustive list of neuron models at different levels of abstraction and computational complexity, appropriate models can be selected depending on the purpose for which the model is built. For instance, modeling specific brain regions such as hippocampal neurons requires that a biophysically realistic model of spiking neurons is chosen with appropriate computational complexity: integrate-and-fire neurons. Conversely, if the goal of building the models is to solve an engineering problem, biophysical realism might take a backseat giving way to LSTMs or multi-layer perceptrons as the model of choice. Finally, understanding general principles underlying neural information processing in living organisms, as is the goal of this dissertation, does not necessarily have to use spiking neuron models and can employs CTRNNs as approximations of rate-coding. Of these, the Izhikevich and CTRNN models lie at an optimal trade-off between biophysical realism and computational tractability. For this reason, they have been employed in the models built in this dissertation. These models are discussed in detail below.

2.2.1 Izhikevich model

As a follow-up to the two-dimensional FitzHugh-Nagumo model, Eugene M. Izhikevich published his model in 2003 (Izhikevich, 2003). This model of spiking neurons included and explicit threshold based spiking in order to produce all-or-none spiking dynamics unlike that of the FitzHugh-Nagumo model. Mathematically, this model defines rates of change in membrane potential, V, and a recovery variable, U, according to

$$\begin{aligned} \dot{V} &= 0.04V^2 + 5V + 140 - U + I \\ \dot{U} &= a(bV - U) \end{aligned} \tag{2.1}$$

with explicit threshold based spiking and reset of state-variables as follows

$$\text{if v } \geq 30 \text{ mV, then} \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases} \tag{2.2}$$

where a, b, c, and d are parameters whose ranges have been thoroughly studied and mapped to characteristic behaviors. As such, this model provides an optimal trade-off between computational complexity and biophysical realism, because it is a 2-D continuous-time, spiking neuron model that can replicate various spiking dynamics observed in cortical neurons such as intrinsic bursting, low-threshold spiking and so on.

This model's utility has been demonstrated in several ways: its ability to perform a range of linear and non-linear pattern recognition tasks (Vazquez, 2010), as models of specific brain regions (Prescott et al., 2006), as generic models of cortical activity to validate analytical methods (Ito et al., 2011a), as well as neural controllers in brain-body-environment models of adaptive behavior (Vasu and Izquierdo, 2017b).

2.2.2 Continuous-Time Recurrent Neural Network model

Moving from spiking to non-spiking neuron models, Continuous-Time Recurrent Neural Networks (CTRNNs) are models of dynamical recurrent neural networks with internal state. CTRNNs are perhaps the simplest non-linear, continuous dynamical neural network models, and yet are universal dynamics approximators (Funahashi and Nakamura, 1993). As their name denotes, they are continuous-time models and are recurrently connected. Thus, they are an obvious choice to build computational models of neural networks to perform tasks that involve stateful interaction with the environment like living organisms do. and a network is defined a vector differential equation:

$$\tau \dot{y} = -y + W\sigma(y + \theta) + I \tag{2.3}$$

where τ, the time-constants, y, the internal-states of the neurons, θ, the biases of the neurons and I, the external input are all N dimensional vectors corresponding to the N neurons that make up the circuit. W is a synaptic weights matrix denoting the fully-connectedness of the circuit, and σ is the sigmoidal function whose value gives the output of the neurons. If interpreted as a spike-rate model, $\dot{y}$ denotes the membrane potentials, $\sigma(.)$, the mean firing rates, θ, the firing thresholds, and W_{ii}, the self-weights represents a simple active conductance.

The fact that that they are not spiking models does not entirely discount their place on biophysical realism because there has been evidence of non-spiking neurons in locusts (Siegler and Burrows, 1979), crayfish (Takahata et al., 1981), and most neurons in *C. elegans* (Lockery and Goodman, 2009). If interpreted as models of non-spiking biological neurons, $\sigma(.)$ can instead represent the saturating non-linearities in the synaptic inputs.

This model's utility has been demonstrated in several ways: to approximate a variety of dynamical systems (Funahashi and Nakamura, 1993), to build biologically-grounded models of living systems (Izquierdo, 2019), as generic models of neural dynamics to validate analytical methods (Candadai and Izquierdo, 2019b), as well as neural controllers in brain-body-environment

models of adaptive behavior (Beer and Gallagher, 1992).

2.3 Model optimization approaches

The variety of computational models of neurons and neural networks described in the previous section, all have parameters that need to be tuned such that the model behaves as desired. Such tuning cannot be performed by hand-designing these parameters. Several optimization methodologies have been proposed for this purpose. Generally speaking, the optimization process involves defining an objective function as a function of the parameters of the model, and identifying the optima on the objective function landscape. Based on the nature of the objective function, model training falls under one of three approaches: supervised, unsupervised and reinforcement learning. Neural network models where the true desired output of the model is known can be optimized using an objective function that minimizes the error between actual and desired outputs. Such a training paradigm is defined as supervised learning. Alternatively, when the ground-truth outputs are not known, under the paradigm of unsupervised learning, objective functions involve inferring statistics from the training data. Finally, reinforcement learning involves optimizing model parameters to maximize the objective function defined as the expected long-term reward in sensorimotor tasks. Based on the objective function, parameter optimization approaches fall under two primary categories: gradient-based and search-based approaches.

Given an objective function, Q, and a set of parameters, θ, gradient-based approaches involve estimating the direction in parameter space to move so as to improve the model performance. This is achieved by estimating the derivative of the objective function with respect to the parameters. Consequently, this requires that the objective function for gradient-based optimization is differentiable. Typically, estimating the true gradient of the objective function is intractable. The gradient is often estimated using a random sample of training data leading to stochastic estimates of the true gradient hence called stochastic gradient-descent. For supervised learning the data is a random sample of training data and labels, and for reinforcement learning, gradients are estimated from sensory input, motor action, and reward data points collected during the course of a sensorimotor

task. Since this approach is sensitive to learning rate, several variants have been proposed to this approach that modulate the learning rate: adding a momentum term where learning rate is increased when updates progressively happen in the same direction, and decreases otherwise (Rumelhart et al., 1986); adaptive gradient, or AdaGrad, where each parameter has its own learning rate that is increased for sparser parameters and decreases learning rate for less sparse parameters (Duchi et al., 2011); Root Mean Square Propagation, or RMSProp, that also has per-parameter learning rates which are modulated based on past-updates to the parameters (Tieleman and Hinton, 2012); and finally Adaptive Momentum Estimation, or Adam, is an update to RMSProp where learning rate of each parameter is not just updated based on average of past gradients but also variance of past gradients (Kingma and Ba, 2014). This general approach to incrementally updating model parameters can be applied to any objective function that is differentiable. Depending on the task, the objective function can either require minimization of an error, or maximization of returns.

An alternative to gradient-based approaches to model optimization are search based approaches. These typically involve a population of solutions searching the parameter space in parallel, and exchanging information on their relative performance in some form according to an objective function as described previously. However, since updates to parameters are not based on estimating gradients, these approaches do not require that the objective function be differentiable. Consequently, the objective function can be more interpretable measures of model behavior. The most popular search based approaches are genetic and evolutionary algorithms (Mitchell, 1998; Holland et al., 1992). These algorithms get their name since they are inspired by the natural evolutionary process of fitness-based selection. The algorithm is inherently a maximization algorithm on a population of solutions, where each individual in the population is represented as a genotype, a point in the N-dimensional parameter space. In its simplest implementation, each step, or generation, involves: first, estimating the performance or fitness of a subset or the entire population (depending on which variant of evolutionary algorithms are employed) using the objective function; second, copying parameters from the highest fitness individuals to the low fitness individuals or recombination; and third, adding random noise to the low fitness individuals, or mutation. This process is repeated until

at least one solution of a desired fitness is found, or until a predetermined computational budget has been exhausted.

Variations to evolutionary approaches typically involve different methods of selecting individuals to recombine: rank-based selection where individuals are ranked based on their fitness and picked with probabilities according to their rank; fitness-proportionate selection where a probability distribution of selecting individuals is constructed proportional to the relative fitness of the different individuals; elitist-selection where a top $X\%$ of high-fitness individuals are selected to be preserved as is for the next generation; and finally, tournament selection where two individuals are picked at random and pit against each other with the high-fitness individual being preserved as is while "transfecting" the low-fitness individual that is then mutated.

Variation to evolutionary approaches also involve multiple recombination strategies: one-point crossover where all genes in a genotype that are on one side of an arbitrarily chosen crossover point are copied; two-point crossover, where all genes between two points are copied; and the more general k-point crossover which is performed over k different crossover points. In the most general case, every gene is independently crossed over with a specified recombination probability.

Evolutionary approaches provide a simple, highly-parallelizable population-based approach to optimization. Since it only requires some measure of performance as a fitness function, the same algorithm can be effectively applied in contexts analogous to supervised, unsupervised or reinforcement learning. Additionally, the fact that these are population based methods enables better exploration of the parameter space in comparison to gradient-based methods and hence have a better shot at escaping local optima. Finally, evolutionary algorithms are only a part of a wide-range of nature inspired population-based stochastic search approaches optimization such as ant-colony optimization (Dorigo and Di Caro, 1999), particle-swarm optimization (Kennedy and Eberhart, 1995), and several others (Beheshti and Shamsuddin, 2013).

In an effort to benefit from the favorable features of both above-mentioned approaches, hybrid approaches of optimization have been developed. One such category of hybrid-approaches are population based gradient-estimation algorithms where performance in a cluster of parameter choices

are compared to estimate the direction of the gradient without actually estimating the derivatives. These estimates have been done using as few as two data points, like in hill-climbing (Russell and Norvig, 2002); a population of randomly sampled data points, like in evolutionary strategies (Beyer and Schwefel, 2002); or when possible, by considering all possible neighboring points, like steepest ascent hill-climbing (Russell and Norvig, 2002). Additionally, hybrid approaches are also employed in optimization methods that involve multiple time-scales. For instance, model parameters may be optimizing using gradient-based approaches, while model hyperparameters may be optimized using evolutionary approaches at a slower time-scale (Young et al., 2015; Loshchilov and Hutter, 2016). Finally, hybrid approaches can involve training different components of a model at different time-scales, which is especially suited when the individual components are mutually-dependent in their learning (Ackley and Littman, 1991). Thus, hybrid approaches to learning span optimization at several levels and provide more sophisticated control over the process.

Altogether, several approaches to model optimization exist and of them, while no single approach is the best (Wolpert and Macready, 1997), some approaches might better suit certain problems. For instance, search based approaches are better suited for discrete parameter spaces since gradients cannot be computed. Similarly, gradient-based approaches may lead to significantly faster convergence in convex landscapes. In any case, a good approach to model optimization involves devising objective functions with known bounds, sufficient random exploration of the parameter space, and conducting several optimization runs to evaluate consistency and variance in results. The work presented in this dissertation involves optimizing neural network models in a way that can explore all possible ways to solve a given problem. To do this, it is crucial to adopt an approach that minimized the bias that an experimenter can introduce. Therefore, evolutionary algorithms have been utilized to optimize models in all work presented in this dissertation.

2.4 Information theory in Neuroscience

Building neural network models and optimizing them to perform tasks gives us models of natural systems. While the resources required to achieve desired task performance is already informative

of the natural system being modeled, these models are most useful when they are analyzed to understand them better. Approaches to understanding neural network operation in the context of behavior requires that it captures linear as well as non-linear interactions between the different components: environment-neural network and neuron-neuron. Several methodologies have been proposed to capture specific features of neural activity such as connectivity, effective dimensionality, encoding, and decoding. In this regard, information theory has emerged as a general framework to quantify stochastic properties and relationships between different variables in a system of interest. It provides tools to measure these quantities in a way that is invariant to the scale of the system and allows comparison across systems. For this reason, the analysis of optimized models in this dissertation has been primarily performed using information theoretic tools. In this section, I describe the main tools used in our analysis as well provide examples of other work within Neuroscience that had utilized information theoretic tools to study the neural basis of adaptive behavior.

Information theory was first introduced by Claude Shannon in his seminal paper "A mathematical theory of communication" (Shannon, 1948), as a methodology to develop efficient coding and communication of data across noisy channels. Its rise to popularity can be primarily attributed to its ability to be applied in any domain, ranging from economics to Neuroscience. It provides a set of tools that enable us to understand the relationships and interactions between arbitrary multivariate random variables. These tools enables us to answer questions such as, "how can we quantify the difference in uncertainty in random process versus another?", and "how much does knowing the value of one random variable, reduce uncertainty about another?". Since Shanon's introduction of information theory, there have been several advancements that allows study of information transfer over time and through the interaction of multiple sources of information. These advancements enable us to ask more involved questions such as, "what is the amount of information transferred from one random process to another over time?", and "Of these multiple sources, what is the amount of information that is redundantly transferred from all sources about a target random variable?". These crucial questions enable us to understand the interactions between different components of

a complex system, and can ultimately lead to a mechanistic understanding of its operation.

2.4.1 Information theoretic measures

This section provides a quick introduction to the information theoretic tools relevant to the analysis conducted in this dissertation. The relevance of each of these measure depends on the data, its source, and the motivation behind the analyses. Refer to Cover and Thomas (2012) for a more detailed account of these measures. More recent papers that introduce information theory to Neuroscientists are McDonnell et al. (2011) and Timme and Lapish (2018).

Entropy

Put simply, entropy can defined as a measure of uncertainty of a random variable. Greater the entropy, the more difficult it is to guess the value the random variable might take. It is function of the probability distribution rather than the actual values taken by the random variable, and it is estimated as follows -

$$H(X) = -\sum_{x \in X} p(x)\ log\ p(x) \tag{2.4}$$

where $H(X)$ denotes entropy of the random variable, X, the summation is over all values the random variable can take, $\forall x \in X$ and $p(x)$ or $p(X = x)$ denotes the probability that the random variable X takes a particular value x. The logarithm is base 2 and so the measured entropy is in the unit of bits. Note that $0log0$ is taken to be 0, and therefore adding new terms to X with probability 0 does not change its entropy.

Intuitively, for a given random variable, a uniform probability distribution over its values would result in the highest entropy since every value is equally likely, making it most difficult to guess. On the other hand, other "peaky" distributions such as Gaussian would have lower entropy since most of the probability mass is near the mean. From a communication systems perspective, where information theory had its origins, entropy can also be interpreted as the average number of bits required to efficiently encode the random variable. For example, a fair coin requires 1 bit (0-heads,

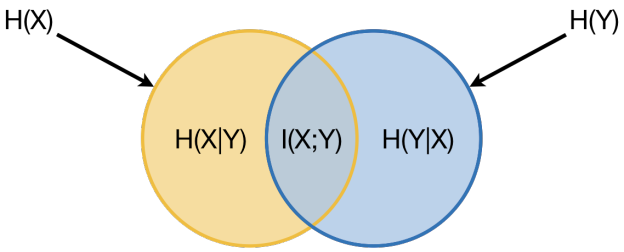

Figure 2.1: Relationship between entropy, conditional entropy and mutual information.

1-tails); a uniform distribution over 8 values of a random variable requires 3 bits, but a non-uniform distribution could possibly be encoded by smaller average number of bits by assigning shorter codes for more likely outcomes.

Mutual Information

One of the most widely used information theoretic measures is mutual information. It is a measure of the amount of information two random variables have about one another - it is a symmetric measure. The information one variable has about the other can be expressed as the reduction in uncertainty about one variable upon knowing the other i.e. the difference between entropy of the variable and the entropy given the other variable.

$$\begin{aligned} I(X;Y) = I(Y;X) = &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{2.5}$$

where $I(X;Y)$ is the mutual information between random variables X and Y, $H(X)$ and $H(Y)$ are their respective entropies, and $H(X|Y)$ and $H(Y|X)$ are their corresponding conditional entropies. The conditional entropy can be estimated from the conditional probability density, which is related to their joint probability density. Upon writing out the entropy expressions about in terms of the marginal and joint densities we arrive at the following expression for mutual information between X and Y

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y)\ log \frac{p(x,y)}{p(x)p(y)} \tag{2.6}$$

The relationship between the individual entropies, the conditional entropies and mutual information can be better understood using a Venn diagram, as shown in figure 2.1.

Specific Information

While mutual information measures the reduction in uncertainty in X, across all values of Y, specific information is a measure of reduction in uncertainty for any given value off $Y = y$. Specific information I_{spec} is defined as follows DeWeese and Meister. (1999)

$$I_{spec}(X; Y = y) = \sum_{x \in X} p(x|y)[log(\frac{1}{p(y)}) - log(\frac{1}{p(y|x)})] \tag{2.7}$$

Based on this formulation, mutual information can be defined as a aggregate of specific information as follows

$$I(X;Y) = \sum_{y} p(Y = y) I_{spec}(X; Y = y) \tag{2.8}$$

Partial Information Decomposition

When there are multiple sources of information (or even a random variable that is 2-dimensional or more) about another random variable, the total mutual information between these sources and the 'target' variable can be decomposed into its non-negative constituents, namely, unique information from each source, redundant information that the sources provide, and synergistic information due to combined information from multiple sources Williams and Beer (2010b). Consider two sources of information (two random variables) $X1$ and $X2$, about the random variable Y. If $X = \{X1, X2\}$, the total mutual information $I(X;Y)$ can be written as follows

$$I(X;Y) = U(X1;Y) + U(X2;Y) + R(X;Y) + S(X;Y) \tag{2.9}$$

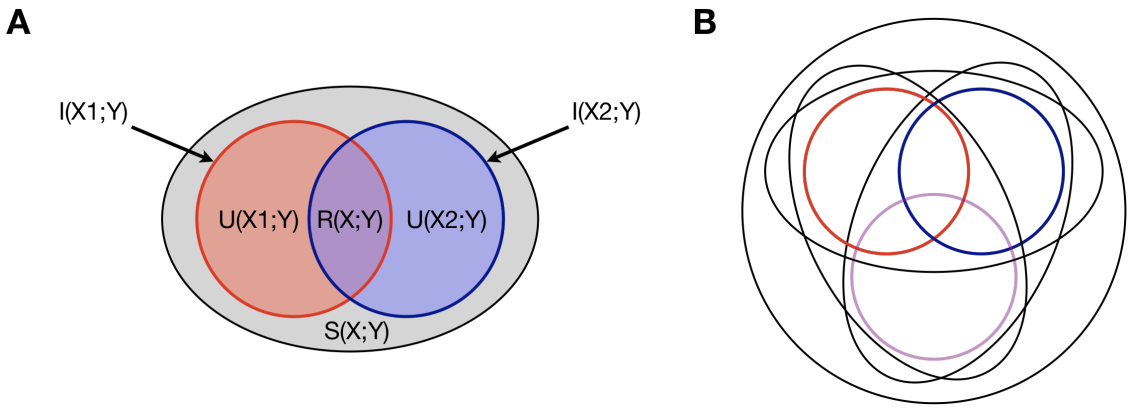

Figure 2.2: Non-negative partial information decomposition of total mutual information. [A] In the case of 3 variables (2 sources $X = [X1, X2]$ and 1 target, Y), each source provides unique information about the target, $U(X1;Y)$ and $U(X2;Y)$, they provide information about the target redundantly, $R(X,Y)$ and finally they provide synergistic information that is only available from the combined knowledge of both sources, $S(X;Y)$.

where U denotes unique information, R denotes redundant information and S denotes synergistic information from these sources about Y. The decomposition can be better understood when visualized as a Venn diagram as shown in figure 2.2. Naturally, with more than two sources, redundant and synergistic information will be available for the all combinations of the different sources. Also, note that each random variable used in these descriptions could be a multi-dimensional. The number of sources are just individual data from the system and can be of any dimensionality.

Redundant Information The sum of the minimum value of specific information each source provides is defined as the redundant information from the two sources. In other words, continuing the $(\{X1, X2\}, Y)$ example, for each value of Y, the specific information $X1$ and $X2$ provide is independently computed, its minimum is found and this values is summed across all values of Y.

$$R(X;Y) = \sum_{y \in Y} p(Y = y) min\{I_{spec}(X1; Y = y), I_{spec}(X2; Y = y)\} \tag{2.10}$$

where specific information is estimated as follows

$$I_{spec}(X1, Y = y) = \sum_{x \in X1} p(X1 = x, Y = y) log \frac{p(X1 = x, Y = y)}{p(X1 = x)p(Y = y)} \tag{2.11}$$

Unique Information The amount of information that each source uniquely contributes, the unique information from that source, is estimated as the difference between the mutual information between that source and the 'target' variable and the redundant information estimated as shown in the previous section. This is also apparent from the Venn diagram representation shown in figure 2.2.

$$\begin{aligned} U(X1;Y) &= I(X1;Y) - R(X;Y) \\ U(X2;Y) &= I(X2;Y) - R(X;Y) \end{aligned} \tag{2.12}$$

Synergistic Information The information about Y that does not come from any of the sources individually, but due to the combined knowledge of multiple sources, is synergistic information. For example, in an XOR operation, knowledge about any one of the inputs is insufficient to predict the output accurately, however, knowing both inputs will determine the output. In contrast, in an AND gate, merely knowing that one of the inputs is $False$ can lead to the conclusion that the output is $False$. Again, based on the Venn diagram in figure 2.2, synergistic information from sources $X = X1, X2$ about Y can be determined based on total, unique and redundant information components as follows

$$S(X;Y) = I(X;Y) - U(X1;Y) - U(X2;Y) - R(X;Y) \tag{2.13}$$

Transfer Entropy

Given two random processes, X and Y, transfer entropy from X to Y, $TE_{X\to Y}$ is defined as the information transferred from from past values of X, X_{t-d} to Y_t, over and above that information that Y's own past, Y_{t-k} provides. Intuitively, transfer entropy is a measure of thee reduction in uncertainty that know the source variable provides over the about the target variable's own reduction in uncertainty.

$$TE_{X \to Y} = I(Y_t; X_{t-d} | Y_{t-k}) \tag{2.14}$$

where k and d are positive values representing possibly different time points in the past depending on the processes. Indeed, they can represent vectors denoting multiple time-points, in order to study the transfer over a range of delays.

Reformulating transfer entropy terms as partial information atoms, Williams and Beer (2011) showed that transfer entropy can be measured as the sum of unique information from X_{t-d} about Y_t that did not come from Y_{t-k}, and the synergistic information that is available due to the combined knowledge of X_{t-d} and Y_{t-k}.

$$TE_{X \to Y} = U(Y_t; X_{t-d}) + S(Y_t; X_{t-d}, Yt - k) \tag{2.15}$$

Estimating informational quantities

One of the major challenges in utilizing information theoretic measures in experimental settings is the availability of sufficient data to infer the data distributions correctly (Paninski, 2003). This is a complex problem and several parametric and non-parametric approaches have been proposed (Silverman, 2018). To estimate data distribution from limited data, we have employed average shifted histograms for its beneficial trade-off between statistical and computational efficiency (Scott, 1985a). This involves discretizing the data space into a number of bins and estimating frequentist probabilities based on the bins occupied by data samples. To reduce the impact of arbitrarily chosen bin boundaries the data distribution is estimated by averaging the bin occupancies across multiple shifted binnings of the data space. This binning based estimator has been shown to approximate a triangle kernel estimator (Scott, 1985a). While the binning provides significant computational advantages, its approximation errors must be considered. Bias properties and guidelines for choosing the parameters for average shifted histograms are given in Scott (1985b); Fernando et al. (2009); Scott (1979, 2012). For a moderate sample size, 5 to 10 shifted histograms has been shown to be adequate (Scott, 1985b). In general, average shifted histograms are best suited for noisy

continuous data where the distribution of the data is unknown. For a more involved discussion on density estimation and its bias properties we point the reader to Scott and Terrell (1987) and Wand and Jones (1994).

Information and time

One variable that is implicit in all formulations of information theoretic measures described above is time. Typically, these information quantities are measured disregarding the time variable. In other words, data distributions are estimated using data across all time-points. Consequently, the information measures are aggregate measures across the time duration over which data was collected. Alternatively, if data was collected in several trials these quantities can be measured *in time*. Information thus measured as a function of time then reveals the dynamics of information in the system under study. This requires that there are a sufficient number of trials to develop reliable estimates of data distributions at each time point. This approach has been applied in several modeling studies, including work presented in this dissertation, as well as in experimental conditions. These studies and the application of information theory in Neuroscience in general is discussed in the next section.

2.4.2 Information theoretic analyses of Neural Networks

The information theoretic principles described above have been used extensively in Neuroscience in several contexts: *in vitro*, *in vivo*, and *in silico* (or computational models). Neural responses are stochastic because of noise and because biological neural networks and their outputs are not purely a function of the input alone, but depend on their internal state as well (Mainen and Sejnowski, 1995). They could produce outputs in the absence of any input, or not produce any response to certain stimuli. This stochastic nature of their behavior makes information theory especially suited for analyzing and interpreting neural activity (Dimitrov et al., 2011; Sayood, 2018; Jung et al., 2014; James et al., 2011; Wibral et al., 2015). In this section I present examples of work at different levels in which information theory has been applied starting from information processing in single

neurons to information processing in neural networks engaged in closed-loop behavior.

Single Neurons

A neuron is typically considered to be the smallest computational element in the animal brain. Although, this is in no way a statement trivializing the computational capacity of the neuron. As revealed by the landmark study of squids' giant axon, and its model (Hodgkin and Huxley, 1952); as well as based on the ever increasing literature on dendritic computation (London and Häusser, 2005), neurons are rather sophisticated information processing systems. Characterizing tuning curves (stimulus-response relationship) has shown that neurons in the visual cortex are selective to orientation of stimulus (Hubel and Wiesel, 1965), place cells in the hippocampus are selective to physical locations in space (O'keefe and Nadel, 1978), cercal neurons in crickets are sensitive to wind direction (Theunissen and Miller, 1991) and so on. Tuning curves were interpreted as neurons being most selective to stimuli that produced to maximum firing rate. However, Butts and Goldman (2006) applied information-theoretic measures to demonstrate the validity of an alternate interpretation - the neuron is most selective to stimuli where the tuning curve has the highest slope. Measuring the specific information that noisy neural activity in medial temporal cortex had about specific values of the stimulus showed that information encoded was maximum near the peak-slope in the tuning curve. Intuitively, this is because there is greatest difference in firing rate for small changes to stimuli at this region of the tuning curve. Further, they showed how the experimental design (range and the number of stimuli provided) can change whether the maximum information lies at the peak firing rate or the highest slope - an idea that is also discussed in this dissertation later.

In addition to identifying the stimulus selectivity of a neuron, information-theoretic analyses have also been extensively applied by viewing neurons as communication channels, and hence estimating their information capacity or channel capacity. This is a measure of the number of distinctly identifiable stimuli that a neuron can encode, or mutual information between the input distribution and output distribution. Automatically, this requires that an assumption is made

about the nature of encoding in order to build the output distribution from spiking activity. Two approaches to encoding that are widely accepted are rate coding and temporal coding (Rieke et al., 1999). Once a specific encoding method has been chosen, the channel capacity will then depend on the noise in the channel, and the distribution of inputs using which the channel capacity will be estimated. Thus, this is a complex problem and even theoretical estimates will strongly depend on the assumptions made. Initial theoretical estimates made using the temporal coding scheme and a Gaussian noise distribution to the inter-spike intervals were made to be 4000 bits/s (MacKay and McCulloch, 1952; Rapoport and Horvath, 1960). This was very optimistic estimate (Stein, 1967) was then re-evaluated using a more biologically realistic gamma distribution for the inter-spike intervals to demonstrate that the capacity was 15-50 bits/s with both the temporal and rate coding schemes (Borst and Theunissen, 1999; Ikeda and Manton, 2009). These results put to rest on which coding scheme was most informative, and the debate now continues in the form of a energy-efficiency (temporal code) versus robustness (rate code). Like most phenomena in nature, animal brains perhaps use both under different conditions.

Populations of Neurons or Brain Regions

Expanding the analysis from individual neurons to groups of neurons, information theory has been applied on two fronts: first, information encoded about stimulus can be measured in a population of neurons similar to how it is measured in a single neuron; and second, by measuring the information that is transferred from one neuron to another.

Several advancements have been made in population coding and decoding approaches that show information being distributed in a population of neurons, and overlapping sensitive ranges across neurons leading to theories of robustness in encoding via population coding. For instance, information about the directionality of motor movement was shown to be best represented as a weighted average of several individual neuron responses (Georgopoulos et al., 1982, 1986; Wessberg et al., 2000). Similarly, saccadic eye movements in monkeys were also shown to be controlled by the weighted sum of the activations of collicular neurons (Lee et al., 1988; Sparks

et al., 1976). While these studies were based on looking at peak-response curve values, an alternative information theoretic approach similar to the one described in the analysis of response curves in single neurons can be adopted. Ince et al. (2010) measured mutual information between the whisker stimuli and the population activity in rat somatosensory cortex for different population sizes. Their results showed that mutual information in populations of size 2 or greater captured the stimulus information very well, without any significant differences in larger populations. This led them to conclude that studying pair-wise interactions in neurons would be sufficient to characterize information transmission in neural networks.

The idea that neural information processing can be understood by studying pair-wise interactions in a population is fundamental to the adoption of network theory in Neuroscience. This involves building interaction networks between neurons in population, as well as between brain regions, based on measuring the interactions between them pair-wise. This approach has flourished in the last decade under the new sub-field of *Network Neuroscience* (Sporns et al., 2005; Bassett and Sporns, 2017; Sporns, 2014). While the literature on how to estimate these interaction networks is continuously growing (Okatan et al., 2005; Hlaváčková-Schindler et al., 2007; Pillow et al., 2008; Gerhard et al., 2011), Transfer entropy (TE) has emerged as one of the most widely used methods in Neuroscience because of its ability to infer directed edges between nodes in a network from node activity alone (Wibral et al., 2014a). As mentioned earlier, these nodes could be anything from individual neurons or macro brain regions. When estimated across all possible pair-wise node combinations, the *effective network* (Friston, 1994) is constructed. The effective network is representative of the interaction between neurons that results in the dynamics on the underlying structurally connected network.

Studying properties of the effective network has yielded several insights on neural information processing, especially at the micro-scale. Multi-electrode array recordings of neural activity of organotypic cortical cultures *in vitro* has shown the presence of rich-club architectures; some highly-connected neurons are more connected to each other than by chance (M. Shimono, 2015; Nigam et al., 2016). This study further showed that information transfer in a neural network is

non-uniform: 70% of the information was propagated through only 20% of the neurons. Following that, Timme et al. (2016) measured the synergistic information in a neuron's activity from the combined sources it receives information from, as an indicator of how much computation was performed by the neuron. Intuitively, this is a measure of how much more information is available in the neuron beyond just the union of information provided by the inputs. This analysis showed that computation in a neuron was more strongly correlated to the number of outgoing connections rather than the inputs it received. Furthermore, it was then shown by cross-referencing effective network properties with synergistic information that it was, in fact, the rich-club neurons that performed a disproportionate amount of computation, about 160% more (Faber et al., 2019).

In larger-scale neural dynamics, where ensemble neural dynamics corresponding to brain regions comprised of several thousand neurons are measured using fMRI, transfer entropy based effective networks can be constructed with nodes denoting brain regions. Such studies have yielded several insights as well into information transfer and its dynamics across the brain. For instance, Joseph T. Lizier and Prokopenko. (2011) estimated effective networks while subject performed a visuo-motor tracking task to show that TE was able to infer the expected directed connectivity between known motor planning regions in the brain. They also showed that increasing task-difficulty resulted in increased amounts of information transfer between the motor cortex and the cerebellum where fine-tuning of motor control occurs indicating the greater error correction happening as the task is more unpredictable. Similarly, by applying TE to magnetoencephalographic data collected during an auditory short-term memory experiment, Wibral et al. (2011) showed that TE was able to capture the expected changes in information transmission left temporal pole and the cerebellum.

Altogether, the integration of network theory with information-theoretic measures is enabling the interpretation of neural activity to go beyond response curves in single neurons to understanding collective information processing in populations of neurons.

Neural Networks engaged in closed-loop behavior

The neural network analyses described in the two previous sections are mostly using resting-state data where the neural network was exhibiting its intrinsic dynamics, or when a subject was passively performing a task. In the real-world animals are in closed-loop interaction with their environments where they act on the environment thereby influencing what they perceive. Perhaps the best examples of information theoretic analysis of neural networks that are involved in closed-loop behavior come from computational modelling studies. Information-theoretic analysis of embodied neural network models that were optimized to perform cognitively interesting tasks such as object categorization, relational categorization etc. have demonstrated that studying information flow through the integrated brain-body-environment system will provide a comprehensive understanding of the neural basis of behaviors (Williams et al., 2008; Williams and Beer, 2010a; Beer and Williams, 2015; Izquierdo et al., 2015a). Such analysis also provide additional insights into how behavior is not entirely controlled by the brain because this allows for considering the environmental and body variables into the information theoretical analysis. For instance, using a model of embodied relational categorization Williams et al. (2008) showed that the optimized agents offloaded information storage to the environment. Agent's were optimized to behave differently in response to the relational category of "smaller" or "larger" based on a visual stimuli of objects falling in sequence. Amongst agents that solved this problem with similar levels of accuracy, some adopted a strategy of moving away from the falling object far enough such that their position was indicative of the size of the object. This behavioral characteristic was made mechanistic explanation by demonstrating the near-perfect amounts of information in the agent's position about the size of the perceived object. In a more biologically grounded model of *C. elegans* klinotaxis, (Izquierdo et al., 2015b) showed that although the underlying neural network parameters might be very different in different model instantiations, they were all identical in the pattern of information flow through the neural network. This was demonstrated by measuring time-varying mutual information in the neural activity about the behaviorally relevant variable, namely change in salt concentration to show that the ensemble of models that were analyzed all had similar information flow patterns. Thus, such an analysis

allows us to go beyond the individual structural differences, and study the common functional characteristics of neural networks thereby enabling us to interpret neural activity in the context of behavior. Altogether, information theoretic analyses of brain-body-environment models takes us one step closer to a comprehensive understanding of the neural basis of adaptive behavior.

2.4.3 Tools for information theoretic analyses

Several researchers have developed software packages that aid information theoretic analyses. Pyentropy is a python package that allows easy estimation of entropies (Ince et al., 2009). JDIT is a JAVA package that was primarily designed for measuring transfer entropy but also includes other measures such as entropies and mutual information (Lizier, 2014). TRENTOOL is a popular MATLAB toolbox which primarily caters to analog neural data such as MEG (Lindner et al., 2011). Ito et al. (2011b) presented a MATLAB toolbox for TE estimation from binary spiking data that accounted for several delays and picked the most reliable transfer entropy value using a data-shuffled baseline. In addition to this MATLAB package, work presented in this dissertation uses, *infotheory*, a packaged developed in-house for PID measures as well as transfer entropy estimation (Candadai and Izquierdo, 2019a). Two existing packages that are most similar to ours are dit (James et al., 2018) and IDTxL (Wollstadt et al., 2019). Unlike dit, our package can also help analyze continuous-valued data and unlike dit and IDTxL we have implemented PID analysis of 4 variables: 3 sources and 1 target. In light of these existing packages and their functionalities, our package primarily focuses on measuring multivariate informational quantities on continuous data where the data distribution is not known a priori. However, it can still be used with discrete data using the same methods.

2.5 Our Approach

This chapter discusses the use of computational models and information theory to advance our understanding of the neural basis of behaviors: a cornerstone in the development of the ideas in this dissertation. Computational models provide the ability to study our biological systems of interest

at a level of abstraction that enables analyses in a way that is not tractable in biological systems. They provide complete access to all variables of a system and, more so, the ability to manipulate them in a way that experimentalists cannot. Information theory, on the other hand, provides the tools for an in-depth analysis of these models. Insights obtained from analysis of the model advance our theoretical understanding of neural information processing. It provides proof-of-existence, and general principles underlying neural information processing in adaptive behavior. Additionally, information-theoretic tools can acts as the bridge between modeling and experimentation. Analyses performed on an idealized model can be directly applied to appropriate experimental data; with the confidence that model-based evaluation of the methods have demonstrated the efficacy of the method to capture the phenomena of interest. Ultimately, the combination of computational modeling and information theory is a potent combination that can advance theoretical as well as experimental Neuroscience for years to come.

The goal of this dissertation is to conduct information-theoretic analyses of computational models of neural networks to advance our understanding of the neural basis of behavior. Consequently, it becomes crucial to identify the level of abstraction at which we build models to study the neural basis of adaptive behaviors. Several perspectives exist in this regard, and perhaps the most widely adopted is from mainstream theoretical Neuroscience where models are built primarily at one of two levels: sensory-response models to describe what the input-output relationship is between stimuli and neural activity; and circuit-level models that describe how these relationships maybe implemented using known physiological data (Abbott, 2008). However, these approaches limit cognition to only the brain. There is an increasing realization that the role played by the environment and the body in behavior needs to be take seriously in order to develop a comprehensive understanding of the neural basis of behavior (Krakauer et al., 2017; Chiel and Beer, 1997; F. J. Varela and Rosch, 1991). This is inline with approaches in Computational Neuroethology where models are built not only of neural circuitry, but also of the body, and ecological context of the model system, all of which are considered to be equally important in behavior (Datta et al., 2019; Chiel and Beer, 2008; R. Pfeifer and Iida, 2007). The work presented in this dissertation adopts this perspective by

considering the environment and environmental interaction to be part of the cognitive system and thus including it in our model and its analysis.

There have been two prominent approaches to building models from a neuroethological perspective: bio-robotics and evolutionary robotics. Bio-robotics, as the name suggests, involves building robots as models of specific animal systems to test and generate hypotheses regarding the control of behavior (Webb, 2002, 2001; Beer and Ritzmann, 1993). Building robots ensures that the physical aspects of an animal's ecological context are faithfully reproduced. Further, robots that are models of specific animals enable targeted understanding of behavior in that animal. For example, Möller et al. (1998) built a robotic model of desert ant navigation to demonstrate that it was a combination of path-integration and visual piloting that enabled these ants to navigate to precise locations over large distances. Similarly, Ijspeert et al. (2007) built a robotic salamander to demonstrate that the same neural oscillators when modulated by environmental feedback can produce the walking and swimming gaits observed in salamanders. These studies not only act as existence-proofs for mechanisms that drive specific behaviors in a target mechanism, but also provide specific testable hypotheses for the animals that the robots are based on. The evolutionary robotics approach, on the other hand, involves building simulated idealistic models of artificial agents with emphasis on building minimal models of cognition that enable in-depth analysis (Harvey et al., 2005; Cliff et al., 1993; Beer and Gallagher, 1992). Integral to evolutionary robotics is use of evolutionary algorithms to optimize model parameters, an approach that minimizes experimenter bias by optimizing neural networks based on *what* they are optimized to do and not *how* the experimenter thinks it should be done. Crucially, in contrast to bio-robotics models, these models are typically not built to physiological data from a specific animal. This is done because freeing the constraint of adherence to a specific model organism enables the construction of minimal artificial agents that nevertheless performing cognitively interesting tasks are tractable for in-depth analysis, and exploring general principles in behavior that go beyond generating hypotheses for specific animals (Beer and Williams, 2009). For instance, using a minimal model of a 1-dimensional agent controlled by a dynamical neural network, Beer and Williams (2015) show that embedded, embodied agents perform a task that

requires memory without an explicit internal storage mechanism; the agent offloaded memory to its position in the environment: a testament to the benefits of adopting a neuroethology perspective. Similarly, Izquierdo and Bührmann (2008) showed that dynamical neural networks can perform multiple tasks even in the absence of neuromodulation or plasticity. Overall, building models that incorporate the environment and agent-environment interaction expand the range of mechanisms that can be discovered to produce behaviors, and utilizing evolutionary algorithms further aid this by enabling the generation of integrated sensorimotor systems with minimal bias. Computational models are especially suited when this approach is taken to research because, although there have been developments in tools for experimental data acquisition, there are still challenges in reliably recording all required data from a freely moving animal. In such cases, theoretical advancements must not wait for experimental technologies to arrive but can instead precede them by advancing our understanding of general principles of adaptive behavior through the analysis of computational models. The work presented in this dissertation involves evolving neural controllers in embedded, embodied and dynamical neural networks to advance our understanding of the neural basis of adaptive behavior.

In summary, the work presented in this dissertation takes the approach of developing computational models of dynamical recurrent neural networks, optimizing them using an evolutionary algorithm to perform tasks that involve embedding or embodying the neural network models, and followed by information-theoretic analysis of the optimized models. In all cases, an ensemble of models were built and analyzed to check if the results are consistent or if several solutions exist that solve the same problem. Specifically, this approach is applied to study predictive coding, multifunctionality and social interaction.

Chapter 3

Predictive Coding

In this chapter, I present our work on inferring the source of predictive information in dynamical neural networks using tools from multivariate information theory, as well as our results from studying the relationship between predictive information and behavioral performance. In the first section of this chapter, I introduce the idea of predictive coding and motivate our work. Following that, I discuss literature from the two research fronts in predictive coding, and the potential for their convergence. Next, the neural network and environment models used in our study are described in detail in the methods section. This is followed by the principal contribution of this work: a framework to identify the source of predictive information and track its dynamics in time. Then, results validating our approach and its application to a structured environment are presented. Finally, I discuss the results and their implications.

3.1 Introduction

The brain is in continuous interaction with the body and in turn with the environment. In this closed-loop agent-environment setup that living organisms are embedded in, it is a challenging research problem to understand how behavior emerges from the interaction of the brain, body and environment. Although not explicitly, experimentalists make assumptions about the roles played by the brain, body and environment implicitly in how they study the brain – e.g. studying the brain alone in isolation is sufficient to understand cognition. These assumptions are made explicit in theoretical neuroscience where the goal is to develop general hypotheses that are consistent with

known neuroscientific principles and physical laws, to answer questions such as: How does the brain process sensory stimuli? How does sensory processing take advantage of the regularities in the environment? How are actions taken based on perception? Several theories have been proposed as to the functional processes ongoing in the brain to produce motor actions based on sensory input. In this work we focus on two complementary theories that fall within the domain of *predictive coding*, a research area that is emerging as a strong candidate for its potential to provide a general framework for understanding the neural basis of behavior (Clark, 2013).

Predictive coding is the idea is that organisms encode information about future environmental stimuli in their neural activity (Helmholtz, 1860; Huang and Rao, 2011; Rao and Ballard, 1999; Srinivasan et al., 1982). Intuitively, an organism that is able to predict the consequences of its action on its future sensory experiences is more likely to be adapted to its environment. There are two prominent research fronts that study the role of predictive coding in behavior: the hierarchical generative predictive processing framework (Friston and Kiebel, 2009; Friston, 2010) and the efficient coding principle (Bialek et al., 2001; Still, 2009). These two fronts are complementary because they address different aspects of how a nervous system acquires predictive information. The hierarchical predictive processing framework focuses on how predictions are generated in the organism's brain. The efficient coding principle, on the other hand, focuses on how the nervous system extracts predictive information from environmental stimuli. Both theories have been supported by experimental evidence, primarily in the visual and auditory systems (Palmer et al., 2015; Chen et al., 2017; Sederberg et al., 2018; Chao et al., 2018; Egner and Summerfield, 2013).

In living organisms, predictive information is likely acquired from a dynamically changing contribution of the environment and the agent's own internal dynamics. Consequently, although different systems may be equally predictive about their future stimuli, the operation of their nervous systems may be entirely different. Therefore, understanding the role of predictive information in behavior requires that the source of information is identified. In this chapter, the following questions are addressed: How do we identify the source of predictive information and study its dynamics

during a behavior? Does tracking the source of predictive information better explain an agent's ability to perform a task? What are the factors that influence the source and magnitude of predictive information encoded in a neural network?

In order to better understand how the nervous system generates predictive information, we propose that it is essential to decompose information transfer across the different components of the system and to track the flow of information in the agent-environment system over time. We present an information-theoretic framework to quantify the contributions from the nervous system and the contributions from the environmental stimuli to the total predictive information in an agent. To do this, we first decompose the total predictive information in the neural system into information that was uniquely transferred from each source using multivariate extensions to information theory (Williams and Beer, 2010b). Second, we unroll information over time to backtrack the origin of the source of predictive information and how it changes over time. This allows us to infer the source of predictive information and its dynamics over the course of a behavior.

To validate our proposed theoretical framework, we examine it on a set of computational models of agent-environment systems, where the agent is a dynamical recurrent neural network (Funahashi and Nakamura, 1993; Beer, 1995c). The systems have been deliberately designed so that the source of predictive information is known and manipulable. We demonstrate that predictive information, as operationalized by Bialek and Tishby (1999) cannot distinguish systems that are at the two extremes of potential agent-environment interaction: a system whose only source of predictive information is the nervous system and a system whose only source of predictive information is the environmental stimuli. We demonstrate how our proposed framework correctly reveals different sources of predictive information in systems with otherwise similar amounts of predictive information. Ultimately, we demonstrate how revealing the flow of information across the agent-environment system can help us to better understand the mechanisms underlying predictive coding.

Predictive information is studied in living organisms because it is considered a signature of their adaptive capacities (Palmer et al., 2015; Still, 2009; Friston and Kiebel, 2009). In this regard, we study the relationship between a system's ability to perform a task and its predictive information.

In order to do this, we turn to a computational model of an agent that is required to process the received stimulus from the environment and make a decision based on it. Specifically, we study predictive information in the context of a relational categorization task (Gentner and Kurtz, 2005; Markman and Stilwell, 2001). We generate model systems that are adapted to their environment and yet remain tractable to analysis by optimizing dynamical recurrent neural networks using an evolutionary algorithm to perform the task (Beer and Gallagher, 1992; Floreano et al., 2008). We then proceed to analyze the resulting systems using predictive information and we compare the results against that of random systems that cannot solve the task. Counterintuitively, we observe that predictive information in trained neural networks is similar to predictive information in random neural networks. This suggests that predictive information alone is not sufficient to distinguish between living organisms that are adapted to their environments and non-adaptive systems. The rest of the paper focuses on an analysis of optimized and random systems using the proposed framework. Altogether, we demonstrate that decomposing predictive information across the components of an agent-environment system, and unrolling it over time reveals its true nature.

3.2 Related work

The idea that prediction forms an important part of cognition goes all the way back to Helmholtz's (1860) theory of perception where he argues that the brain learns to represent the causes of stimuli in the environment thereby enabling it to link any incoming stimuli to existing internal models of environments. This idea has inspired research in Neuroscience and machine learning. Importantly, research in this domain has followed two independent paths: first, research that aims to flesh out Helmholtz's ideas about the encoding of causes of stimuli and thereby generating predictions internally; and second, research that tackles prediction as a broader problem without necessarily subscribing to the specific mechanisms that the former approach proposes. This section will outline relevant work along each of these research directions and discuss paths that can bridge them.

3.2.1 Predictive coding as a hierarchical generative process

Helmholtz's theory inspired work that further develop ideas of top-down predictive mechanisms: predictions are *generated* from internal models of causes of stimuli to explain away stimuli from lower-levels. This involves a hierarchical architecture where higher levels (deeper layers of the brain) generate predictions for the input to levels below them based on an internal model of the world, and in turn receive errors in predictions to update their internal model. The first evidence for the feasibility of such a top-down predictive system came from advances in machine learning. The Helmholtz machine was a multi-layered bi-directional artificial neural network that was able to learn the underlying causes in a binary image set (Dayan et al., 1995; Hinton et al., 1995). The Helmholtz machine was setup using two sets of weights: one set of generative weights that generate predictions for lower-layers and one set of recognition weights that feed activity from the lower-level to the higher levels leading to the recognition of the cause of that particular stimulus. While the upstream and downstream information flow was not happening concurrently in the model, the Helmholtz machine was a milestone in demonstrating that such bidirectional information flows based on generative models can perform interesting tasks such as binary image classification tasks.

From a neuroscientific perspective, the idea of successive layers in the brain generating predictions to the lower-level layers is counter to traditional understanding of bottom-up processing of stimuli, for example, as shown by progressive detection of bright spots followed by edges and then shapes in visual processing (Hubel and Wiesel, 1965; Riesenhuber and Poggio, 2000). This reversal in the flow of information was studied in visual perception by demonstrating that a hierarchical model of the visual cortex where each level in this hierarchical model generated predictions of the activity in the lower levels and in turn received residual errors in predictions (Rao and Ballard, 1999; Jehee and Ballard, 2009; Huang and Rao, 2011). This model was shown to reproduce the receptive fields of neurons in the visual cortex, and also better capture extra-classical effects to the receptive fields such as endstopping (Bolz and Gilbert, 1986; Desimone and Schein, 1987). This work was extended by Lee and Mumford (2003) to incorporate probabilistic generative models at each level of the hierarchy thereby placing these approaches within the Bayesian framework (Doya

et al., 2007) and was further explored by Friston (2005). Interestingly, work involving hierarchical probabilistic generative models of predictive coding has predominantly focused on visual perception, perhaps because the hierarchical structure of the visual cortex fits these theories the best.

Moving beyond visual perception, the idea of predictions and prediction error has also been applied to rewards in behavior (Schultz et al., 1997), specifically in the anterior cingulate cortex (ACC) (Alexander and Brown, 2019; Brown and Braver, 2005). The Predicted Response-Outcome(PRO) model attributed the function of the ACC to be an estimator of surprise in rewards without considering its valence i.e. unexpected reward or positive surprise versus absence of expected reward or negative surprise (Alexander and Brown, 2010, 2011). Abandoning the valence of the surprise and only considering the prediction error puts this model in line with previous work in visual cortex, thereby acting as an account of predictive coding in the cortex to perform higher-level cognitive processes beyond perception. Further, to generalize predictive coding across cortical regions, Bastos et al. (2012) proposed canonical microcircuits based on the idea that a cortical column has the necessary computational primitives for hierarchical predictive coding. If each cortical column is considered one level of a hierarchy, predictions from higher-levels are received in layers 2/3, and predictions for lower-levels are made and passed down from layer 5. Layer 4 receives prediction error from the lower-layers and layers 2/3 returns the prediction error for this layer back to the higher-level layer.

Predictive coding as a theory for perception, and its generalization into the Bayesian framework has culminated in a "grand unified theory of the brain": the Free-energy Principle (Friston, 2009, 2010). The free-energy principle aims to provide a general theory for the neural basis of behavior by providing a theoretical basis for perception, and action (Hohwy, 2013). While perception is modeled as a hierarchical generative predictive model, the role of action is proposed to be essentially a prediction-fulfilling mechanism: actions are taken so as to minimize prediction error by confirming to the causes of stimuli as expected by the internal generative model (Brown et al., 2011; Buckley et al., 2008). The perception and action mechanisms are optimized so as to minimize

prediction-error, or otherwise denoted by surprise or free-energy. Recently, Friston et al. (2009) and Baltieri and Buckley (2017) showed that such an optimization process can lead to behaviors similar to those learned using traditional reinforcement learning approaches using the Mountain Car and Phototaxis tasks respectively, albeit in a non-neural model.

Ultimately, the hierarchical generative predictive model provides a framework where perceptions are mediated by predictions generated internally in the brain and by errors in those prediction. Essentially, these internally generated predictions are proxies to the sensory stimuli and the only information extracted from the actual stimuli are errors in prediction. The idea of hierarchical predictive processing can be applied to all levels of neural processing and has the potential to provide an elegant general theory of the brain. While the ideas presented retroactively explained several experimental results (Clark, 2013; Egner and Summerfield, 2013), there is still no conclusive test of the specific ideas themselves. For instance, studies that involve analysis of event-related potentials in EEG data have shown the presence of behavioral error related activity in decision making tasks in the pre-frontal cortex (Gehring et al., 1993); in motor tasks in the limbic regions (Gemba et al., 1986); and in language, where reading a sentence that deviated from the "preferred" structure elicited a distinct brain potential presumably because it encodes surprise (Osterhout and Holcomb, 1992). Additionally, functional neuroimaging studies have shown increased activation in the visual cortex corresponding to surprising visual stimuli presumably due to prediction errors (Alink et al., 2010; Den Ouden et al., 2009; Egner et al., 2010); increased activation in the medial prefrontal cortex corresponding to difference between expected and actual outcome of a gambling task (Jessup et al., 2010). Finally, the effect of reduced activation after repeated stimuli presentation due to reduction in prediction error (Friston, 2005) has been shown in functional Magnetic Resonance Imaging (fMRI) (Summerfield et al., 2008), Electroencephalographic (EEG) (Summerfield et al., 2011) as well as in Magnetoencephalographic (MEG) (Todorovic et al., 2011) experiments. While all these results demonstrate effects that can be attributed to surprise or prediction error, some of them (especially the perception related studies) could also be attributed to other mechanisms such as attention. More specifically, a direct test of the idea that sub-populations in the same region

of the brain simultaneously encode expectation as well as error would significantly strengthen the position of generative predictive coding as a theory of cortical function. Perhaps, the strongest support in this regard was work done with high-density electrocorticography in monkeys by Chao et al. (2018). In this work, they provided auditory stimuli that had correlations at two levels of hierarchy and found prediction and prediction error signals in the α/β and γ bands of auditory cortical activity respectively. However, experimental results in this domain involve attributing observations of increased or decreased activations to predictive coding. A significant methodological contribution in this regard would be an approach to disentangle detection of attention from prediction or prediction error.

3.2.2 Predictive coding as efficient information encoding

An alternative branch of research whose roots can also be traced back to Helmholtz provides a complementary account of predictive coding. The distinction being, while living organisms do encode predictive information it is *extracted* from the stimulus as opposed being generated internally in the brain (Bialek et al., 2006). There is correlation in the natural world and it has been shown that nervous systems are sensitive to spatial and temporal regularities in the stimulus (Montague and Sejnowski, 1994; Srinivasan et al., 1982; Huang and Rao, 2011). Based on this, Bialek et al. (2001) argued that the most important features from sensory stimuli are those that are maximally predictive of the future. In this framework, perception can then be formulated as efficiently separating the predictive features from non-predictive noise in the environmental stimuli (Bialek et al., 2006).

The mathematical slant of the research in this domain, largely due to its origins in the physical sciences, has led to proposal of methods for quantification of predictive information. It is defined as the mutual information between the current neural activity and future stimulus, $I_{pred} = I(S_{t'}, N_t)$ (Bialek and Tishby, 1999; Rieke et al., 1999; Cover and Thomas, 2012; Shannon, 1948).

$$I_{pred} = I(S_{t'}, N_t) = \sum_{s_{t'}, n_t} P_N(n_t) P(s_{t'}|n_t) \log_2 \frac{P(s_{t'}|n_t)}{P_S(s_{t'})} \tag{3.1}$$

where $t' = t + \delta t$ with $\delta t > 0$, P_S is the distribution of environmental stimuli, P_N is the distribution of neural activity across the entire experiment, $P(s_{t'}|n_t)$ is the conditional probability that the stimulus is s at a future time t' given that we have observed a neural activity of n at time t. Optimization of simulated agents with neural controllers to maximize predictive information have demonstrated the viability of this theory by producing interesting exploratory behaviors (Ay et al., 2008) and multi-jointed locomotion (Martius et al., 2013). However, these behaviors have been limited to behaviors that could broadly be classified as entropy maximizing behaviors and demonstrations of more cognitively sophisticated behaviors such as decision making has been limited to theoretical derivations of optimal non-neural behavioral policies from assumed general prior distributions that make up an environment (Still, 2009).

Experimental evidence for the efficient extraction of predictive information has been provided using *in vivo* experiments where the amount of predictive information encoded in neural activity in response to stimuli was estimated. By measuring neural activity in the retina of a clamped salamander while moving bars of light were shown, Palmer et al. (2015) showed that neural activity encoded near-optimal information when compared to the analytically estimated information bounds. Moreover, they report identification of retinal neurons that are selective to reversal in direction of the stimulus' movement. While the authors attribute this as extraction of information about this change in direction to then be able to predict the future states of the stimuli, proponents of the generative predictive coding hypothesis might interpret this as prediction error. Chen et al. (2017) confirmed this study by demonstrating similar results in bullfrog retina. Interestingly, experimental work in this domain is limited to visual processing as well. Additional support for this theory would require the demonstration of information about higher order environmental variables in deeper cortical layers. Perhaps the best efforts in this regard come from Sederberg et al. (2018) who used data recorded in salamanders to train an artificial neuron to learn an efficient encoding of neural activity in the retina based on correlations in their spike trains. This computation model demonstrates that training a binary neuron using spike timing–dependent learning rules enables a "deeper" layer to perform near optimal encoding of information available in the correlations of its

lower layer thereby providing a proof-of-concept for efficient coding to be applicable at every level of cortical processing.

3.2.3 Potential for convergence

Do hierarchical generative predictive coding theory and the efficient coding principle have a chance to converge into a single theory? They are similar in the core idea that agents encode information about the future states of the environment. However, the generative predictive coding theory and the efficient coding principle require reconciliation of the following points: First, while the former states that the predictions are purely made in the brain, the latter states that predictive information is extracted from the stimuli; and second, while the former reduces perception to the transmission of prediction error up the hierarchy, the latter proposes an optimal information extraction at each layer. Each of these theories have their individual strengths: the generative predictive coding theory is strongly rooted in experimental neuroscientific findings and the efficient coding hypothesis is grounded in mathematical foundations. As a result, the former provides a clear biologically plausible hypothesis and the latter provides methodologies for effectively measuring and testing the proposed phenomenon.

The two theories are offering fundamentally different levels of explanation as defined by Marr (Marr, 1980; Dawson, 1998): generative predictive coding at the level of implementation and efficient coding principle at the computational level. So they cannot be directly compared. If we borrow the computational aspects in the generative predictive coding hypothesis, we have the idea that predictions are internally generated in the brain, and an internal model is constantly updated based on errors in prediction. If we relax the idea that prediction errors are the only signals that formulate perception, we can also accommodate the efficient coding principle as perception involving the extraction of predictive information from the stimulus. This composition of theories would mean that, agents make predictions about the future using a combination of internally generated and externally extracted information. In fact, extensions to the efficient coding principle have led to theories that present adaptive behavior as learning to optimizing the trade-off

between maximizing predictive power while minimizing complexity of the internal model (Still, 2009; Bialek, 2012; Still and Precup, 2012). Importantly, this framework encourages the agent to acquire an internal (explicit or implicit) predictive model that enables maximization of predictive information. Thus, it allows for predictive information to be generated internally by the agent while at the same time extracting as much as possible from the stimulus. Therefore, with some relaxation of the specific architectural prescriptions of the generative hierarchical predictive coding hypothesis, this framework fits both theories. Along these lines, although not intended as a convergence of these theories, based on predictive information formulated in equation 3.1, Still (2014) proposed a constrained optimization problem for the combined perception-action systems as follows:

$$max\Big(I(S_{t'}, N_t) - \lambda I(S_{t_{past}}, N_t)\Big) \tag{3.2}$$

where $I(S_{t'}, N_t)$ is the predictive information that is maximized constrained by the complexity of the agent's internal model, $I(S_{t_{past}}, N_t)$, and λ is the Lagrange multiplier that regulates the trade-off between the two (Still and Crutchfield, 2007; Still et al., 2010). Notably, this formulation is a variation of the information bottleneck paradigm for supervised learning (Tishby et al., 2000). There are two reasons why adopting this common denominator might be useful: first, this does not require any specific neuroanatomical implementation and is open to discovering how the cortical architecture might implement these processes (the aspects of generative predictive coding that do not have experimental evidence); and second, this framework provides a rigorous mathematical framework within which experimental data can be analyzed to obtain reliable insights.

The bottom-line of this discussion about the potential for convergence of these two theories is the idea that living organisms can generate predictive information internally while at the same time extracting predictive information available from regularities in the environment. However, predictive information given by equation 3.1 cannot differentiate between the two. The rest of this chapter presents our framework for analyzing predictive information in agent-environment systems that allows us to distinguish the contribution of each source, namely internally by the agent and externally from the environment (Candadai and Izquierdo, 2019b).

3.2.4 Measuring information encoding in neural networks

Information theory provides a general theoretical framework to study the relationship between complex systems from observed data in a scale invariant manner (Shannon, 1948). In Neuroscience, information theory has a long history of being applied to study various aspects of neural computation. For instance, mutual information has been used widely in the study of neural coding (Rieke et al., 1999; Dayan and Abbott, 2001). Mutual information enables us to quantify the amount of information that the neural network has about another variable of interest such as an environmental feature, muscle or body states, or even activity of other neurons. However, mutual information is a bivariate symmetric measure while most interactions in living systems are multivariate. Furthermore, the informational structure in living systems changes in time. These characteristics have led to extensions to Shannon information in two ways: multivariate information metrics, and time-dependent information measures.

Recent extensions to mutual information termed Partial Information Decomposition (PID) has provided methods to decompose multivariate information into its component terms that include the individual contributions of each variable, and the different combinations between them (Williams and Beer, 2010b; Bertschinger et al., 2014; Griffith and Koch, 2014). Briefly, the total information that a neural network has about, say 3 variables, is measured using mutual information between the neural network activity and a random variable that is a concatenation of the 3 variables of interest. This total information can be decomposed into information that was uniquely contributed by each variable, the redundant contributions by different combinations of the 3 variables, and finally the synergistic contributions of different combinations of the 3 variables. When applied to the study of multivariate animal-environment systems PID allows us to study the interactions between its different components and assign credit appropriately to the contribution of the different components, a feature that is utilized in the current study of predictive information.

Temporal dependencies between components of a complex system can be studied using Shannon information measures as well its extensions (PID) by studying relationships between data at different time points. Such a formulation had led to the development of two kinds of information measures:

time-delayed measures and measures of information flow. One example of time-delayed measure that was described previously is predictive coding; mutual information between current neural activity and the future stimulus (Bialek et al., 2001). Similarly, mutual information between time-delayed states of the same neural network conditioned on another variable's state, enables measuring the transfer of information from the other variable to the neural network, also known as transfer entropy (TE) (Schreiber, 2000a; Wibral et al., 2014a). Measuring TE between neurons in a network yields a "functional network" that captures the directed informational interaction between neurons in a network. Importantly, while these measures are time-dependent because they measure associations between variables delayed in time, the measures are still atemporal in the sense that they still aggregate information across all data-points separated time by the specified delay. On the other hand, study of information flow involves considering the variables of interest as random processes rather than random variables (Williams and Beer, 2010a). Thus, neural activity at a specific time point is a random variable whose information components can be measured with the state of the environment at a different time point or even that of the same network at a different time point. This is then repeated for each time point to obtain the temporal change in the measured information quantity over the time course of the data. Thus, all information theoretic measures could be estimated in a temporal manner to study the dynamics of information in a system, another feature that is utilized in the current study of predictive information.

Analysis of computational models of embodied neural networks optimized to perform cognitively interesting tasks have demonstrated the explanatory power that extensions to Shannon information have to provide a mechanistic explanation of neural dynamics as it is related to a behavior. Williams and Beer (2010a) demonstrate that tracking information in a neural network over the course of a relational categorization task delineates the role of each neuron in the behavior as well as the temporal changes to their role over the course of the behavior. They further expanded on this work by measuring the flow of decomposed information in the same relational categorization task, and demonstrated that such a framework provides an analogue to a dynamical systems analyses of the behavior (Beer and Williams, 2015). This work provides the crucial insight that

these extensions to Shannon information can provide a scalable alternative to dynamical systems theoretical analyses of neural networks which is known to provide a complete characterization of an agent-environment system (Beer, 1995a; Beer et al., 1996; Beer, 2000). The approaches to measure multivariate information and over time are increasingly adopted to study information processing in biological neural networks. Izquierdo et al. (2015a) built a computational model of klinotaxis grounded in the known biophysical properties of the nematode *C. elegans* and tracked the flow of information flow through a complete sensorimotor circuit: from stimulus, to sensory neurons, to interneurons, to motor neurons, to muscles. This analysis identified key neurons that act as integrators of information, and demonstrated possible hypotheses for the degree to which gap junctions and sensory neurons play a role in the behavior. Utilizing Williams and Beer's (2011) measures of decomposed information transfer, Timme et al. (2016) demonstrated that maximal transfer of unique information in *in vivo* organotypic cultures was occurring in neurons that receive connections from high out-degree neurons. Wibral et al. (2017) showed that different informational objective that have been attributed to the neo-cortex such as predictive coding, infomax and coherent infomax, efficient coding can be formulated using PID terms. Furthermore, they propose an objective function called "coding with synergy", maximizing synergy between input and output, as a global objective across layers of the neocortex. in related work, Faber et al. (2019) measured synergistic information as a proxy for computation in neural networks to demonstrate that neurons in an *in vivo* organotypic cortical network that are disproportionately connected to other high-degree neurons perform most of the computation. In the work presented in this chapter, we utilize PID measures in-time to measure the unique, and redundant transfer of predictive information from the environment and the neural network to identify the contributions of the different source of predictive information.

3.3 Methods

In the agent-environment models used throughout this work, the agents were modeled using dynamical recurrent neural networks. The parameters of the neural network were optimized using an

evolutionary algorithm such that it was able perform the required task. In this section, we specify implementation details about the neural network model, the tasks, and the optimization algorithm.

3.3.1 Neural network model

A Continuous-Time Recurrent Neural Network (CTRNN) was used as the model neural network (Funahashi and Nakamura, 1993; Beer, 1995c). The neural network consisted of three layers: the input layer which was connected by a set of feed-forward weights to the interneuron layer; the interneuron layer was a CTRNN which fed into the output layer; the output layer produced the output of the neural network which was given by a weighted combination of the interneurons' output. The dynamics of each interneuron was governed based on state equations given by

$$\tau_i \frac{dy_i}{dt} = -y_i + \sum_{j=1}^{N} w_{ij} o_j + w_i^{in} I \tag{3.3}$$

$$o_j = \sigma(y_j + \theta_j) \tag{3.4}$$

where y_i refers to the internal state of neuron i; τ_i, the time-constant; w_{ij}, the strength of connection from neuron j to neuron i; o_j, the output of the neuron; I, the input and w_i^{in}, the weight from the input to the neuron. Based on the state of the neuron its output is given by equation 3.4, where $\sigma()$ refers to the sigmoid activation function given by $\sigma(x) = 1/(1 + e^{-x})$, and θ_j refers to the bias of neuron j. The output of the network at any time t, $O(t)$, is estimated as a weighted sum of the outputs of each neuron (weights given by w_i^o), passed through a sigmoid function and scaled to be in the range $[-1, 1]$.

$$O(t) = 2 * \sigma\left(\sum_{i=1}^{N} w_i^o o_i(t)\right) - 1 \tag{3.5}$$

All neural networks described in this chapter were made up of $N = 3$ neurons. The tunable parameters of such a model include the weights between the neurons (w_{ij}), the input weights (w_i^{in}), the output weights (w_i^o), time-constants (τ_i) and biases ($\theta_i j$) of each neuron. The model was

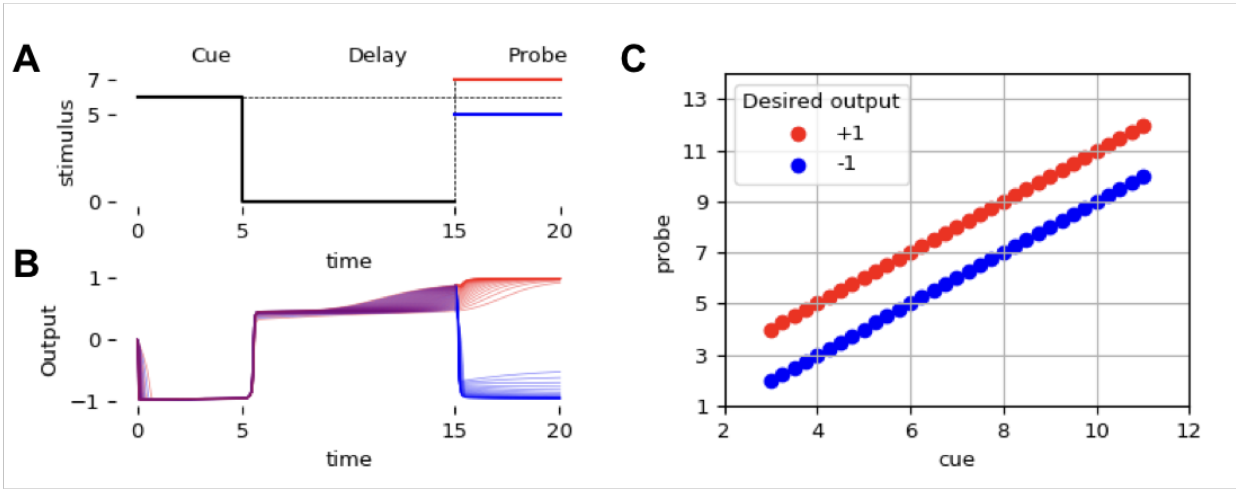

Figure 3.1: Relational categorization task design. [A] Illustration of one trial of the relational categorization task showing the three stages: cue, delay and probe. [B] Sample output for trials corresponding to the two relational categories demonstrating an output of 1 when probe was greater than the cue (red), and an output of -1 when probe was less than the cue (blue). [C] Distribution of cues and probes showing two possible probe values for each cue except in the boundary conditions.

simulated using Euler integration with a step-size of 0.02.

3.3.2 CPG task

The neural network model described above is capable of intrinsically producing oscillations. To create Central Pattern Generators (CPGs), neural networks were optimized to produce oscillations from a range of initial conditions. The neural network was started at 100 different initial conditions by systematically setting the neuron outputs in the range $[0, 1]$. For each condition, the neural activity was recorded for 10 simulation seconds. The ability to generate oscillations was assessed by measuring the absolute difference in each neuron's as well as the neural network's output in consecutive time-steps across all time-points in a trial, and then across trials. The neural network's output was fed to an environment governed by

$$\tau \frac{ds}{dt} = -s + O \tag{3.6}$$

where s refers to the state of the environment, τ refers to its time-constant which was set to 0.5, and O refers to the output of the neural network given by equation 3.5.

3.3.3 Relational categorization task

We adapted the relational categorization task to provide neural networks with structured stimuli (Gentner and Kurtz, 2005; Williams et al., 2008; Markman and Stilwell, 2001). This task involves first providing the neural network with a cue stimulus in the range $[3, 11]$ for 5 units of time. This is followed by a delay period when no stimulus is provided for 10 units of time. Finally, a probe stimulus that is of magnitude greater or less than the cue is provided for 5 units of time. The goal of the task is for the neural network to distinguish probes that were larger than the cue or smaller than the cue, by producing an output of $+1$ or -1 respectively. In the first version of this task, the probe can take one of only two values, either $cue + 1$ or $cue - 1$. In the second version of the task, the probe can take any value in $[3, 11]$. While the goal of the task remains the same in both versions, the distribution of the probes given the cue, and therefore information that the cue gives about the probe is significantly different (Fig. 3.1). Performance of a neural network in this task was estimated by measuring absolute deviation of the network's output from the desired output of $+1$ or -1 during the probe stage. Time-averaged deviation was also averaged across all trials of cue-probe values, to obtain a score in the range $[0, 1]$.

3.3.4 Neural network optimization

Neural network models described previously were optimized to perform the relational categorization and CPG tasks using an evolutionary algorithm (Mitchell, 1998; Goldberg and Holland, 1988). This optimization methodology involves instantiating a population of 100 random solutions that evolves over several generations to produce solutions capable of performing the task. A generation is defined as the process of creating a new population of solutions that has improved in "fitness" (task performance) from the last. Each solution, referred to as a genotype, is an N dimensional vector corresponding to the parameters to be optimized. The parameters were encoded to be in the range $[0, 1]$ and scaled to produce the neural network that the genotype encoded. In each generation, the fitness of every genotype is evaluated and a new population is created using a fitness-based selection and mutation strategy as follows: The genotypes that perform in the top 1%

were retained as is for the next generation. The rest of the individuals were created by selecting two genotypes preferentially in proportion to their fitness and combining them. To these offspring, Gaussian mutation noise with mean 0 and standard deviation 0.01 was added before being added to the population of genotypes for the next generation. After a fixed number of generations, the best individual in the population was selected as the representative solution from that optimization run. 100 such runs were conducted to obtain an ensemble of 100 neural network models that successfully performed each task. For the relational categorization task, optimization was carried out for 500 generations. In the case of the CPG task, at the end of 50 generations the optimization process was terminated and deemed successful if the best agent in the population reached a fitness of 30 or greater. This was repeated until 100 CPGs were produced.

3.3.5 Random neural networks

Matched random neural networks were created for the relational categorization task by shuffling the parameters of the optimized neural networks. All parameter groups, namely time-constants, input weights, recurrent weights, output weights, and biases were randomly shuffled within themselves rather than across groups. Thus, the ranges of parameters were preserved in each group but their associations with neurons were randomly shuffled.

3.4 Identifying the source of predictive information

Predictive information is the information encoded in neural activity about its future stimulus. Formally, it is defined as mutual information between current neural activity (N_t) and the stimulus at a future time ($S_{t'}$) (Palmer et al., 2015; Bialek, 2012; Rieke et al., 1999; Cover and Thomas, 2012; Shannon, 1948), according to:

$$I(S_{t'}, N_t) = \sum_{s_{t'}, n_t} P_N(n_t) P(s_{t'}|n_t) \log_2 \frac{P(s_{t'}|n_t)}{P_S(s_{t'})} \tag{3.7}$$

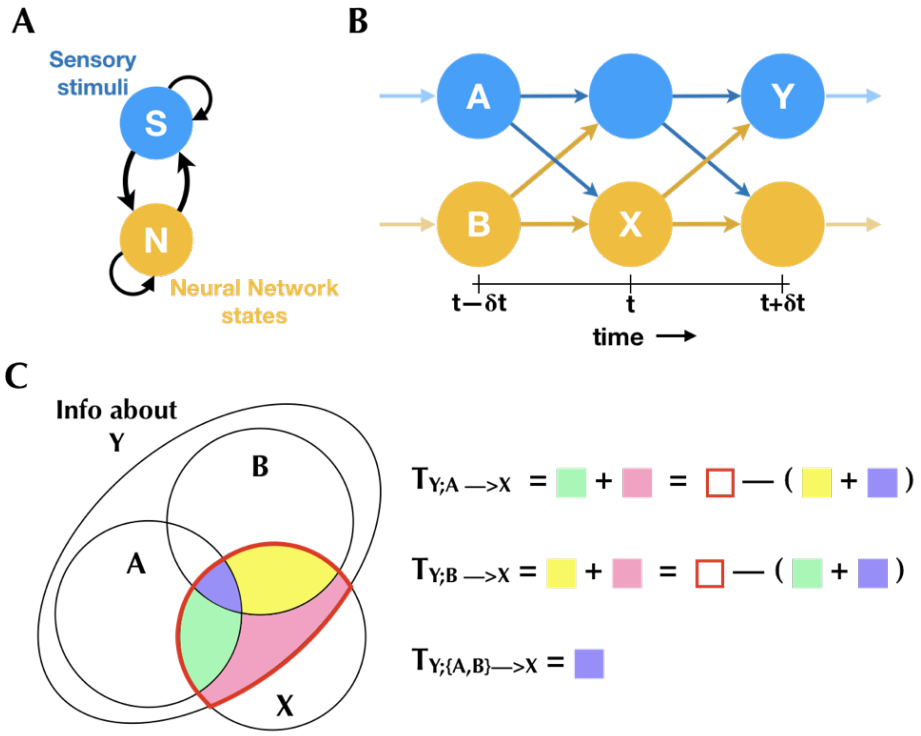

Figure 3.2: Predictive information source estimation based on idealized agent-environment interaction. [A] Sensory stimuli (S) and neural activity (N) are two coupled dynamical systems. [B] Agent-environment interaction unrolled over time. X represents current neural activity, $N(t)$, Y, future environmental state, $S(t+\delta t)$, and A and B represent the sources, namely past neural activity $N(t-\delta t)$ and past environmental state, $s(t-\delta t)$ respectively. [C] Partial information diagram for calculating the sources of predictive information in an agent-environment system. The total information that X has about Y is a combination of information that is available uniquely from A alone (green), uniquely from B alone (yellow), synergistically from their combination $[A,B]$ (pink), and redundantly from both of them (purple). PID allows us to measure information transfer using these components. Alternatively, they can also be measured by estimating the total redundant information from both sources combined (red) and removing the information from the other source.

where $t' = t + \delta t$ with $\delta t > 0$, P_S is the distribution of environmental stimuli, P_N is the distribution of neural activity across the entire experiment, $P(s_{t'}|n_t)$ is the conditional probability that the stimulus is s at a future time t' given that we have observed a neural activity of n at time t. When this measure is estimated using the stimulus and neural activity across all data points separated in time by some δt, it is a measure of reduction in uncertainty in future stimulus given the current neural activity.

The presence of predictive information in a neural network suggests there is a source where this

information was generated. In an idealized agent-environment system (Fig. 3.2A), the source of information can be either the neural activity in the previous time step, the environmental stimulus in the previous time step, or both (Fig. 3.2B). Measuring predictive information as defined in equation 3.7 requires that we examine two variables: current neural activity (N_t, henceforth X) and future stimulus ($S_{t+\delta t}$, henceforth Y). Identifying the source of this predictive information requires that we examine two additional variables: past neural network activity ($N_{t-\delta t}$, henceforth A) and past stimulus ($S_{t-\delta t}$, henceforth B). Such a four-variable analysis requires that we adopt multivariate extensions to information theory. We focus specifically on Partial Information Decomposition (PID) (Williams and Beer, 2010b), a method for decomposing multivariate mutual information into combinations of unique, redundant and synergistic contributions, as well as measures of information gain, loss and transfer (Bertschinger et al., 2014; Faber et al., 2019; James et al., 2016, 2011; James and Crutchfield, 2017; Lizier et al., 2018; Timme et al., 2014; Wibral et al., 2017; Williams and Beer, 2011, 2010b). In order to identify the source of predictive information, we can decompose the total information that the current neural activity has about the future stimulus into three components: (a) information uniquely transferred from past environmental stimulus, $T_{Y;A\to X}$; (b) information uniquely transferred from past neural network activity, $T_{Y;B\to X}$; and (c) information redundantly transferred from past environment stimulus and past neural network activity, $T_{Y;\{A,B\}\to X}$, according to:

$$\begin{aligned} T_{Y;A\to X} &= \Pi_R(Y;\{[A,B],X\}) - \Pi_R(Y;\{B,X\}) \\ T_{Y;B\to X} &= \Pi_R(Y;\{[A,B],X\}) - \Pi_R(Y;\{A,X\}) \\ T_{Y;\{A,B\}\to X} &= \Pi_R(Y;\{A,B,X\}) \end{aligned} \tag{3.8}$$

where $\Pi_R(Y;\{A_1,A_2,..A_k\})$ is the redundant information that random variables A_1 through A_k have about the random variable Y and $[A,B]$ refers to a random variable that is a concatenation of A and B. In words, information about Y transferred uniquely from source A to X is estimated as the total redundant information from the combined sources $[A,B]$ *minus* the information that

is redundant with the other source B. This decomposition of the total information into different contributions is typically represented using a PI-decomposition diagram (Fig. 3.2C). The PID measures described here, as described in Williams and Beer (2010b), use I_{min} to estimate redundant information, Π_R. Since these metrics have been proposed, there have been a few alternative approaches proposed (Harder et al., 2013; Griffith et al., 2014; Ince, 2017) but we use I_{min} because to date, this is the only approach that can guarantee non-negative decomposition of information in a four-variable system.

During the course of behavior, the flow of information in a system changes over time (Izquierdo et al., 2015a; Williams and Beer, 2010a). In order to understand the source of predictive information for any agent-environment system, it is not enough to decompose information from multiple sources; we must also track its flow of information over time. Although information theoretic measures are typically averaged over time, the measures described above can be unrolled over time (Williams and Beer, 2010a; Williams, 2011). This is done by measuring information transfer at each time-point using data collected across several trials thereby allowing us to study the dynamics of predictive information sources.

3.4.1 Measuring information transfer

To identify the source of information over time, information transfer measures were estimated independently at each time point. For any given time step, data for environmental stimulus at the previous time step, neural activity of previous time step, current neural activity, and stimulus at a future time step, was collected across multiple trials. Probability densities were estimated from this data using a kernel density estimation technique known as average shifted-histograms (Scott, 1985a) with 7 shifted binnings of 100 bins along each dimension of the data space. These probability density estimates were then used to measure the redundant information terms in equation 3.8. All information theoretic quantities were estimated from raw data using the *infotheory* package (Candadai and Izquierdo, 2019a).

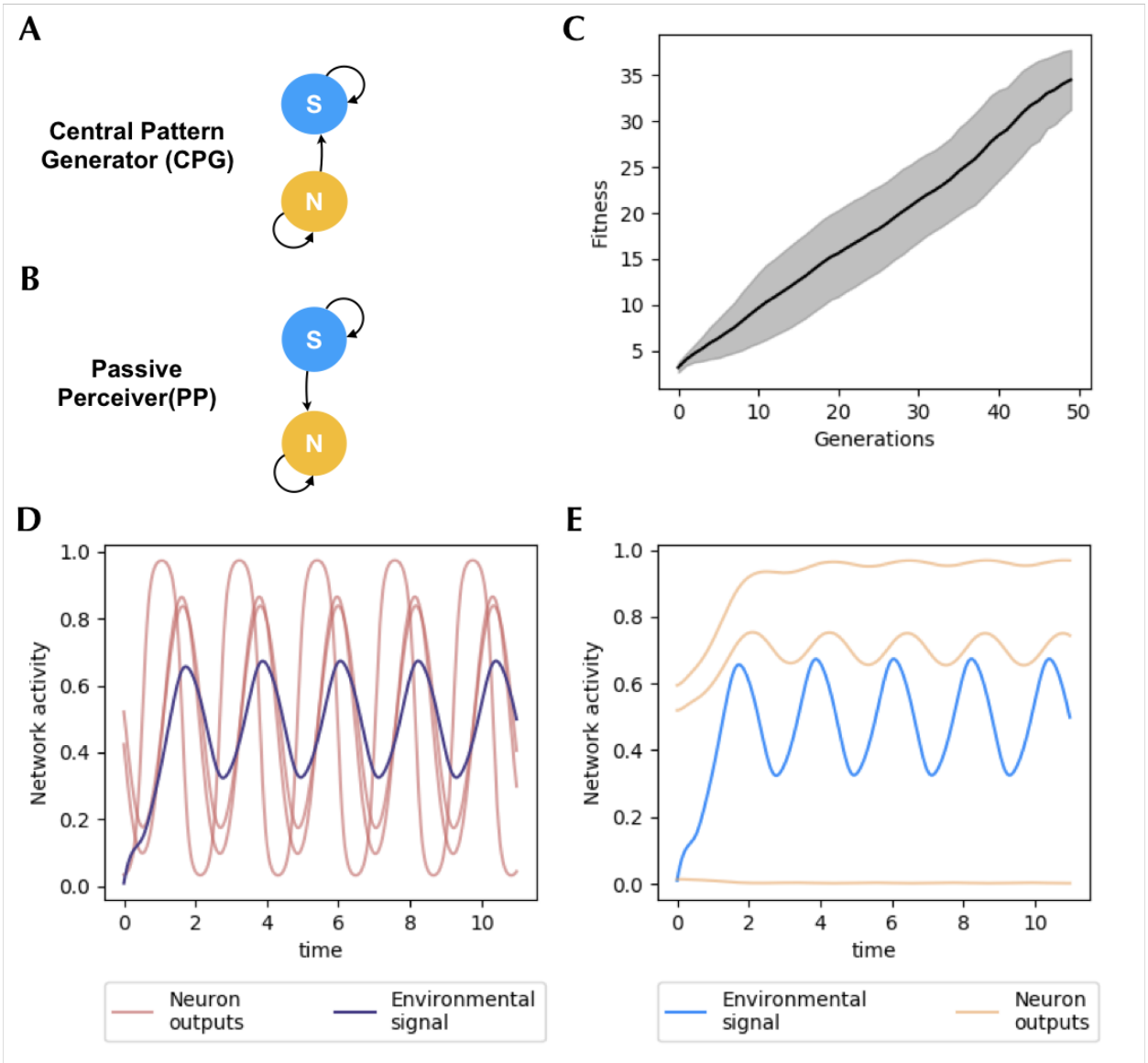

Figure 3.3: Optimization and neural traces of CPG and PP. [A] Schematic and traces of a Central Pattern Generator (CPG) that influences the environment through intrinsically generated oscillations. [B] Schematic and traces of a Passive Perceiver (PP) that is driven by oscillatory inputs from the environment (in this case, by the environmental signals recorded from the CPGs) [C] Fitness over time for 100 valid runs of optimizing a CPG model. Only runs that achieved a fitness greater than 30 were deemed valid. [D] Neural traces from one trial of the best CPG demonstrating that all neurons (red) as well as the neural network output (blue) oscillate. [E] Neural traces (orange) when the output from the CPG shown in panel B was fed to a random neural network in the PP condition demonstrating input driven oscillation in the random neural network.

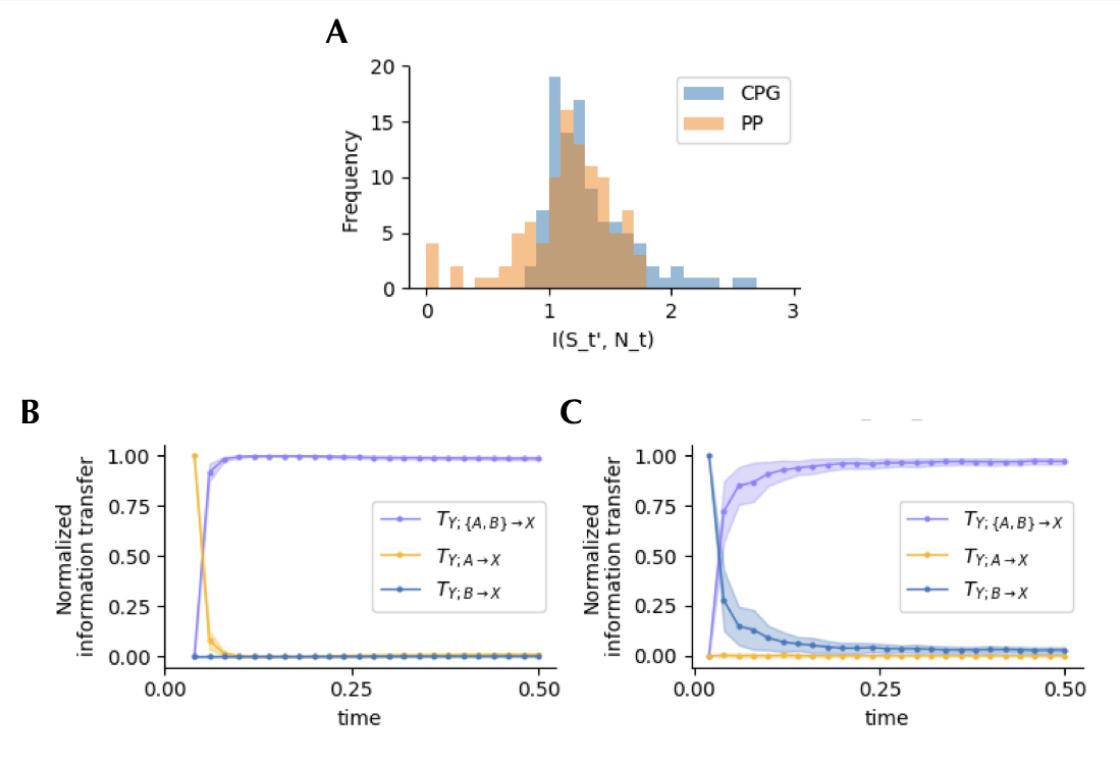

Figure 3.4: Predictive information in systems on the extremes of the range of possible agent-environment interactions [A] Estimating total predictive information as shown in equation 3.7 shows that CPG and PP models encode similar amounts of predictive information about environmental state in the next time-step. [B] Decomposing that total information into information that came from the environment and the neural network consistently showed that information about the next time-step in the CPG originated in the neural network (yellow) before becoming redundant (purple) as the environment and the neural network synchronized. [C] Conversely, with PPs, the environment was consistently shown to be the source of information (blue) before they environment and neural network synchronize and become redundant (purple).

3.5 Disparate systems with similar predictive information

Neural systems can be predictive in fundamentally different ways: they can generate predictive information internally or they can extract it from environmental stimulus. We use computational models of two extreme conditions where the ground-truth predictive information source is known to be the environment in one condition and the neural network in the other, to demonstrate that (a) predictive information cannot distinguish between these different kinds of systems and (b) it is only through decomposing the information across sources and unrolling over time that we can distinguish the two systems based on their operation. The two conditions we consider are

agent-environment interactions at two extremes of the range of possible interactions: a central pattern generator (CPG) and a passive perceiver (PP). In the CPG condition, the neural network influences the environment by producing spontaneous oscillatory activity but receives no input from the environment (Fig. 3.3A). In the PP condition, the neural network is influenced by input from the environment, but it does not affect the environment (Fig. 3.3B). We evolved 100 different dynamical recurrent neural network CPGs, and in each case, we fed the sum of the neurons' outputs to the environment. All runs produced CPGs that could reliably produce oscillations (Fig. 3.3C,D). For the PPs, we generated 100 random neural networks and fed them an oscillatory input. In order to provide the same distribution of activity as the CPG condition, we provided the random neural networks with the same oscillatory environmental signal that CPGs generated which caused the random PP networks to oscillate (Fig. 3.3E). The environmental signal and neural data were recorded from each instance for 500 trials where, in each trial the environment started with a different initial condition. Although, the environmental signal and the neural activity exhibit oscillatory activity in both conditions, the key difference in the operation of these systems is that in the CPGs the neural network drives its own activity and in the PP, the environment drives the neural network. Therefore, the neural network is the source of predictive information in the CPGs and the environment is the source of predictive information in the PPs.

As a first step in the analysis of these two systems, we used the recorded data to measure predictive information in the neural network about the environmental signal in the next time-step ($\delta t = 0.02s$). To calculate predictive information, data distributions were constructed using all tuples of neural activity at time t and environmental signal at time $t + \delta t$, averaged across time and trials. The analysis revealed that the neural networks, in these two otherwise diametrically opposed systems, encoded similar levels of information about stimulus in the next time step (Fig. 3.4A). From this first experiment, we conclude that predictive information is not sufficient to distinguish systems that generate their own predictive information from systems that encode the information available from the environmental stimuli.

To understand what makes these two neural systems different, it is necessary to identify the

source of their predictive information. As a next step in our analysis, we decomposed the information in the neural system about the future stimuli across the different possible sources and we unrolled the analysis over time. At each time-point, we measured information in the neural network about the environmental signal in the next time-step that was uniquely transferred from the environment, uniquely transferred from the neural network and redundantly from both.

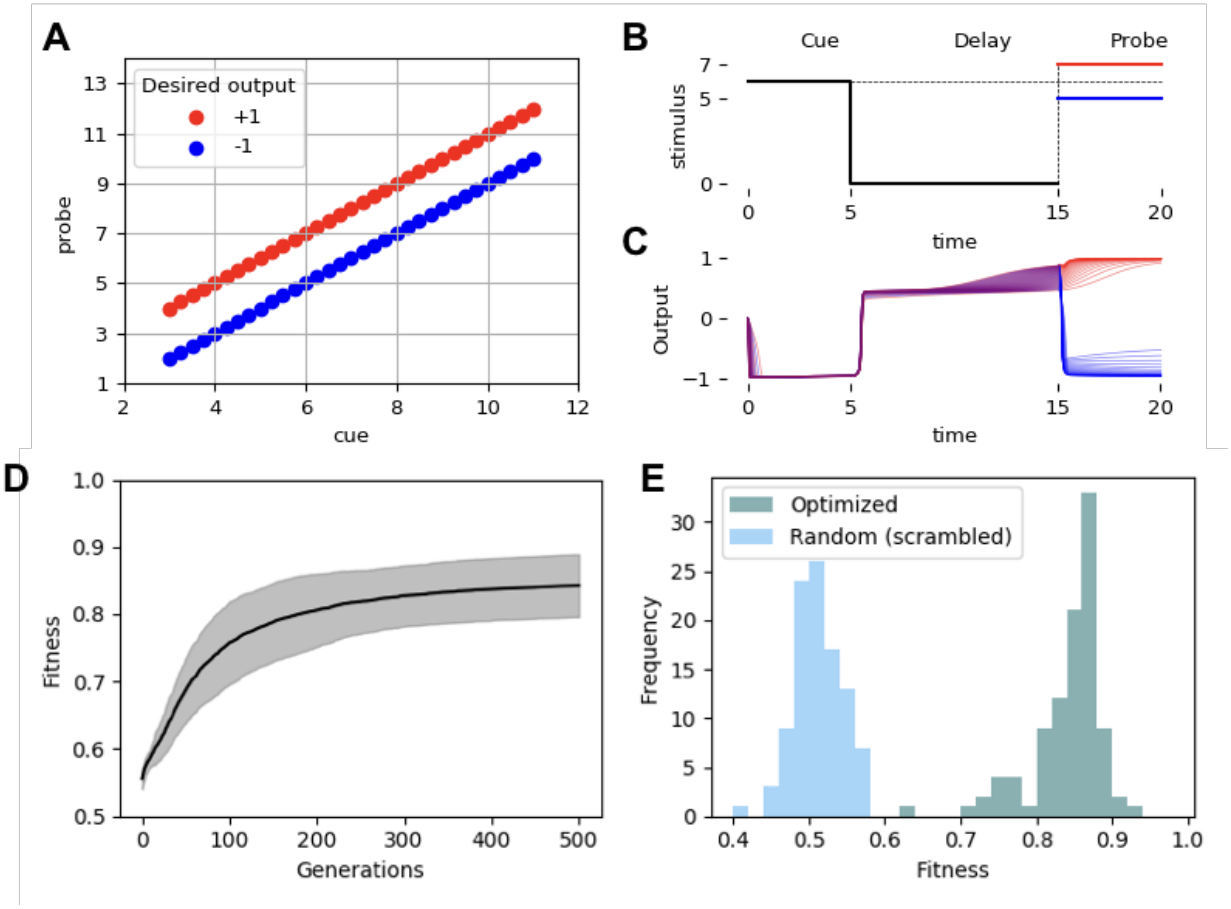

Figure 3.5: Optimizing neural networks to perform relational categorization. [A] Distribution of cue and corresponding probes in the relational categorization task. For each cue, the probe can be one of two values: greater, $cue + 1$, or lesser, $cue - 1$, with the expected outputs of +1 (red) and −1 (blue) respectively. [B] One trial of the relational categorization task. The cue stimulus is presented till t=5, followed by a delay period with no stimulus (t=5 to t=15) and then a probe that is greater (red) or lesser (blue) than the cue is provided. [C] Behavior of the best out of 100 dynamical neural networks optimized to perform this task showing perfect categorization of the relational value from 35 trials where the probe was greater (red) and 35 where the probe was lesser (blue). [D] 100 independent runs all converged to near-perfect performance with deviation from a perfect score only due to small deviations from expected output and not mis-categorization. [E] Neural networks whose weights and time-constants were scrambled lost their ability to perform the task.

In the CPG condition, since the neural networks are not influenced by the environment (Fig. 3.3A), the only source of information about the future environmental signal is from the neural network itself. Accordingly, the dynamics of information transfer for CPG systems reveals correctly that the neural network is the source of predictive information (Fig. 3.4B). At the start of the interaction between agent and environment, the neural network uniquely transfers information about the future environmental state to the environment. Following that, the environment quickly becomes synchronized with the neural activity. This means that the state of the environment becomes informative of its own future state. This results in the environment and the neural network becoming redundant sources of predictive information. Crucially, however, the environment never provides any unique information to the neural network about its future stimulus.

In the PP condition, since the neural networks are driven by the environment (Fig. 3.3B), the only source of information about the future environmental signal is the stimulus from the environment itself. Accordingly, the dynamics of information transfer for PP systems reveals correctly that the environment is the source of predictive information (Fig. 3.4C). As opposed to the CPG systems, at the start of the interaction between the neural network and the environment, it is the environment that transfers unique information to the neural network. Subsequently, and similarly to the CPG condition, as the state of the neural network begins to encode the information from the environmental stimulus, the predictive information is redundantly transferred by both the neural network and the environmental stimulus. Consistent with our expectation, the neural network never provides any unique information to itself about the future of the stimulus.

In summary, in this section we show that predictive information alone cannot distinguish between two extremely different kinds of neural systems, both of which encode predictive information about the future of the environment. This is because when the entire time course of the data is considered, the environment and neural network are synchronized for a majority of the time. Information uniquely transferred from any source is only detectable within a short time window before they synchronize. In this section, we have shown that decomposing information across sources and unrolling over time allows us to study information source dynamics at every perturbation to the

agent-environment interaction and hence reveals the source of predictive information.

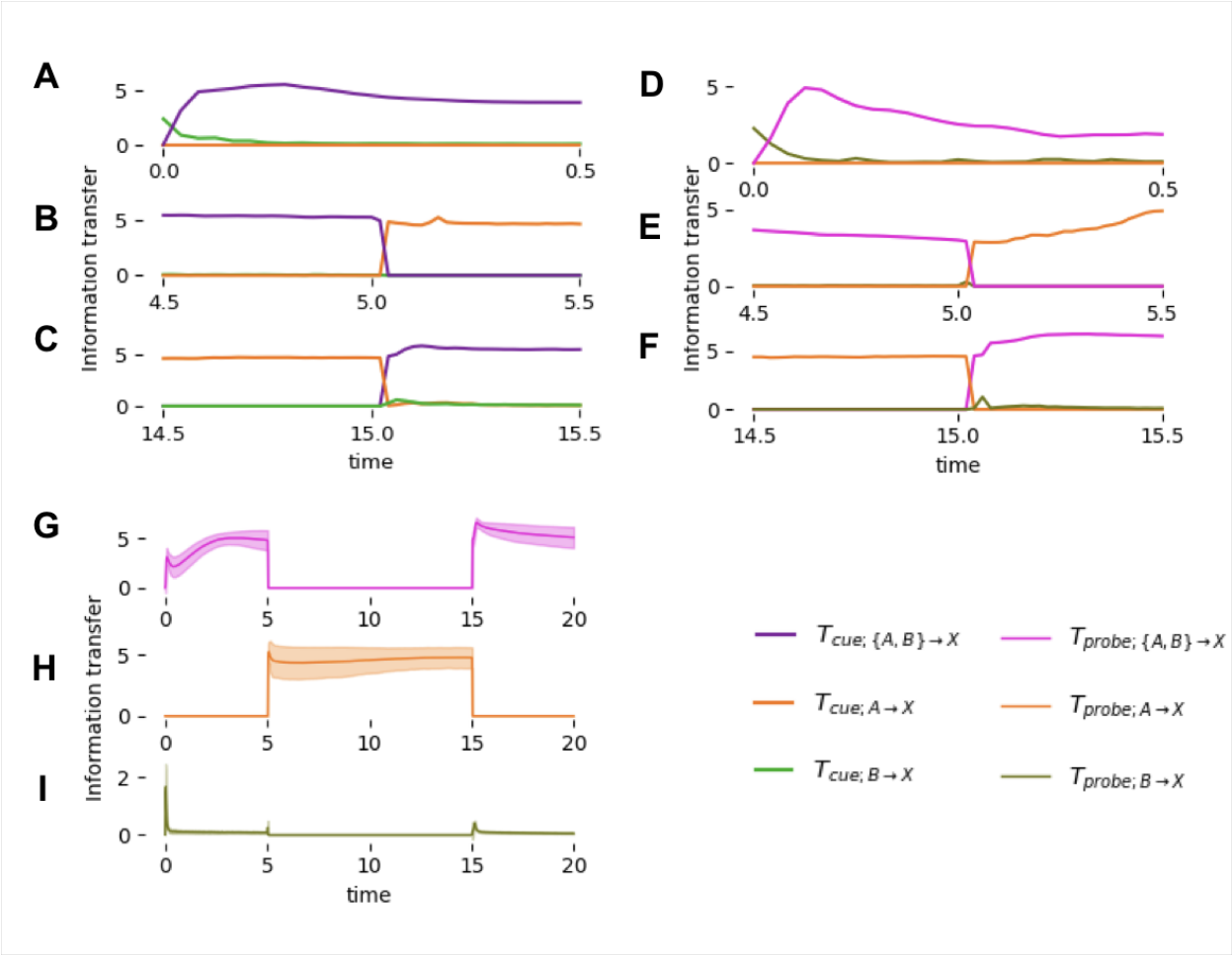

Figure 3.6: Predictive information source dynamics with structured stimuli. [A] Dynamics of information about the cue during the cue stage show information uniquely provided by the environment (green) initially, but becoming redundantly available in the neural network and environment (purple) as it encoded the cue. [B] Towards the end of the cue stage, information is entirely redundant (purple). When the stimulus stops being provided at t=5, the neural network is the unique source of information about the cue (orange). [C] Dynamics of information about the cue just before the probe arrives showing that the neural network continues to retain information about the cue (orange). At t=15, when the probe is provided, information quickly becomes redundant (purple) denoting that the probe has information about the cue.

3.6 Predictive information with structured stimuli

The natural environment is not uniformly random but is in fact highly structured with spatial and temporal regularities (Barlow, 2001; Graham and Field, 2007; Huang and Rao, 2011). This

structure is reflected in the stimulus that agents receive from the environment. Accordingly, this is emulated in most preparations in neuroscience, where a neural system is presented with artificial stimuli with some underlying structure designed by the experimenter. We posit that the structure in the environment will strongly influence the amount of predictive information encoded by the neural network and its sources. In order to study this, we examined the flow of information in a neural network model trained to solve a relational categorization task.

Relational categorization is the ability to discriminate objects based on the relative value of their attributes (Gentner and Kurtz, 2005; Markman and Stilwell, 2001). This task allows us to specify the inherent structure in the environment by changing the distribution of objects whose attributes are compared thus making it especially suited for studying the influence of environmental structure on predictive information. It involves providing the neural network with stimuli across three stages: cue, delay, and probe. In the cue stage, the neural network is provided with a stimulus of specific magnitude for a duration of time. This is followed by a delay stage, where no stimulus is provided. Finally, in the probe stage, the neural network is provided with a second stimulus. The magnitudes for the cue and probe stage stimuli are picked from a predesignated distribution (Fig. 3.5A). It is this distribution that defines the structure in the environment. For this study, we design it such that the stimulus in the probe stage can have a magnitude that is one of two values: smaller ($cue - 1$) or larger ($cue + 1$) than the stimulus provided during the cue stage (Fig. 3.5B). The goal of the neural network in this task to perform a relational categorization of "greater than" or "lesser than" by producing an output of $+1$ or -1 respectively, during the probe phase. This task has been widely studied in a variety of contexts including in humans (Kurtz and Boukrina, 2004), pigeons (Wills, 1999), rats (Saldanha and Bitterman, 1951), insects (Giurfa et al., 2001), as well as using computational models (Williams et al., 2008; Izquierdo-Torres and Harvey, 2005).

In this section, we show results from analysis of neural network models performing the relational categorization task. We demonstrate that decomposing information across the sources and unrolling over time reveals that the environment is structured by appropriately attributing the observed predictive information to either the environment or the dynamics of the neural network.

Furthermore, we demonstrate that encoding predictive information alone is not indicative of task performance and that the magnitude and source of predictive information can change during the course of a behavior depending on environmental structure and neural network dynamics.

3.6.1 Characterizing information source dynamics in the best optimized neural network

Dynamical recurrent neural networks were optimized using an evolutionary algorithm to perform the relational categorization task. A total of 100 independent evolutionary runs yielded an ensemble of 100 different neural networks that could successfully perform the task (Fig. 3.5D). The best neural network from this ensemble achieved a performance of 93.12%. Although this neural network correctly classified all probes, the performance score was not perfect due to slight deviations in the output (Fig. 3.5C).

In order to better understand how a neural network performed this task, we can characterize the flow of information across the agent-environment system. To this end, we decomposed the total information that the best neural network from the ensemble had about the cue into information uniquely transferred from the environment, uniquely transferred from the neural network, and redundantly from both, during the course of the task. During the cue stage, the environment was initially the unique source of information about the cue (Fig. 3.6A). As the neural network encoded the stimulus, the source became redundant. During the delay stage, the environment ceases to be a source of information. As the neural network had already encoded information about the cue, it becomes the unique source (Fig. 3.6B). Crucially, the neural network preserves this information throughout the delay stage. Finally, during the probe stage, the environment once again becomes a source, and therefore the source is redundant (Fig. 3.6C). Note that when the environment provides the probe stimulus it became the source of information about the cue. Since the neural network already contained information about the cue, the neural network and the environment both redundantly act as the source.

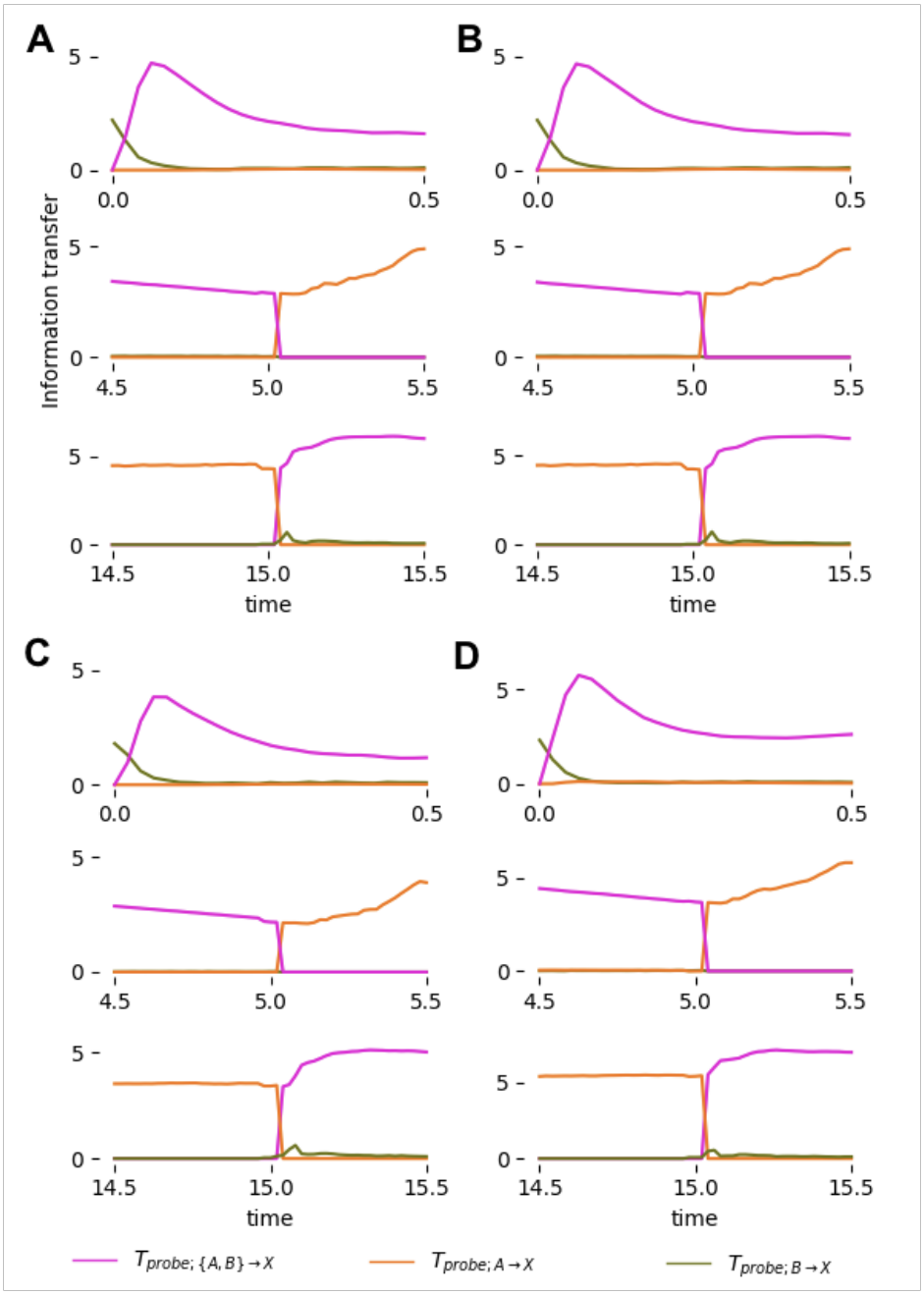

Figure 3.7: Inferring the source of predictive information is robust to different binning and shifted-histograms. Results are qualitatively similar to results from fig. 3.6A after changing [A] number of shifted bins to 3 [B] number of shifted bins to 11 [C] number of bins per dimension to 50 and [D] number of bins per dimension to 200, for cue (top row), delay (middle row) and probe (bottom row) stages of the task.

As explained previously, predictive information in this task arises from the relationship between cue and probe stimuli. Encoding information about the cue automatically results in encoding

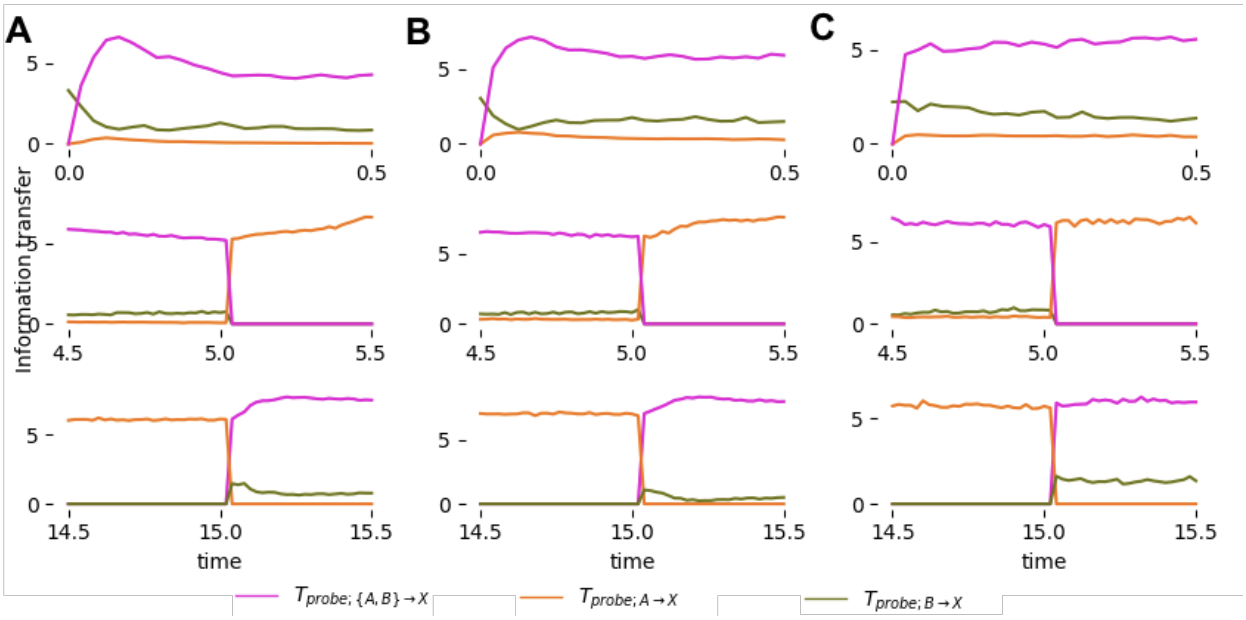

Figure 3.8: Inferring the source of predictive information is robust to zero-mean Gaussian noise with standard deviation [A] 0.01, [B] 0.05 and [C] 0.1. Results are qualitatively similar to results from fig. 3.6A for cue (top row), delay (middle row) and probe (bottom row) stages of the task.

information about the probe (and vice versa). This is because knowing the cue significantly reduces uncertainty about the probe; the probe can only be one of two values given a cue. Predictive information that the neural network has about the probe and its sources is qualitatively similar to the information it has encoded about the cue (Fig. 3.6D-F). The neural network encodes information about the probe stimulus upon receiving the cue, and retains that predictive information during the delay stage. This is merely a consequence of encoding and retaining the cue. The entire ensemble of neural networks optimized to perform this task consistently exhibit this phenomenon of encoding information about the probe transferred uniquely from the cue stimulus (Fig. 3.6G-I). Similar results were observed when these information quantities were measured using 5 and 11 average shifted-histograms and with 50 and 200 bins per dimension (Fig. 3.7). Moreover, these results were robust to noise in the neural network (Fig. 3.8).

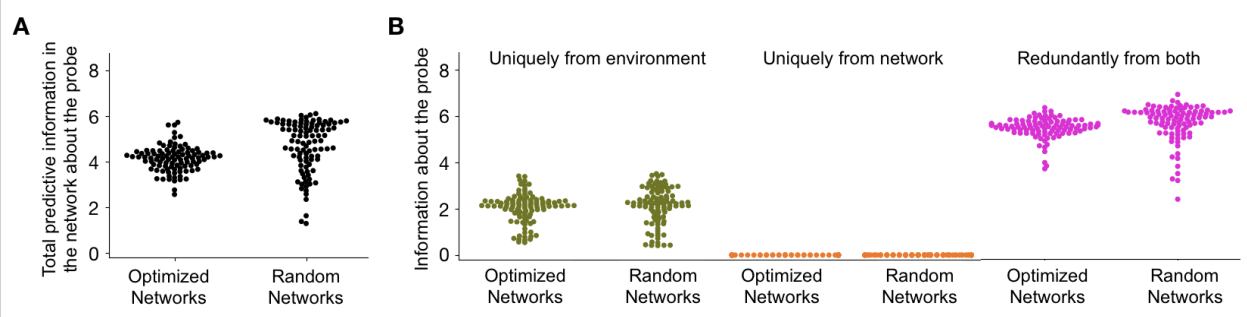

Figure 3.9: Comparison of predictive information sources in optimized and random neural networks. [A] Total predictive information estimated by averaging over the entire course of the task is similar in random and optimized neural networks. [B] Total predictive information about the probe averaged across the cue stage of the task, is the same in random and optimized neural networks. [C] Decomposition of that total predictive information showing that information about the probe in both random and optimized neural networks was from the environment (green), eventually becoming redundant as they both encoded the cue stimulus (pink). The neural network had no role to play in its encoding of predictive information about the probe during the cue stage (orange).

3.6.2 Environmental regularities induces predictive information in any neural network

Since optimized neural networks encode information about the probe merely by encoding the cue, does any neural network that encodes the cue also encode information about the probe, and therefore have similar predictive information? In order to study this, we created 100 random neural networks and presented them with the same task. Although these neural networks were not able to perform the relational categorization task (Fig. 3.5B), they encoded similar amounts of total predictive information as the trained neural networks (Fig. 3.9A). Specifically, they encode the same amount of information about the probe during the cue stage (Fig. 3.9B). Furthermore, decomposing that information revealed that the information originated from the environmental stimulus and that the neural network dynamics had no role in its encoding of predictive information in both random and optimized neural networks (Fig. 3.9C). Thus, predictive information alone is not sufficient to distinguish neural networks optimized to perform specific tasks from random neural networks that are merely reflecting the information provided by the environment.

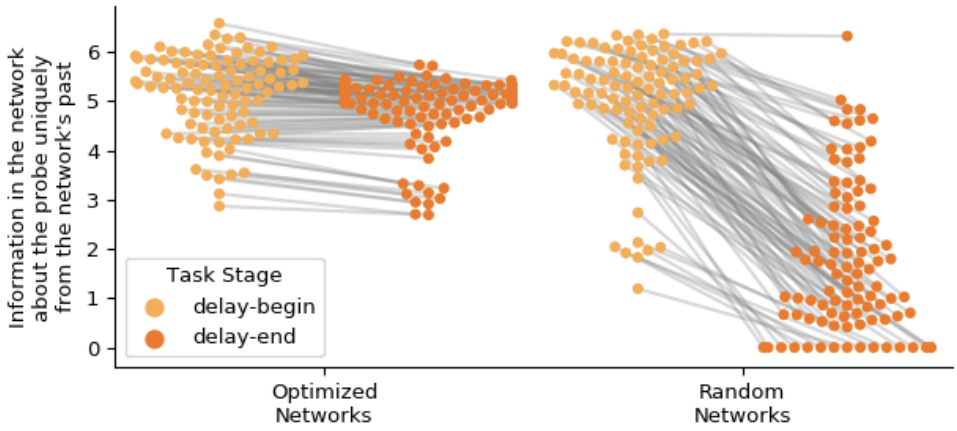

Figure 3.10: Influence of neural network and environmental properties on predictive Information. [A] Both random and optimized neural networks have similar levels of information about the probe at the beginning of the delay stage (light orange), but unlike optimized neural networks, random neural networks lose that information by the end of the delay stage (dark orange). [B] Total predictive information in the optimized neural networks about the probe during the cue stage showed a significant drop upon changing environmental statistics from 2 probes/cue to 9 probes/cue. [C] Drop in total information show in B can be attributed to the drop in information uniquely from the environment about the probe in the 9 probes/cue setting.

3.6.3 Information decomposition distinguishes between random and optimized neural networks

Unlike CPG and PP that were distinguished based on having different information sources, random and optimized neural networks in the relational categorization task have the same information sources. Even under this condition, decomposing the total information across sources and unrolling over time helps distinguish them by revealing differences in the magnitude of information transferred from each source over time. Specifically, predictive information sourced by the neural network during the delay stage is markedly different between random and optimized neural networks. As discussed in the previous section, optimized neural networks preserve information about the cue (and hence predictive information about the probe) during the delay stage. In contrast, random neural networks tend to lose that information. As a consequence, the amount of unique information provided by the neural network at the end of the delay period is higher for the trained neural networks than for the random neural networks (Fig. 3.10A). This difference disappears when information is

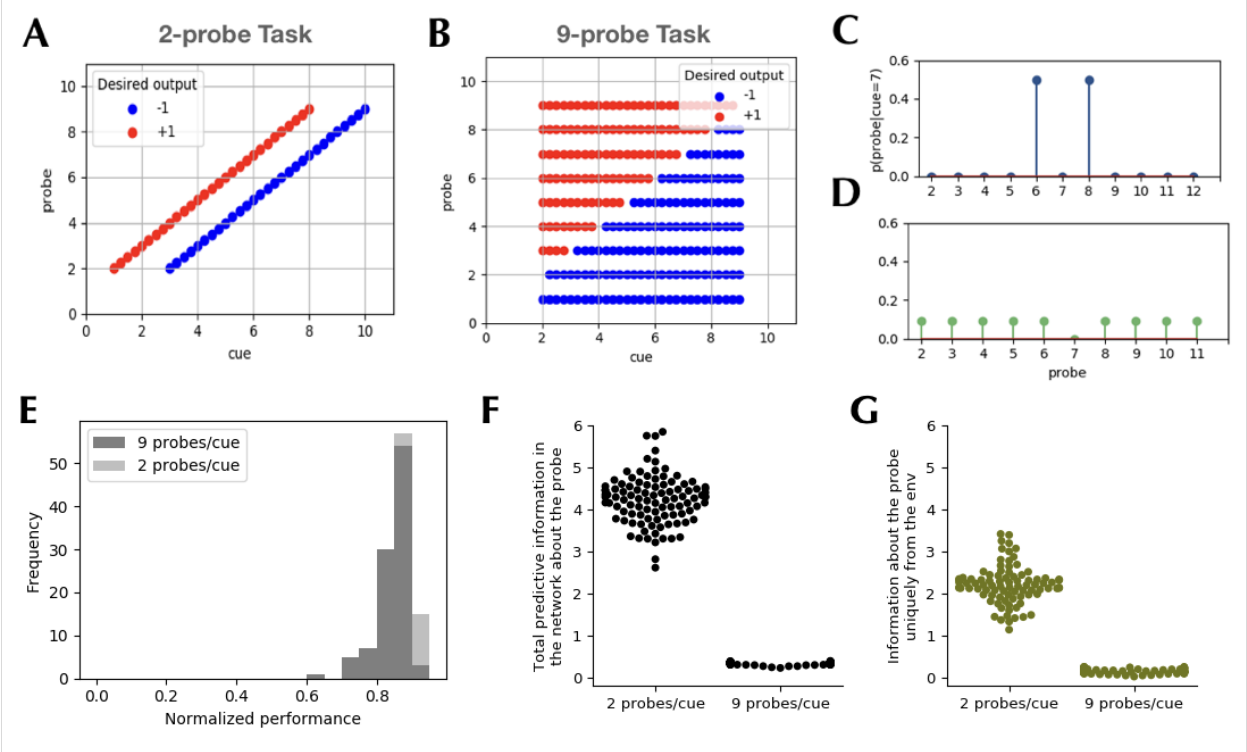

Figure 3.11: Different environmental structures within the relational categorization task [A] Relational categorization task with highly structured stimuli; for each cue probe is one of two possible values. [B] Relational categorization task with minimal structure in stimuli; probe can be one of 9 values for a given cue. [C] Conditional probability of probes given a cue for environmental structure shown in panel A, demonstrating the significant reduction in uncertainty of the probe given the cue. [D] Conditional probability of probe values given a cue under the environmental structure in panel B shows that probe values still have a nearly uniform distribution, and hence very less reduction in uncertainty. [E] Neural networks optimized to perform under the distribution shown in panel A perform just as well under the distribution shown in panel B.

measured across time, and can only be observed by unrolling it over time.

3.6.4 Statistics of the environment influences magnitude of predictive information

Encoding the cue results in encoding information about the probe in this task because of the relationship between them. How does changing this relationship impact predictive information in the neural networks? In order to study this, without changing the nature of the relational categorization task we merely changed the structure in the environment. This was achieved by modifying the task such that the probe could be one of 9 possible values for a given cue, rather than one of two possible values (Fig. 3.11B). Reduction in uncertainty about the probe's value given the

cue is now much less compared to the original environmental structure (Fig. 3.11C,D). This will be reflected in the information that the cue can provide about the probe. However, this came at no cost to performance because the neural networks were still encoding the cue just as well. The same ensemble of optimized neural networks were able to perform this task successfully without any more training (Fig. 3.11E). Information dynamics was then measured using data recorded under this 9-probe condition. Measuring the total information in the neural network during the cue stage about the probe revealed that there was significantly less information in the neural network in 9 probes per cue condition (Fig. 3.11F). The reduction in total predictive information can be wholly attributed to the reduction in information about the probe (Fig. 3.11G). Thus, differences in environmental structure can result in significantly different amounts of predictive information encoded in neural networks without any behavioral differences.

3.7 Discussion

The study of predictive coding and its relevance to behavior has been studied from multiple perspectives in the literature with regards to the source of information: predictive information can be generated by the neural network (Friston, 2010; Friston and Kiebel, 2009) and predictive information can be provided by the environment (Bialek, 2012; Bialek et al., 2001). In this work, using computational models where the ground-truth about the source of information was known, we demonstrate that predictive information can originate from either the environment or the neural network or both, and that the source of information can dynamically change during the course of a behavior. In order to do this, we first presented a theoretical framework based on multivariate information theory that allows us to infer the source of predictive information and its dynamics. This involved decomposing the total information that neural networks encode about a future stimulus into information transferred uniquely from the neural network, uniquely from the environment and redundantly from both sources. We validated this framework using the CPG and PP models where information is known to originate from the neural network and the environment respectively. Second, using the more structured relational categorization task, we demonstrated that (a) amount

of predictive information encoded in a neural network is not indicative of its performance; (b) the source of information about a future stimulus can change during the course of the task; and (c) the source of information about a future stimulus can change within the same task depending on the regularities of the environment. Thus, predictive information might be necessary but is not sufficient to explain the neural basis of a behavior. Decomposing information across sources and studying its dynamics over time takes us one step further in understanding the role of predictive information in a behavior.

The framework presented here for inferring the source of predictive information takes us beyond general correlations that information theoretic measured are known to capture by capturing the effects of perturbation on the neural system. Identifying the sources of predictive information requires that the system under study be perturbed. The presentation, removal or sudden change of a stimulus is a perturbation. This causes the system to break the redundant encoding observed in a steady-state. It is during such a perturbation that we can use partial information decomposition to determine the source of information in a coupled system. Once the neural network and the environment settle into the next steady-state after the transient due to the perturbation, information once again becomes redundant between them. Thus, through the combination of information decomposition, time-unrolling and perturbation we are able to infer the ground-truth causal influences in the models we have analyzed.

The framework presented here can be applied to experimental data across multiple scales. In fact, it can be applied to any time-series data spanning multiple trials corresponding to several perturbations from the steady state. However, in this work, we focus on open-loop systems. Specifically, we focus on agent-environment systems where the agent influences its environment or where the agent is influenced by the environment. Such an open-loop setup is typical in experiments in neuroscience, where the subject receives a stimulus, but does not have the ability to influence the future stimulus through their state or actions. In natural behavior, the agent and environment are in closed-loop interaction. The analysis of closed-loop systems introduces an added complexity. The regularities of the environment can be generated by the regularities of the neural network's

dynamics and vice-versa. As a result, the distribution of environmental stimuli and the distribution of the neural activity are dependent on each other, unlike the open-loop setup where one of them is independent of the other. As it is, the framework requires that one of the distributions be fixed across time in order to make fair comparisons of information at different time-points. Future work in this direction will involve extending the framework and designing the experimental setting that would allow us to infer the source of predictive information in a freely moving animal.

The general idea of living organisms implementing predictive coding has strong theoretical and experimental support. Importantly, experimental evidence and theoretical proposals at the moment have the potential to encourage the notion that predictive information can be acquired from the environment while also being generated internally. Our work provides a framework within which such a dual contribution, externally from the environment and internally by the living organism, can be studied. Moreover, it enables a study of temporal dynamics in the contribution of these two sources. We also show that environmental regularities can significantly offset effort required to acquire this information by simply providing it at no (or minimal) cost to the animal. This makes the case for taking into consideration environmental variables into our purview of analysis to delineate the role played by the brain in behavior. Ultimately, including the environment in the study of neural basis of behavior will enable a comprehensive study of embedded embodied living systems.

Chapter 4

Multifunctionality

In this chapter, I discuss our work analyzing the extent to which neural resources can be reused across tasks in multifunctional neural networks. In the first section of this chapter, I introduce our work and its motivation. Following that, I discuss the literature involving reuse of neural circuitry across behaviors, and its associated mechanisms. Following that, I explain the methodological details involved in building models of multifunctional neural networks, and analyzing them. Finally, I explain the results of the analyses and then discuss the implications of our results.

4.1 Introduction

A crucial aspect to adaptation in cognitive beings is their ability to exploit regularities in the environment and reuse existing resources across multiple behaviors. Extensive empirical evidence shows that neural resources optimized during the course of learning one behavior are reused for others (Anderson, 2010a). This multi-functional ability of neural circuits has been demonstrated in the small nervous systems of the nematode worm *Caenorhabditis elegans* (302 neurons) (Hobert, 2003) as well as in the macro scale of the human brain (100 billion neurons) (Lizier et al., 2011). The mechanisms that facilitate this phenomenon have largely been attributed to neuromodulation and synaptic plasticity (Briggman and Kristan, 2008; Getting, 1989; Morton and Chiel, 1994). However, it is still unclear if neuromodulation is necessary for multifunctionality, and to what extent neural resources are shared across multiple behaviors.

The goal of this work is to show a concrete example of how the interaction between brain,

body and environment enables neural networks to perform multiple behaviors and elucidate the extent of neural reuse. The work presented in this chapter approaches this from two perspectives: information-theoretic and dynamical systems-theoretic. One of the most useful tools in understanding how neural networks process information is the ability to quantify information transfer by action potentials through the use of information theory (F. Rieke and Bialek., 1996). Transfer entropy (TE) (Schreiber, 2000b; Kaiser and Schreiber, 2002) is an information-theoretic measure that is being used extensively for estimating effective networks that emerge from task-specific interactions between the underlying structural network elements at different spatial and temporal scales (R. Vicente and Pipa., 2011; S. Ito and Beggs, 2011; Rubinov and Sporns., 2010; Nigam et al., 2016; Gourévitch and Eggermont., 2007a; M. Garofalo and Martinoia., 2009; N. Timme and Beggs., 2014; M. Shimono, 2015). Most of the work analyzing information processing in the brain has been experimental, focusing on three main techniques: fMRI, *in vitro*, and *in vivo* studies. While fMRI studies have yielded several insights into whole-brain organization, development and pathology, network nodes are defined at the macro or meso-scale where each node corresponds to hundreds or thousands of neurons (R. Vicente and Pipa., 2011; N. Kriegeskorte and Bandettini., 2006). *In vitro* studies have helped understand the micro-level structural organization and flow of information across small networks of neurons, but it is not possible to study neural dynamics in the context of behavior (N. Timme and Beggs., 2014; S. Ito and Beggs, 2011). Finally, *in vivo* studies make it possible to record micro-scale activity from behaving animals, but resolving the cortical network that is involved in a behavior and recording from all neurons involved in that particular behavior is not feasible yet. From a dynamical systems theoretic perspective, reuse in embodied recurrent neural networks unfold over three levels: structural network, autonomous dynamics of the neural network, and transient dynamics of the neural network. Structure is defined by the neural circuit itself, the intrinsic parameters of the neurons, and the synaptic strength of connectivity between them. While it is possible that an agent possesses specialized circuits for performing different behaviors, reuse at this level involves utilizing overlapping circuits to produce multiple behaviors. The next level, when structure is reused, is that of the neural network's autonomous

dynamics isolated from the body. Each behavior is associated with a set of phase-portraits corresponding to the inputs the agent experiences while performing them. The sets of phase-portraits (and the attractors therein) could be overlapping or could be unique to each behavior. The set of all attractors from all phase-portraits corresponding to a behavior are also referred to as the attractor set of the behavior in this paper. The third level of reuse is that of ongoing transient dynamics as the agent is in continuous closed-loop interaction with the environment. When there is attractor reuse from the previous level, it is possible that multiple behaviors navigate different transients around those attractors or, they might be reused too. In this work, the information theoretic analyses was employed to capture the overlap task-specific informational architecture, and the dynamical systems theoretic analyses was employed to study temporal aspects of reuse over the course of each behavior.

Over the past few decades, theoretical neuroscientists have begun to study computational models of networks of model neurons with the aim of understanding the relationship between neural activity and function. However, most of this work has focused on characterizing the dynamics of the abstract network, without any substantial connections to function (van Vreeswijk and Sompolinsky, 1996; N. Kopell and Traub, 2000; Brunel, 2000; E. S. Schaffer and Abbott, 2013). More recently, work has begun to focus on developing functional spiking neural networks (L. F. Abbott and Memmesheimer, 2016). However, this has been generated largely for disembodied networks: a network receives a time-series input, and its task is to generate a specified output (D. Thalmeier and Memmesheimer, 2016). These models address how specific patterns of neural activity are generated by sensory stimuli or as part of motor actions, but lack the continuous closed-loop interaction between the brain, body and an environment. Incorporating these components to address how a neural circuit works, requires us to develop behaviorally-functional spiking networks. The view that the body and the environment play a crucial role in the understanding of behavior is increasingly accepted (Chiel and Beer., 1997; Izquierdo and Beer., 2016) and there has been some work that has focused on understanding behavior through the development and analysis of whole brain-body-environment models (Beer, 2003).

Optimizing dynamical neural networks to perform two closed-loop tasks, object-categorization and pole-balancing, revealed that neural networks are capable of performing multiple tasks in the absence of neuromodulation or plasticity. Information-theoretic analysis of the multifunctional neural networks revealed that neural networks that reused the same structure such that the TE networks associated for each task was distinct, consistently performed the tasks well. However, the converse was not true; agents that did not have distinct TE networks per task also performed both tasks well. Furthermore, based on a dynamical systems theoretic analysis, we show for the first time to our knowledge that reuse of transients, namely indistinguishable neural activity, can produce starkly distinct behaviors in embodied dynamical neural networks. This can only be observed when the brain is studied in conjunction with the closed-loop interaction between the body and the environment.

4.2 Related work

The ability of neural circuits to dynamically reconfigure themselves to perform multiple functions has been studied under various contexts. Peter A. Getting introduced the term "polymorphic network" to denote networks that could be switched into different modes of operation based on a modulatory signal (Getting and Dekin, 1985; Getting, 1989; Marder, 1994). Kristan et al. (1988), based on their study of three different behaviors in the medicinal leech: the shortening reflex, crawling and swimming, demonstrate the presence of "multifunctional interneurons" that are active at different levels for each behavior and result in driving the appropriate central pattern generators Morton and Chiel (1994) describe "reorganizing circuits", based on experimental observations in crustacean stomatogastric (more on this later), as neural circuits where (a) single neurons switch from one pattern to another or (b) multiple neurons switch from one pattern to another or (c) two pattern generators fuse to form a new pattern generator or (d) many neurons form distinct pattern generators fuse to form one new pattern generator. Finally, Anderson (2010a) proposed a theory based on "neural reuse" as a principle for the functional organization of the brain. This was based on the idea that the brain's architecture is based on optimization to co-opt and reuse of neural

resources across tasks. Ultimately, albeit discussed under different but synonymous terminologies, the idea of functional reorganization of neural resources across tasks is pervasive in Neuroscience and is increasingly studied (Briggman and Kristan, 2008; Lurie et al., 2019). In the rest of this section I first discuss the different extents to which neural resources can be theoretically reused in a multifunctional circuit, followed by a discussion of existing work on different mechanisms that may enable multifunctionality and the computational models that demonstrate their feasibility, including experimental evidence that supports each of them.

4.2.1 Organization of multifunctional circuits

In order to study multifunctional neural networks it is useful to first make the distinction between anatomical and functional connectivity. Anatomical or structural connectivity refers to the physical synaptic connections and their connection strengths in a neural networks. On the other hand, functional connectivity refers to the connectivity inferred from interactions between neurons during the course of a behavior. Reuse of neural resources can be studied across both these levels and lies on a spectrum. In one end of the spectrum lies anatomical reuse, multiple tasks could be performed using distinct neural circuits. Alternatively, the anatomical neural networks could be overlapping i.e. some of the neurons might be active during both tasks (Figure 4.1A). This phenomenon has been observed in several circuits, for example several neurons that generate the pyloric rhythm in the stomatogastric ganglion of the crab are also involved in generating gastric mill patterns (Weimann et al., 1991; Katz and Harris-Warrick, 1991). Similarly, scratching and swimming in turtles (Berkowitz, 2002), and swimming and crawling in the medicinal leech (Briggman and Kristan, 2006a) has been shown to involve a shared pool of neurons while each behavior still having specialized neurons. Finally, the entire circuit might be reused across tasks, meaning all neurons would be active in multiple behaviors. This has been observed primarily in pattern generating circuits that have been observed to produce different oscillatory patterns under different conditions leading to different behaviors. For instance, in the sea slug, *Tritonia diomedia*, the same circuit drives independent muscles to perform swimming and crawling respectively (Popescu and

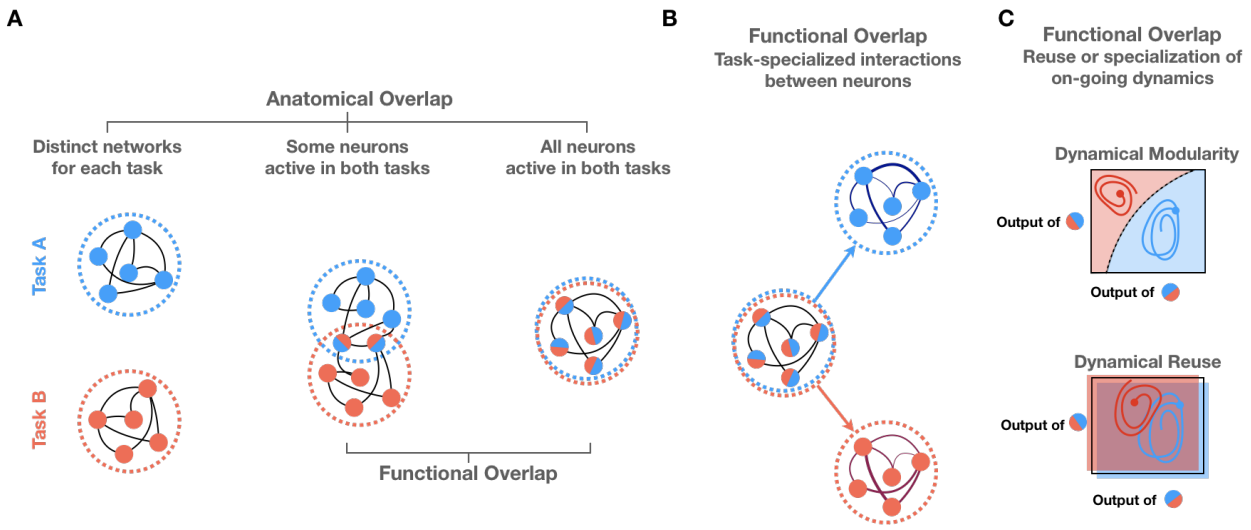

Figure 4.1: Extents of anatomical and functional overlap in neural networks. [A] At the anatomical level, different behaviors could be performed by distinct neural circuits; some of the neurons could be shared and reused across multiple behaviors; or all neurons could be active during multiple behaviors. The latter two conditions demonstrate a overlap off functionality amongst those shared neurons. [B] When there is anatomical overlap, it is often assumed that the circuit operates in distinct modes for each behavior. As we show in this chapter, this can be captured by the specialization in the interaction between neurons specific to each task. [C] Finally, the specialization in the interactions could arise from entirely distinct modes of operation denoting the presence of functional modules for each task. Alternatively, studying the dynamics over time might reveal an overlap in the dynamics of the circuit as it performs multiple tasks.

Frost, 2002); and the pre-Bötzinger complex in the brain stem has been shown to be responsible for different respiratory patterns such as eupneic breathing, gasps, and sighs (Lieske et al., 2000). These experimental observations shed light on the wide range of structural overlap that is possible in multifunctional networks based on which neurons are active during different behaviors.

The adoption of network science into Neuroscience provided a plethora of theoretical tools to study neural networks (Bassett and Sporns, 2017). One such tool that is especially useful for the study of multifunctional networks is the concept of functional networks. Estimating the functional connectivity network from neural activity captures the task-dependent dynamics in the neural network that emerges over the anatomical network (more on methods for estimating functional networks is discussed later). Thus, it enables us to go beyond identifying neurons that are active in multiple tasks, and investigate how the co-active neurons interact over the course of multiple tasks (Figure 4.1B). While functional networks have been estimated across multiple tasks in the

human brain most studies are focused on identifying the task-specialized regions of the brain (See Power et al. (2011) and Anderson et al. (2013) for an integration of such studies). Work that involves the study of overlap in the functional networks, while few, have shown interesting results. Language and executive functions have been shown to share functional networks in a way that can explain recovery from aphasia (Cahana-Amitay and Albert, 2014). Conversely, multifunctionality has been shown to arise in the "disgust" related regions of the brain in patients suffering from Obsessive Compulsive Disorder (Viol et al., 2019). In the work presented in this chapter, using a computational model we investigate the extent to which functional networks can overlap across tasks.

When there is anatomical overlap, from a dynamical systems theoretic perspective, there are two possibilities: the dynamics of the overlapping neurons could be in distinct parts of the state-space, dynamical modularity, or the dynamics could be reuse such that the ongoing neural activity is reuse across tasks, dynamical reuse (Figure 4.1C). While anatomical overlap has been widely studied in the literature, it is often assumed that anatomical overlap involves functional modularity. Perhaps the best example of such a study would be a computational model based study of legged-walking in Beer et al. (1999) and Chiel et al. (1999). They demonstrate that the different phases of a walking behavior such as swing and stance could be attributed to distinct dynamical modules, and that, performing the walking behavior involved switching between them based on intrinsic dynamics or environmental feedback. In the work presented in this chapter, using a computational model of multifunctional neural network we demonstrate that neural reuse occur down to the level of dynamical reuse - no modularity at any level.

4.2.2 Mechanisms for multifunctionality

In this section, I discuss mechanisms that have been attributed to the emergence of task-specific neural dynamics in anatomically overlapping networks. The three primary mechanisms that have been attributed to enabling multifunctionality are: switching behaviors based on the activation, or lack thereof, of an interneuron; neuromodulation; and environmental feedback. While the first two

have been widely studied in experimental neuroscience, the ability of environmental feedback to enable multifunctionality even in the absence of other plasticity mechanisms has not been studied as much.

Multifunctionality through switching neurons

A multifunctional circuit's dynamics can be altered by the activity of projection neurons that can selectively inhibit/excite neurons in the multifunctional pool of neurons. For instance, the stomatogastric ganglion (STG) of crabs exhibits two kinds of gastric mill rhythms, type I and type II, that has been shown to be modulated by projection neurons CG and GI. Type I oscillations have been observed when CG inhibits particular STG neurons (Simmers and Moulins, 1988) and type II oscillations have been observed due the combined activation of CG and GI which results in disinhibition of the STG neurons (Combes et al., 1999b). While this example shows evidence of multiple projection neurons acting in parallel, this is not always the case. In the tadpole of the frog *Xenopus*, oscillatory activity in a group of neurons that govern swimming and slower oscillations in the same neuron observed during struggling are mediated by a single projection neuron. While activation of the projection neuron elicited swimming, prolonged activation results in the struggling behavior (Soffe, 1997). Similarly, in zebrafish, the level of activity of a single neuron results in the activation one of two different central pattern generating circuits that lead to escape versus swimming (Ritter et al., 2001). Thus, different interactions between the multifunctional circuit and its projection neurons have been shown to reliably elicit multiple behaviors.

Multifunctionality through neuromodulation

Perhaps the most studied mechanism that enables the same circuit to perform multiple behaviors is neuromodulation. As the name suggests, neuromodulators act on neural circuits effectively changing the dynamical landscape over which the circuit is operating. Thus, the same network can perform different tasks as they operate under the influence of different neuromodulators or in their absence. Two neuromodulators, serotonin and dopamine have been shown to enable switching be-

tween swimming (Nusbaum and Kristan, 1986) and crawling (Crisp and Mesce, 2004) respectively in the leech. While serotonin promotes swimming, dopamine terminates swimming and initiates crawling in the "swim" central pattern generating neurons of the leech. As such, this is an example of two different neuromodulators acting on the same circuit to elicit different behaviors. Additionally, studies of the stomatogastric ganglion (STG) of crabs has yielded several insights into the role of neuromodulation in multifunctionality (Swensen and Marder, 2000). Specifically, the STG produces three different kinds of rhythms: pyloric, gastric mill and cardiac sac rhythms (Hooper and DiCaprio, 2004; Marder and Bucher, 2001). It has been shown that the proctolin, released from an interneuron identified as MPN, can initiate or speed up pyloric rhythms (Nusbaum and Marder, 1989a,b). While the pyloric rhythm is controlled by the neuromodulator, the same interneuron MPN inhibits gastric rhythm (Blitz and Nusbaum, 1997, 1999) indicating that multifunctionality can emerge from the convergence of multiple mechanisms on the same neural circuit - in this case, transmitters and neuromodulators.

Multifunctionality through brain-body-environment interaction

An animal that is embedded in an environment is involved in a closed-loop interaction with it. The environmental feedback that an animal receives in response to its actions is constantly influencing its behavior. Consequently, the feedback an animal would receive in one task is characteristic of that task and is qualitatively distinct from feedback in a different behavior. Based on this, it can be posited that the same circuit could perform multiple behaviors purely due the difference in closed-loop environmental feedback eliciting different dynamics. Importantly, this can happen in the absence of biophysical mechanisms such as neuromodulation. Multifunctionality emerging from animal-environment interaction is the focus of this chapter.

Most of neuroscience research has only focused on one of the two mechanisms discussed above. Perhaps the most significant support for this third mechanism comes again from studies of the crab STG. As discussed previously, two gastric mill rhythms are elicited via the activation of project neurons CG and GI. However, these projection neurons are in turn influenced by the mechanosensory

neuron, AGR (Combes et al., 1999a). Thus effectively, the the switching between the gastric mill rhythms is due to mechanosensory neurons that feed into the overall gastric circuit, which can be construed as environmental inputs to the STG. Inspired by multi-stable central pattern generating circuits in the crab, there have be computational modeling studies that have replicated the different rhythms that would lead to swimming or walking using neural network models. Specifically, a coupled-oscillator model was demonstrated to show newt-like swimming and walking rhythms based on input from the stretch-receptors and limb proprioception (Bem et al., 2003).

The work presented here builds on previous work using brain-body-environment computational models for multiple tasks by developing a computational model of a brain-body-environment system that performs multiple behaviors using the same sensory and motor capacities. In the work by Izquierdo and Buhrmann (2008), the same neural network without any changes in parameters was shown to perform two qualitatively different behaviors while placed in two different bodies. Williams and Beer (2013) showed that when different motor systems are used for different tasks, the qualitative difference in environmental feedback drives the same network differently to produce different behaviors. Agmon and Beer (2014) presented a model where different sensory apparatus in the agent, sensitive to different stimuli, performed different associated behaviors using the same motor control systems. In these models, although the neural network remained the same, the body was changing. The work presented in this chapter, uses the same sensory and motor control mechanisms for two tasks - object categorization and pole-balancing. Using a combination of information-theoretic and dynamical systems theoretic analyses, we show that while tasks can have task-specialized dynamics on a time-aggregated metric, reuse to the level of transient dynamics can be observed when the brain, body, environment and their interaction are taken into account.

4.2.3 Estimating task-specific effective networks

Irrespective of the mechanisms that may underlie multifunctionality, the primary feature of interest of multifunctional neural circuits is their dynamic reorganization to perform multiple functions. Estimation a directed network of associations between the neurons in the network that based on their

interactions during the course of a task allows us to capture the specific reorganization associate with that task. This is the task-specific effective network. The effective network arising from the same anatomical network based on interaction during another task can be estimated. Comparing the two task-specific effective networks emerging from the qualitatively different dynamics in the same circuit during different behaviors gives us the extent to which neural resources are reused across tasks. Several methods exist to estimate effective networks from neural activity, some of which are point-process General Linear Models (Paninski, 2004), Dynamic Causal Modeling (Friston et al., 2003), Granger causality (Granger, 1969), and Transfer Entropy (Schreiber., 2000). In this section, I review the approach utilized in this work, Transfer Entropy. Note that the other recent methodology for measuring information transfer using partial information decomposition has already been discussed in the previous chapter.

Transfer Entropy

Information theory, ever since its introduction by Shannon (Shannon and W., 1949) has been employed in a variety of fields ranging from neuroscience to the social sciences. Transfer entropy (TE) was proposed by Schrebier as a model-free information theoretic method to find the amount of information flowing from one random process to another (Schreiber., 2000). In its simplest form, the transfer entropy from a source process X to a target process Y is a measure of the reduction in the uncertainty in predicting future values of Y given current value of X, over predicting future value of Y purely based on Y's current value.

$$TE_{X\to Y} = \sum_{x_t, y_{t+1}, y_t} p(x_t, y_t, y_{t+1}) log(\frac{p(y_{t+1}|y_t, x_t)}{p(y_{t+1}|y_t)}) \tag{4.1}$$

TE has been used for primarily two kinds of applications in neuroscience - the TE between a stimulus and neural activity gives the amount of information about the stimuli encoded by the neurons; TE between spiking activities of neurons in a network gives a picture of the information flow through a circuit. Among the first studies that used this method on spike trains was to identify the neural coding of auditory stimulus in cats by Gourévitch and Eggermont. (2007b).

This approach used different sized temporal binning of spiking data, thereby giving the number of spikes in a period $[t-\tau, t)$ in a bin of size τ. Applying TE between auditory stimuli and this data over different values of τ revealed that auditory stimuli were encoded in cortical neurons in 2-15ms time windows.

TE has also been applied in the visual cortex by Besserve et al. (2010), where TE between stimuli and the different frequency bands of the field potentials were associated to show that the gamma band has most TE suggesting its involvement in stimuli processing. In this work, the TE formulation shown in equation 4.1 was modified to not necessarily use the previous time bin, but any arbitrarily delayed time bin albeit only one time bin was used at a time.

While the aforementioned methods have used TE to decipher neural coding of stimuli it has been most widely used for studying information flow through neural networks. TE has been applied at the macro scale to estimate directed connectivity from BOLD signals to identify functional associations over the known structural connections in macaque neocortex (Honey et al., 2007). This study, that came out almost at the same time as Gourévitch et. al.'s work, validated TE using a baseline generated by averaging TE values from multiple time-offsets of the data rather than a random shuffling. Lizier et al. (2011) applied a multivariate extension of this method to fMRI data as the subject was performing a variety of sensorimotor tasks and found task relevant changes in the effective network. Their method involved $TE_{\{X_1, X_2..X_n\} \to Y}$ and also included higher order terms for each covariate, i.e. it included terms over a window of k time bins rather than just one time bin. While, they proposed this in their methods, they used a simplified version in order due to its computational complexity. This simpler version involved down sampling the k time bins and hand-picking the covariates. Following this, Vincent and Wibral's group used higher-order TE with delays, albeit univariate, and applied it to magnetoencephalography (MEG) data (Wibral et al., 2011; Vicente et al., 2011). They also made a toolbox available for this purpose called the TRENTOOL (Lindner et al., 2011) which primarily caters to analog neural data such as MEG.

All these modifications proposed to TE from its original definition in equation 4.1 has mostly been made to address two specific drawbacks in the original definition. Firstly, it does not capture

varying synaptic delays since values only 1 time bin in the past are being taken into account. Secondly, since it only considers the activity of the source or the target's own past in only one time bin, effects that involve the combined information from multiple time bins such as bursts cannot be captured. It has been shown that synaptic delays in biological neural networks range between 1-20ms (Swadlow., 1994, 1985; Ferster. and Lindstrom, 1983) and also that bursts are a robust mode of communication between neurons (Lisman, 1997; Eugene M. Izhikevich and Hoppensteadt., 2003). However, incorporating these into TE increases its computational complexity exponentially. For a k self-history order and l source history order there are 2^{k+l+1} combinations of states over which the probabilities have to be computed for TE estimation. In order to account for these in a computationally efficient manner, Ito et al. (2011b) presented a MATLAB toolbox for TE estimation from binary spiking data. The computational efficiency comes from the sparsity of the binary data, where only time bins that had spikes were processed which saved time spent on checking empty bins. The higher-order TE that computes over various delays is given by

$$TE_{X\to Y}(d) = \sum p(y_{t+1}, y_t^l, x_{t+1-d}^k) log_2 \frac{p(y_{t+1}|y_t^l, x_{t+1-d}^k)}{p(y_{t+1}|y_t^l)} \tag{4.2}$$

where d is the delay over which TE is computed and k and l define the order on the self-history of process Y and order on the history of source X respectively. The delay term accounts for various synaptic delays and the different orders accommodate for information over multiple time bins. This method involves recomputing the transfer entropy at several delays and different orders. This results in a number of TE estimations from which the most appropriate one has to be chosen. A straightforward option is to choose the maximum TE and corresponding delay, or the peak TE (TE_{pk}). Ito et. al. also used a Coincidence Index (Chiappalone et al., 2006; Tateno and Jimbo., 1999; Juergens and Eckhorn., 1997) which was defined as follows

$$CI = \frac{\sum_{d=t_p-\tau/2}^{t_p+\tau/2} TE(d)}{\sum_{d=1}^{T} TE(d)} \tag{4.3}$$

where t_p is the time at which TE_{pk} occurred, T is total time of the window of all delays for which TE

was estimated, and τ is the coincidence window size. The numerator in this RHS of this equation is an approximation of the area under the TE(d) vs d curve around a window of size τ centered at t_p and the denominator is the area under the curve over the whole duration T. Therefore, CI is the ratio of the area within τ of the peak and this is taken as a measure of the strength of the connection between two neurons, denoted by TE_{CI}. In their work, Ito et. al. used $T = 30ms$, and tested $\tau = 1, 3, 5, 7ms$. The intuition behind this measure is as follows - when network activity is in bursts, there are a lot of neurons that fire coincidentally, which may be captured as false positives in the TE analysis. However, these are in fact due to the common network drive and not due to synaptic drive and there is a temporal difference between these two kinds of drives which is captured by CI (Beggs and Plenz., 2003). Common drives show up distributed over several delays in TE estimates thereby making a plot of TE(d) vs delay rather flat. Synaptic drives, on the other hand are quite reliable in their delay and hence produce a peak in the same plot corresponding to the actual delay of the synapse. Now, it can be seen that CI measures the peakiness of this plot and hence controls for network drive in TE estimates.

While Ito et. al. applied a threshold on CI to pick valid connections (strength of connection was given by TE_{pk}), Shimono et. al. used a shuffling method to propose a more systematic way to identify the significant connections (Shimono and Beggs., 2014). Spikes of the source neuron were jittered such that firing rate was preserved and only the times were changed. This disconnected the source from the target and not jittering the target neuron meant that the self-predictability of the target remained the same. The TE was calculated for several such jittered datasets and $log(TE_{pk})$ vs CI for the jittered and original dataset were plotted overlapping each other. The objective was then to identify the region where connections from original dataset did not overlap with the jittered data. It can be seen that these connections are expected to be in the high $log(TE_{pk})$ and high CI region of the graph (top right corner of the graph). This is because connections that are extremely specific in their delays (very peaky CI plots) and also have high values of transfer entropies are expected to be true connections. A decision boundary was defined enveloping the points from the jittered datasets along the top right corner and all points from the original dataset beyond this

boundary were deemed significant connections. Analysis of the TE network of few hundred neurons revealed specific structure in cortical functional connectivity such as hubs, communities. The same method applied to cortical cultures and calcium imaging data revealed a rich club topology in the microconnectome (Nigam et al., 2016).

One of the main advantages of this measure is that it is model-free unlike the previous methods discussed here. Further, it is computationally more tractable than any GLM method because the latter requires that a model be optimized to fit the data from each neuron but the former is purely data driven and only requires the computation of probabilities. However, sufficient amount of data is required to estimate these probabilities reliably. Further, using transfer entropy it is difficult to identify inhibitory connections because an absence of a spike in a sparse dataset, that is typical of cortical spiking data, could be simply due to the inherent sparsity of firing or due to inhibition and that is difficult to disambiguate.

As a measure of effective network connectivity, although Transfer Entropy has been most widely used in neuroscience, it has been used in evolutionary robotics in some instances - synchronization dynamics of neurons and its influence on evolvability and behavior has been studied, and TE has also been used in analysis of neural networks in supervised and unsupervised learning contexts (Moioli and Husbands., 2013).

The work discussed in this chapter employs Transfer Entropy to infer the effective networks from neural activity recorded in a computational model of a multifunctional neural network. The effective networks associated with each task is compared to study the extent of reuse of neural resources across the task.

4.3 Methods

In the agent-environment models used throughout this work, the agents were modeled using dynamical recurrent neural networks. The parameters of the neural network were optimized using an evolutionary algorithm such that it was able perform the required task. In this section, we specify implementation details about the neural network model, the tasks, and the optimization algorithm.

4.3.1 Agent design

The agent is circular with a diameter of 30 units. It can move back and forth along a one-dimensional axis, and its behavior is controlled by a 3-layer neural network (Figure 4.2A). The network architecture consists of seven sensory neurons fully connected to N fully interconnected interneurons, which are in turn fully connected to two motor neurons.

There are 7 sensory neurons in the top layer which are stimulated by the agent's vision ray. They follow the state equation:

$$\tau_s \dot{s}_i = -s_i + K_i(x, y) \quad i = 1, \ldots, 7 \tag{4.4}$$

where s_i is the state of sensory neuron i, τ_s is the time constant that is shares across all sensory neurons, $K_i(x, y)$ is the sensory input from the i^{th} ray due to an object at location (x, y) in agent-centered coordinates, and the dot notation over the state variable indicates the time differential $\frac{d}{dt}$.

The seven sensory neurons project down to a middle layer of N fully interconnected Izhikevich spiking neurons with the following two-dimensional system of ordinary differential equations (Izhikevich, 2003):

$$\dot{v}_i = 0.04v_i^2 + 5v_i + 140 - u_i + S_i + I_i \quad i = 1, \ldots, N \tag{4.5}$$

$$\dot{u}_i = a(bv_i - u_i) \tag{4.6}$$

with the auxiliary after-spike resetting

$$\text{if } v \geq 30\text{mV}, \quad \text{then} \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases}$$

with each interneuron receiving weighted input from each sensory neuron:

$$S_i = \sum_{j=1}^{7} w_{ji}^{s} \sigma(s_j + \theta_s) \tag{4.7}$$

and from other spiking interneurons:

$$I_i = \sum_{j=1}^{N} w_{ji}^{i} o_i \tag{4.8}$$

where v_i is representative of the membrane potential of spiking neuron i, u_i represents its membrane recovery variable, w_{ji}^{s} is the strength of the connection from the j^{th} sensory neuron to the i^{th} spiking interneuron, θ_s is a bias term shared by all sensory neurons, $\sigma(x) = 1/(1 + e^{-x})$ is the standard logistic activation function, w_{ji}^{i} is the strength of the recurrent connections from the j^{th} to the i^{th} spiking neuron, and o_i is the output of the neuron: 1 if $v_i \geq 30$mV, and 0 otherwise. The sign of all outgoing connections from an interneuron depend on its excitatory or inhibitory nature, as identified by a binary parameter. Parameters a,b,c and d control the type of spiking dynamics. For inhibitory neurons $a \in [0.02, 0.1]$,$b \in [0.2, 0.25]$,$c = -65$ and $d = 2$, whereas for excitatory neurons $a = 0.02$, $b = 0.2$, $c \in [-65, -50]$ and $d \in [2, 8]$.

Finally, the layer of interneurons feeds into the two motor neurons, with the following state equation:

$$\tau_m \dot{m}_i = -m_i + \sum_{j=1}^{N} w_{ji}^{m} \bar{o}_j \quad i = 1, 2 \tag{4.9}$$

$$\bar{o}_j(t) = \frac{1}{h_j} \sum_{k=0}^{h_j} o_j(t - k) \tag{4.10}$$

where m_i represents the motor neurons, w_{ji}^{m} is the strength of the connection from the j^{th} spiking interneuron to the i^{th} motor neuron, $\bar{o}_j$ represents the moving average over a window of length h_j for the output of spiking interneuron j.

Finally, the difference in output between the motor neurons results in an instantaneous horizontal velocity that moves the agent in one direction or the other.

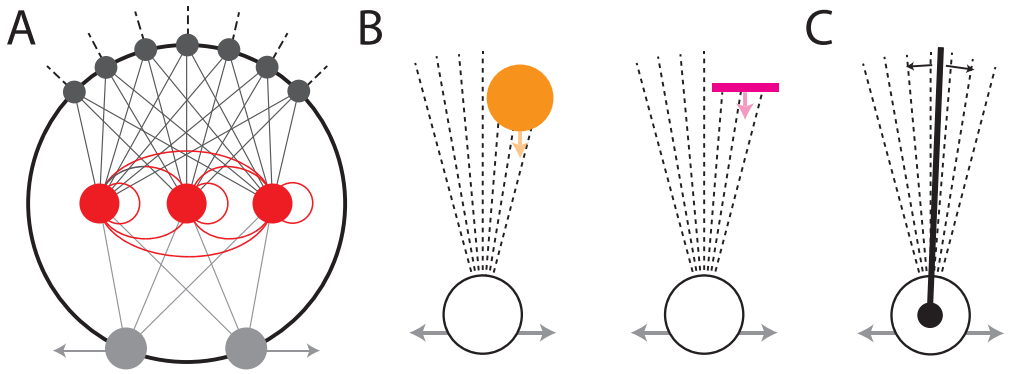

Figure 4.2: Agent design and task setup. [A] 7 rays of vision feed into sensory neurons (black). These neurons are fully connected to the recurrent interneuron layer (red), that in turn feed the left and right motor neurons (grey). [B] Categorization task with circle and line trial. The falling object needs to be caught if it is a circle and avoided if it is a line. [C] Pole-balancing task. The pole attached to the agent's center is expected it to keep balanced within the rays.

4.3.2 Object categorization

We replicated the categorization task first introduced in Beer (1996). The task involves discriminating between the shape of falling objects (circles and lines) by moving towards one (circles) and away from the other (lines) (Fig. 4.2B). The circles' diameter and the line's length were both set at 30 units. To encourage generalization, each evaluation of the agent's performance was conducted over 8 different trials for each type of object, with the objects' initial horizontal offset from the agent uniformly distributed in the range $[-50, 50]$. The objects fall with a constant velocity of 0.3 units per second. Performance in this task was quantified by averaging over $1 - |d_i|$ for the circle trials and $|d_i|$ for the line trials; where d_i is the normalized distance between the center of the agent and the center of the object when the vertical offset between the agent and the object reaches 0 (offset of over 45 units was clipped at 45).

4.3.3 Pole-balancing

We adapted the original pole-balancing task (Barto et al., 1984) such that the agent has the pole attached to its center and senses it through the same rays used for sensing the falling objects (Fig. 4.2C) (Vasu and Izquierdo, 2017c). The sensory input, as the pole sweeps across a ray at angle ϕ, increased linearly from 0 at $(\phi - 1)$ reaching the maximum value at ϕ and falling back to

0 at (ϕ + 1) and vice verse. Note that the agent sensed the pole only when it intersected a ray but it "disappeared" from view while passing between rays. The pole was considered dropped if it fell beyond the rays, or if the agent moved farther than 45 units on either side from its starting position. Performance in this task was calculated by averaging cosine($\theta * 6$) at each time step of the 500s evaluation duration, where θ is the pole angle with the vertical axis. To promote generalization of the behavior, performance was averaged over 16 trials with the pole starting from 4 different angles on either side of the agent in $[-9, 9]$ with angular velocity -0.1 or 0.1.

4.3.4 Evolutionary optimization

An evolutionary search algorithm was used to optimize the parameters of the agent: time-constant, gain and bias for sensory neurons (3), weights from sensory layer to N-interneurons ($N*7$), recurrent weights between interneurons (N^2), bias and time-constant for each interneuron ($2N$), weights from interneurons to motor neurons ($2N$) and gain, bias and time-constant for motor neurons (3): totaling $D = 3 + 7N + N^2 + 2N + 2N + 3$ parameters. A search started with a random population of 100 solutions encoded as D-dimensional genotype vectors with each element in [-1,1]. These elements were scaled and mapped on to the different parameters to build the agent. Gains are scaled to be in [1,20], time constants in [1,2], biases in [-4,4] and all weights were scaled to be in [-5,5]. The fitness of agents was evaluated based on their performance in each task. Based on fitness, an elitist fraction of the top 4% solutions were retained while their copies were subject to a Gaussian mutation noise with mean 0 and variance 0.3 to produce a new population of solutions. This was repeated for a fixed number of generations.

Since optimization is stochastic, 100 independent runs were carried out for the single and multi-task scenarios. For the individual tasks, optimization was carried out for 1000 generations in each run. In the multi-task setting, these experiments were conducted in three different task presentation paradigms: (1) evolved for both categorization and pole-balancing for 2000 generations, (2) evolved only for pole-balancing for the first 500 generations, and then evolved for both tasks for 1500 generations and (3) evolved for categorization for the first 1000 generations, and then for both tasks

for another 1000. The 500 generation limit for paradigm 2 and 1000 for 3 was based on the number of generations required to acquire good performance in each task when optimized individually. Agents were reset between all trials of all tasks. In the multifunctional cases, the product of the individual task fitnesses was used as opposed to sum or average because it guarantees good performance in both tasks, while still keeping the fitness in $[0, 1]$. All three optimization paradigms gave similar results.

4.3.5 Transfer Entropy

Transfer entropy (TE) is an information-theoretic measure which has received recent attention in neuroscience for its potential to identify effective connectivity between neurons (Wibral et al., 2014b). Intuitively, TE from neuron J to I is a measure of the additional information provided by the activity in neuron J over and above the information from I's own history of activity that helps predict the activity of neuron I. This proportional increase, when high, corresponds to a causal influence of J over I. TE is especially useful in taking into account non-linear interactions between neural units and provides a directed measure of influence from one neuron to another. Further, due to synaptic delays, the causal influence of one neuron over another can only be detected if tested at that corresponding delay. In order to account for this, a modified version of TE was proposed by S. Ito and Beggs (2011). We used the MATLAB toolbox provided by the these authors to estimate TE. This involved utilizing the history of neuron J over different time delays to predict the future activity of I and then picking the peak-TE over all delays, as follows:

$$\mathrm{TE}_{J\to I}(d) = \sum p(i_t, i_{t-1}, jt-d)\log_2 \frac{p(i_t|i_{t-1}, j_{t-d})}{p(i_t|i_{t-1})} \tag{4.11}$$

where i_t denotes a binary spike/no-spike activity of neuron I at time t, where j_t denotes a binary spike/no-spike activity of neuron J at time t, d denotes the synaptic time delay, $p(x)$ is the probability of that set of spiking events occurring at that particular times, and $p(x|y)$ is the conditional probability that a set of spiking events occur given that certain other events have occurred.

4.3.6 Cluster Specialization Coefficient

Successful agents were subject to TE analyses on a trial-by-trial basis; the TE network was estimated for one trial of falling line or circle. This produces 48 TE networks from the 24 circle and 24 line trials. These networks are then hierarchically clustered using an agglomerative clustering algorithm in MATLAB. This procedure produces a clustering tree diagram where the leaves of the tree correspond to one of the 48 networks. The leaves are successively connected to one another depending on how close they are in TE network space, until they are all in one cluster. These trees are called dendrograms. For each successful agent we produced a trial-by-trial TE dendrogram.

In order to quantify the task-specificity of the TE networks, we measured the number of clusters that were unique to each task. We cut the dendrogram at different places to partition the 48 networks into multiple clusters. Then the members of the cluster were sorted based on task. We identified the smallest number of clusters that can be formed, such that all the networks in the cluster correspond to networks that were inferred from the neural activity of one and only one of the tasks. These clusters are called task-specialized clusters. For example, the dendrogram in Figure 4.6C can be dissected at level 1, to partition the data into two task-specialized clusters, each containing only networks corresponding to one task. Further dissection of these clusters will yield more number of task-specialized clusters but the minimum number of task-specialized that can be formed is 2. CSC is defined based on the ratio between the minimum number of task-specialized clusters to the maximum number of task-specialized clusters that can be formed. The maximum number of task-specialized clusters that can be formed is equal to the number of networks i.e. one network per cluster. Therefore, the CSC for an agent is defined as CSC $= 1 - c_{\text{Min}}/c_{\text{Max}}$, where c_{Min} is the minimum number of task-specialized clusters and c_{Max} is the maximum number of task-specialized clusters.

4.4 Fixed neural circuits perform multiple behaviors without neuromodulation or plasticity

The highest level of reuse is that of overlap in the anatomical circuit; an agent performing more than one behavior could acquire specialized circuitry to perform each behavior or could share

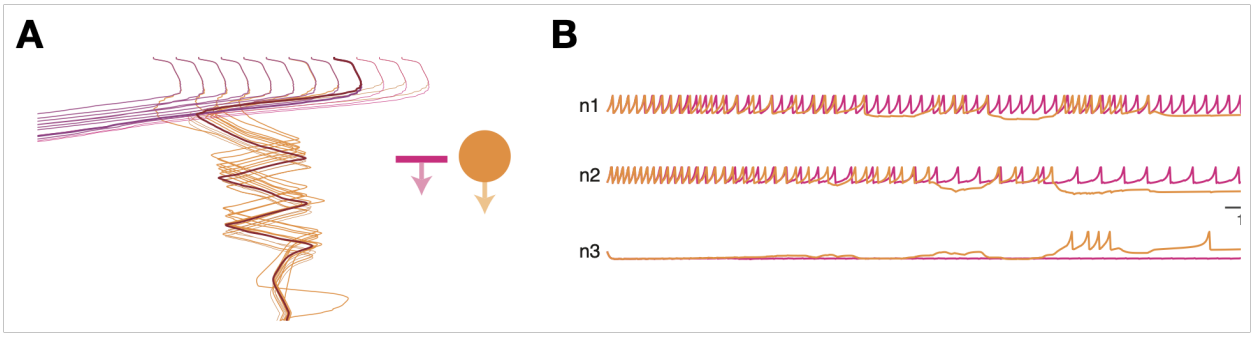

Figure 4.3: Behavior of the best multifunctional Izhikevich agent from 100 runs. [A] Behavior. The horizontal offset between the agent and the falling object is shown along the horizontal access. The vertical axis denotes the time. As the object falls, it can be seen that the orange traces (representing the circles) end up with 0 offset, meaning the the agent "caught" them and vice versa for lines. Only 24 of the 48 trials are shown for clarity. Two traces are highlighted. [B] Neural traces for example runs highlighted in [A]. The neural activity is similar at the beginning of the object's fall. Note that neuron 3 does not show any activity for the line avoiding task but does spike during the circle catching task.

neural circuits between them. In order to test this, using an evolutionary optimization approach, we evolved networks of different sizes to perform the two categorization behaviors of circle-catching and line-avoiding. 100 independent evolutionary runs optimizing agents controlled by a fully recurrently connected Izhikevich interneuron layer yielded solutions with neural networks as small as 2 Izhikevich neurons that could perform both tasks. The best 3-interneuron agent that achieved a performance of 97.8% and was able to successfully catch the circles while avoiding the lines. This agent perfectly differentiated the circles and the lines by "catching" all circles and moving away from all the lines (Figure. 4.3A). The horizontal offset distance between the agent and falling object as it approaches the agent is 0 for all circles and is large for all lines. The fitness of the agent is not a perfect 100% simply because that would entail catching the circles at exactly the center of its 30-unit wide body. Visualizing the corresponding spiking neuron activity (Figure 4.3B) shows that the inhibitory neurons ($n1$ and $n2$) have a higher firing rate than the excitatory neuron. Importantly, in this agent, only two of the three neurons were active during the line-avoiding task and all three interneurons were active for the circle-catching task. Thus, this multifunctional circuit involved a sharing of neurons such that the latter task is performed using entire circuit of the former plus an additional neuron. It is to be noted that the agent had no external signal indicating which task

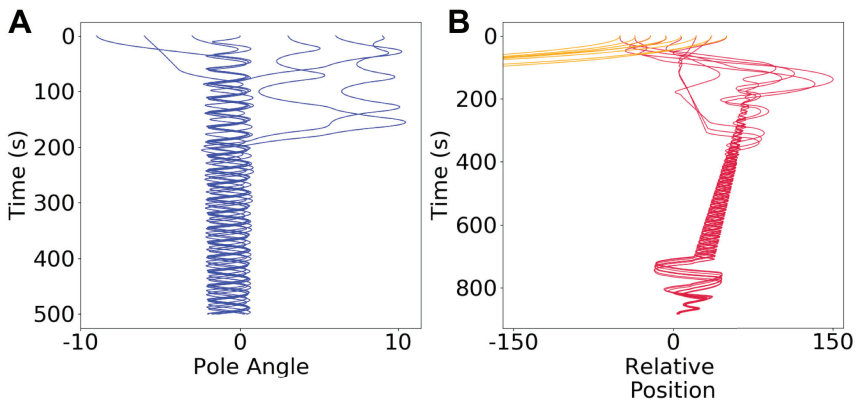

Figure 4.4: Behavior of the best multifunctional CTRNN agent from 100 runs. [A] The agent was able to bring to center and balance the pole starting from different pole angles. [B] The same agent, using the same neural network was also able to catch the circles (red) while avoiding lines (orange).

to solve. Furthermore, the agent also had no synaptic plasticity or neuromodulatory signals that could be responsible for re-configuring the circuit for the different tasks. It is in these Izhikevich neuron multifunctional agents that effective network analysis using transfer entropy was performed to analyze the extent of functional overlap.

Following the study of functional overlap using effective networks, temporal analysis of the reuse of ongoing neural activity was performed. Being continuous-valued (representing firing rate), since CTRNNs are most conducive to a dynamical analysis, we utilized agents with CTRNN interneuron layers. Furthermore, since the best multifunctional agent demonstrated only partial overlap of the anatomical circuit, the CTRNN agents were optimized to perform the pole-balancing task in addition to the circle-catching and line-avoiding tasks. Over 100 independent evolutionary runs, CTRNNs no larger than the ones that could solve the individual tasks could also solve both tasks. The best 2-interneuron multifunctional CTRNN agent could perform categorization with a fitness of 95.8% and pole-balancing with a fitness of 95.4% (Figure. 4.4). The optimization scheme that led to this agent was composed of evolving for pole-balancing for the first 500 generations, followed by evolving for both tasks. This agent used the same circuit to successfully catch circles while avoiding lines and also balance a pole. Observing the neural traces of the interneurons shows that these multifunctional agents used fully-overlapping anatomical networks to perform

circle-catching and pole-balancing (Figure. 4.11).

In order to validate the multifunctional nature of the neural networks, we first systematically explored the minimal resources required to solve each task individually. 100 independent evolutionary runs were performed for CTRNNs of different sizes for each task. The smallest CTRNN that could perform pole-balancing had 2 interneurons. The best of these agents achieved 98.44% fitness and was able to move the pole to its upright position from a broad range of initial positions and keep it balanced for an extended duration of time (Fig. 4.5A). The smallest network that could perform the categorization task also had 2 interneurons. The best of these agents had a fitness of 98.5% and was able to successfully catch all circles and avoid all lines falling from the full range of starting positions (Fig. 4.5D). In order to address multifunctionality using these two tasks, it is important to demonstrate that they indeed require their own set of sensorimotor transformations. In other words, circuits that solve one task, should not be able to solve the other task, and vice versa. To demonstrate this, all agents that were optimized to perform one task were evaluated on the other. Agents that were trained to balance the pole were as good as random agents at the categorization task (Fig. 4.5D,E). Agents that were trained to categorize could balance the pole only slightly better than random controllers (Fig. 4.5C,E). This suggests each task requires its own unique set of sensorimotor transformations and that ultimately solving one task does not guarantee good performance in the other.

4.5 Effective network clusters show task specialization

When anatomical networks overlap, the same neurons could be interacting differently in each task (Figure 4.1B). In order to answer this question, we consider the circle-catching and line-avoiding tasks and analyze multifunctional Izhikevich neural networks that were optimized to perform both tasks. Transfer entropy analysis of the multifunctional Izhikevich agent reveals that different behaviors (e.g., catching circles versus avoiding lines) have their own specialized effective networks. In other words, the effective networks for minor variations of one task (circle catching) are more similar to each other than the effective networks of minor variations of the other

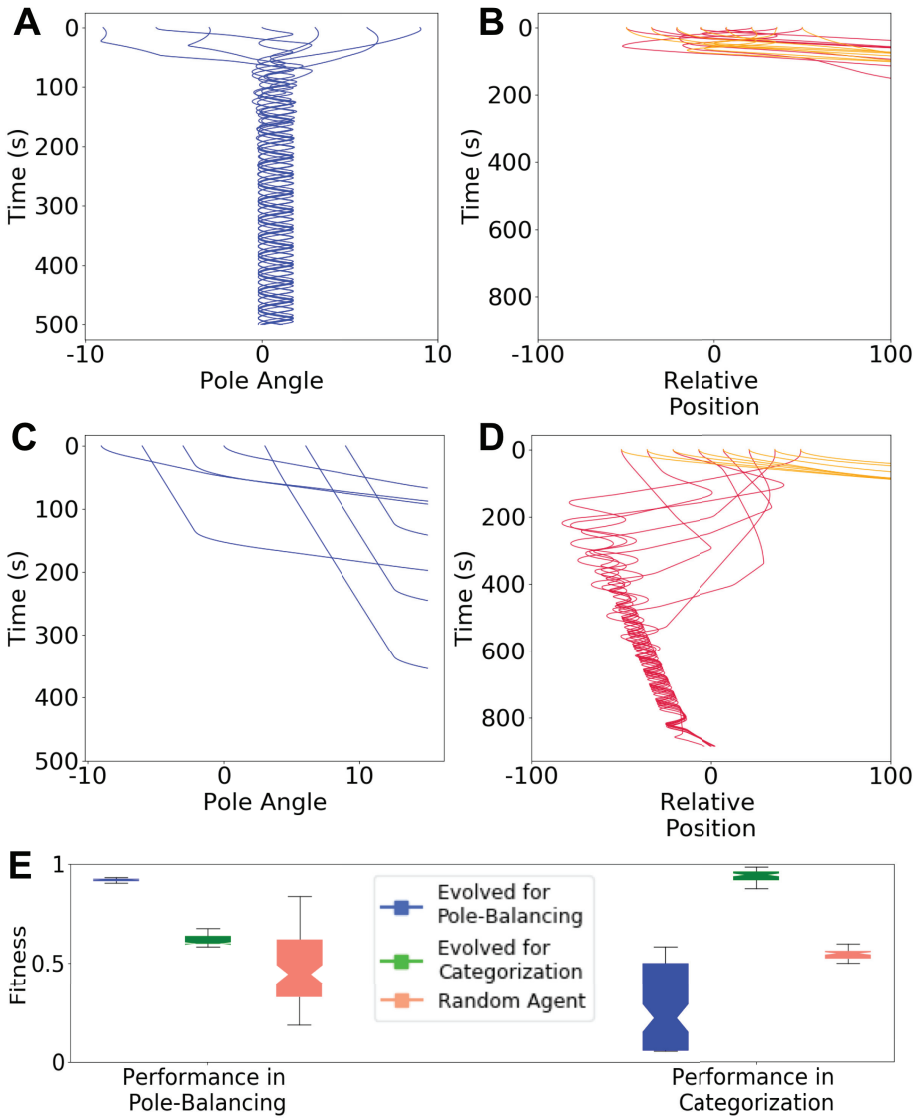

Figure 4.5: Behavior and performance on individual tasks. [A] Best agent from 100 runs of optimizing for pole-balancing alone. The agent was able to bring the pole to the center and keep it balanced from different initial pole angles. [B] Best pole-balancer shown in A is unable to categorize circles (red) from lines (orange) and avoids both. [C] Best agent from 100 runs evolved for categorization alone is unable to balance the pole. [D] Best categorization agent shown in C, demonstrating its ability to catch circles (red) while avoiding lines (orange). [E] Optimizing for one task results in performance similar to a random agent on the other task. Fitness distribution from 100 runs of agents evolved for pole-balancing (blue) in pole-balancing and categorization, and similarly that of the agents evolved for categorization only (green) in pole-balancing and categorization, and random agents (salmon) on both tasks.

task (line avoiding). Further, having task-specific effective networks was indicative of the agent's performance. More generally, the degree of task-specificity of the effective networks correlated with behavioral performance.

We selected the highest fitness individual (fitness=97.8%) over all 100 runs from the network size $N = 3$ batch to analyze first. This agent perfectly differentiated the circles and the lines by "catching" all circles and moving away from all the lines (Figure 4.3A). In order to probe the difference in neural dynamics between the circle catching and line avoiding tasks, the agent's neural activity from each of the 48 trials of falling lines and circles was recorded and analyzed. We used Transfer Entropy to estimate the task specific effective network for each trial. Since TE quantifies information transfer in the comparable units of bits, we can directly compare TE networks from neural activity recorded while the agent is performing different tasks. The effective network estimated from one line-avoiding trial (Figure 4.6A) shows that $n3$ is not part of the network. This can be reconciled with the fact that neuron 3 was completely dormant during this task. The effective network for one circle catching trial (Figure 4.6B) on the other hand, shows a fully connected network. Spiking activity during one of the circle catching trials shows that the neural activity in $n1$ and $n2$ control the scanning behavior of the agent, while $n3$'s activity coincides with the agent's fine motor control at the end of the trial when it needs to center itself with the circle. This explains the lack of participation by $n3$ in the line avoiding task because such fine motor skills are not required.

It is to be noted that depending on the activity in the neural network, the estimated TE networks significantly differ from the structural network. Another point to note is that the TE networks also include recurrent self-connections and these are computed simply based on a neuron's ability to predict its own behavior. While TE networks for only one circle and one line trial are shown here, we performed the same analysis for all trials on this best agent. We found that all the observations made for the individual trials were consistent across the rest of the trials. The relationship between different TE networks across all trials for the best $N = 3$ agent was determined by clustering of the 48 TE networks. This resulted in the networks getting organized by task. This observation was made

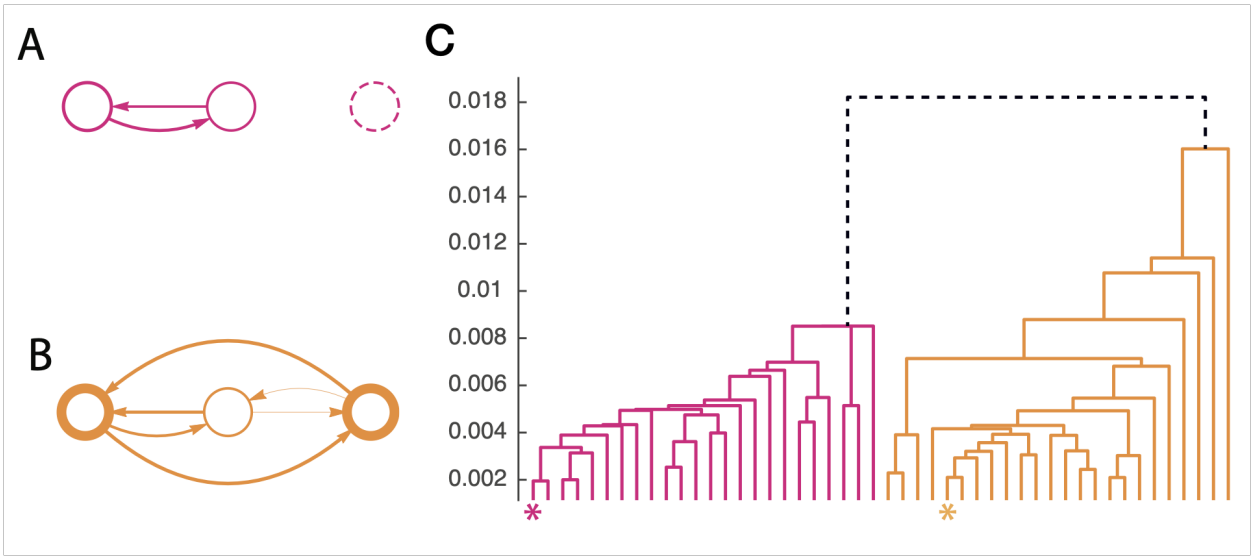

Figure 4.6: Transfer entropy networks of the best $N = 3$ agent. [A] Effective network for one line avoiding trial. Note that neuron 3 is not part of this effective network. [B] Effective network for one circle catching trial. [C] A dendrogram of the hierarchical clustering of trial-by-trial TE networks. The branches of the tree are color-coded by whether the TE network was inferred from a circle catching task (orange) or line avoiding task (magenta). It can be seen here that the tree can be cut at the top level to produce 2 clusters whose members are of only one task. The TE networks shown in panels [A] and [B] are identified in this tree with color matched asterisks.

by first estimating the task-specific TE networks for each of the 48 trials (24 circle TE networks and 24 line TE networks). These 48 networks were then clustered using a simple hierarchical clustering algorithm. The dendrogram of the clustered TE networks shown in Figure 4.6C shows an interesting property. All TE networks of the same task (either circles or lines) fall under one cluster before being grouped into the cluster of the other task. This means that the same structural network effectively presents itself as very similar networks for variations of one task, which are all different from the very similar effective networks for the other task. In other words, there is high within-task homogeneity and high across-task heterogeneity in the task-specific effective networks. Naturally, the next step is to look for this phenomenon in other $N = 3$ agents.

4.5.1 High CSC agents have high fitness

We developed a metric to quantify and compare task-specialization in TE network clusters, which henceforth we will refer to as cluster specialization coefficient, or CSC. $N = 3$ agents that had high

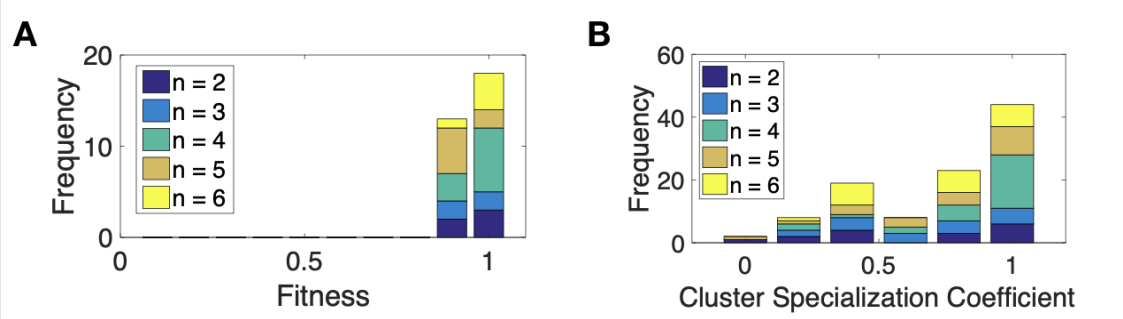

Figure 4.7: Performance and CSC distributions of the high-performing agents for all N. [A] A histogram of the fitness of all individuals that have the highest CSC of 0.95 showing that all agents that had a high CSC value performed well. [B] A histogram of the CSC values for agents that evolved to have a fitness $\geq$ 90%, showing that while there are agents that perform well with low CSC values, it is more likely that an agent with high fitness also has a high CSC value.

CSC showed very high behavioral performance. The greater the value of CSC, the greater the expression of within-task homogeneity and across-task heterogeneity, with the maximum value being 0.95 corresponding to the 2 specialized clusters that purely have effective networks corresponding to one and only one task in each of them. All agents, from the 100 evolutionary runs that had a CSC of 0.95 were collected and they were all high-performing agents. It is to be noted that the structural networks of agents that had a high CSC were highly degenerate in terms of the ratio of inhibitory-excitatory neurons and yet exhibited this phenomenon.

4.5.2 High fitness with high CSC is independent of network size

The same analyses that were performed on the best $N = 3$ agent and other $N = 3$ agents, were repeated for smaller ($N = 2$) and larger networks ($N =$ 4,5 and 6). Irrespective of the network size, agents that have the maximum possible CSC of 0.95, consistently showed high behavioral performance (Figure 4.7A). This establishes the idea that high within-task homogeneity and high across-task heterogeneity of task-specific effective networks yields high performance in this task. The obvious next question is to ask if the converse is true. Plotting the distribution of CSCs for agents that have $\geq$90% fitness (Figure 4.7B) revealed that there exists solutions that have CSCs as low as 0 that still perform well. However, a majority of agents that perform well have a high CSC.

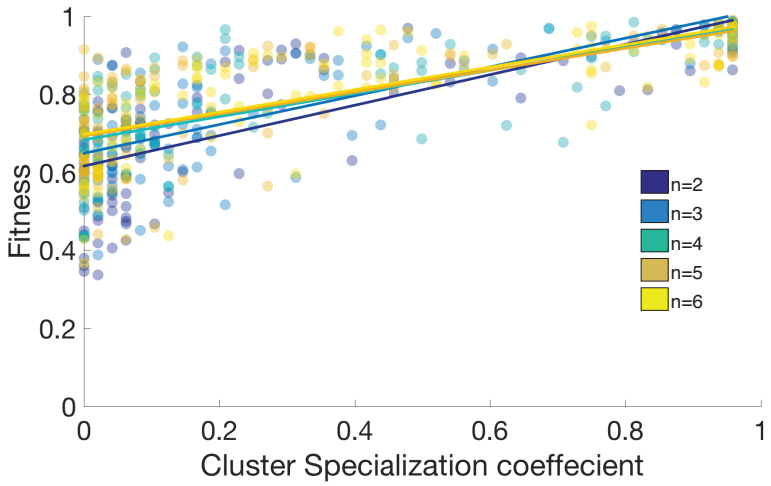

Figure 4.8: High task-specificity in the effective network suggests high performance. Fitness as a function of CSC for all best-agents across the ensemble, across different network sizes. The high-concentration of points in the right top corner reveals that a high CSC makes it more likely that behavioral performance is also high. Linear regression lines show a positive correlation between fitness and CSC for all network sizes. Interestingly there is no significant difference in this trend for different network sizes.

4.5.3 CSC positively correlates with behavioral performance

In order to further ascertain the relationship between CSC and fitness we looked at the fitness of all agents as a function of their CSC. As shown in Figure 4.8 fitting a straight line to the points on a CSC versus Fitness plot, shows a positive correlation between an agent's CSC and fitness for all values of N. Based on this and the previous result, we conclude that expressing specialized task-specific effective networks is not a necessary condition for high performance but conversely, it is sufficient for high performance. Given the generic nature of the network and it's evolutionary optimization process, it can also be said that this is true for all categorization tasks. It is intuitive that networks that can effectively express themselves distinctly for each category can perform well.

4.6 Environmental degeneracy enables reuse of ongoing neural activity

When the same neural network is involved in performing multiple behaviors, it is often assumed that there exists distinct modes of operation corresponding to each behavior (Getting and Dekin, 1985) and the results demonstrated so far are in line with those intuitions. However, it is possible

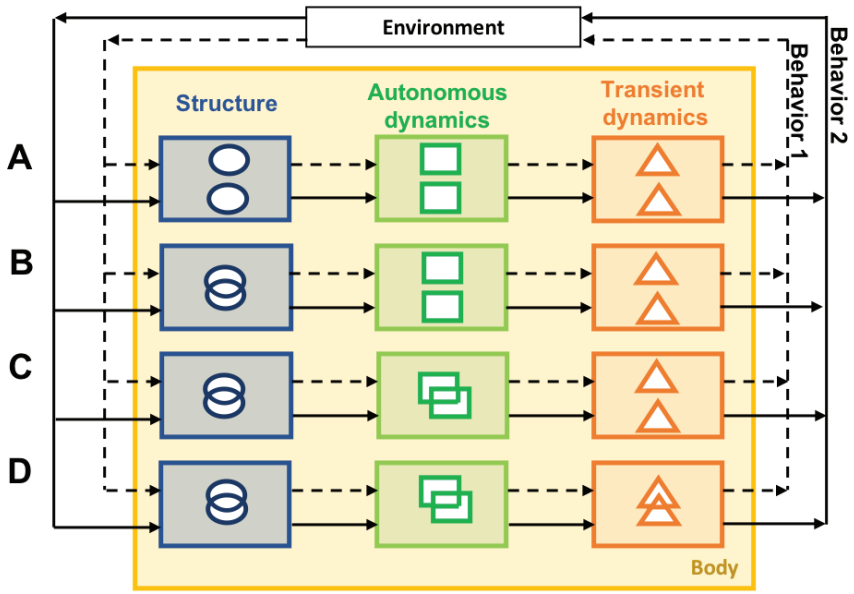

Figure 4.9: Three levels of neural reuse in multifunctional agents: structure (blue), autonomous dynamics (green), and transient dynamics (orange). [A] Dedicated circuits for each behavior: non-overlapping structures, and thus non-overlapping autonomous or transient dynamics. [B-D] Multifunctional circuits: overlapping structures (i.e., shared neural resources for multiple behaviors). [B] Autonomous dynamics are unique to each behavior, and thus transient dynamics are also unique (i.e., different set of attractors for each behavior). [C] Autonomous dynamics are shared across multiple behaviors, but transient dynamics are unique to each behavior (i.e., overlapping set of attractors for multiple behaviors, but different dynamics when coupled with the body and environment). [D] Both autonomous and transient dynamics are shared across multiple behaviors (i.e., overlapping set of attractors and similar overall dynamics when coupled with the body and environment for multiple behaviors).

that besides such a "functional modularity", ongoing neural dynamics could be reused across tasks (Figure 4.1C). In order to study this, we perform dynamical systems theoretical analyses as outlined below on an CTRNN agents optimized to perform the categorization and pole-balancing tasks. This analyses reveals that neural reuse could occur down to the level of ongoing neural activity. The same neural activity could lead to distinct behaviors due to degeneracies in the environment. Importantly, this can only be observed when the brain is studied in conjunction with the closed-loop interaction between the body and the environment.

From a dynamical systems perspective, reuse in embodied recurrent neural networks unfold over three levels: structural network, autonomous dynamics of the neural network, and transient dynamics of the neural network (Fig. 4.9). Structure is defined by the neural circuit itself, the intrinsic

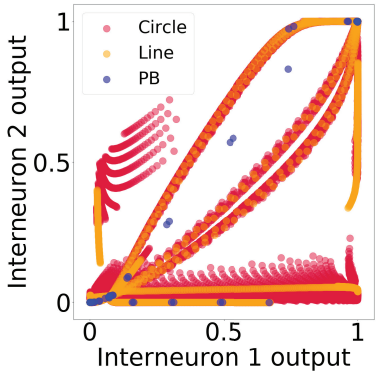

Figure 4.10: Attractor reuse. Locations of attractors from the three sets of phase-portraits corresponding to circle-catching (red), line-avoiding (orange) and pole-balancing (blue) tasks are overlapped. This shows that each behavior has its own set of unique attractors as well as shares them with other behaviors.

parameters of the neurons, and the synaptic strength of connectivity between them. While it is possible that an agent possesses specialized circuits for performing different behaviors (Fig. 4.9A), reuse at this level involves utilizing overlapping circuits to produce multiple behaviors (Fig. 4.9B,C,D). The next level, when structure is reused, is that of the neural network's autonomous dynamics isolated from the body. Each behavior is associated with a set of phase-portraits corresponding to the inputs the agent experiences while performing them. The sets of phase-portraits (and the attractors therein) could be overlapping (Fig. 4.9C,D) or could be unique to each behavior (Fig. 4.9B). The set of all attractors from all phase-portraits corresponding to a behavior are also referred to as the attractor set of the behavior in this paper. The third level of reuse is that of ongoing transient dynamics as the agent is in continuous closed-loop interaction with the environment. When there is attractor reuse from the previous level, it is possible that multiple behaviors navigate different transients around those attractors (Fig. 4.9C) or, they might be reused too (Fig. 4.9D).

4.6.1 Autonomous dynamics reuse: Overlapping sets of attractors were used to perform both behaviors

Given that the circuit is the same across the two behaviors (object categorization and pole-balancing), we wanted to evaluate if there was reuse in the autonomous dynamics of the neural network of this agent. In order to do this, we first constructed the phase-portraits for several inputs

across each task. A phase-portrait for a particular input can have one or more attractors. The set of phase-portraits associated with a behavior (say circle catching) can be obtained by fixing the inputs to what the agent experiences during that behavior (circle at fixed positions relative to the agent), and allowing the network to settle into its attractors from different initial states. The three sets of phase-portraits corresponding to circle catching, line avoiding and pole-balancing were compared based on attractor composition, basins of attraction, and location of attractors to evaluate reuse.

Attractor compositions refers to the type of attractors that were present in the set of phase-portrait for each behavior (i.e. fixed-points, limit-cycles etc.). In this agent, all phase-portraits associated with all behaviors in the best multifunctional agent were only composed of fixed-point attractors. Even though attractor composition is the same across behaviors, they could have different basins of attractions around those attractors. This could lead to different behaviors operating in its own region of the phase-space. However, for this agent, since only one fixed-point attractor existed in all phase-portraits, there exists only one basin of attraction which is the same across all behaviors. Thus, with same attractor composition and basins of attraction, the phase-portraits were qualitatively similar across all behaviors (i.e. no bifurcation).

These qualitatively similar phase-portraits could further be quantitatively compared based on the location of attractors in them. Differentiated by their location, each behavior could have a unique set of attractors or they could overlap to different extents; the exact locations of the fixed-point attractors on all phase-portraits of different behaviors do not have to be the same. Upon analyzing their locations we discovered that the multifunctional agent reused attractors identical in location between these behaviors (Fig. 4.10). This reuse was only partial since each behavior also had its own set of unique attractor locations that were not shared. Reusing the same attractors means that different inputs from different behaviors were mapped to the same phase-portrait, which suggests that there is an inherent degeneracy between the sensory inputs and the requisite behavioral pattern.

4.6.2 Transient dynamics reuse: Phases of the behaviors reused the same transient dynamics

Our analysis started at the level of structural reuse and went on to discover reuse at the level of autonomous dynamics in the best multifunctional agent. The next level is that of ongoing dynamics as the nervous system coupled with its body and environment performs the behaviors. Note that in the previous level, attractors were identified by fixing the relative position of the agent with the object and then allowing the dynamics of the network to settle to their attractors. However, during behavior, both the agent and the object are in constant movement. Therefore, at any given time, for a particular relative positioning of the agent and the object, the sensory input might change before the network settles into the attractor associated with that fixed input. As a result, dynamics of the network are in constant transient movement across the phase-portraits (and the attractors therein) associated with that behavior. Since attractors in the same location were partially reused in this agent, the relevant question to explore reuse at the next level is whether multiple behaviors have unique transients or if they could be shared.

Transient dynamics were shared partially between the circle-catching and pole-balancing tasks in the best multifunctional agent. In order to understand the behavioral implications of this, we evaluated the entire sensorimotor loop as this agent performed both behaviors. While in each case there were times when their transients were different, for a particular phase during these behaviors, the dynamics almost exactly matched (Fig. 4.11A). Inputs to the interneurons, their outputs, and motor neuron outputs were all identical. This suggests that the agent's nervous system does not differentiate between these phases of the two behaviors. This leads to two interesting questions: (1) Are the two behaviors indistinguishable during this phase? (2) If not, where does the difference come from?

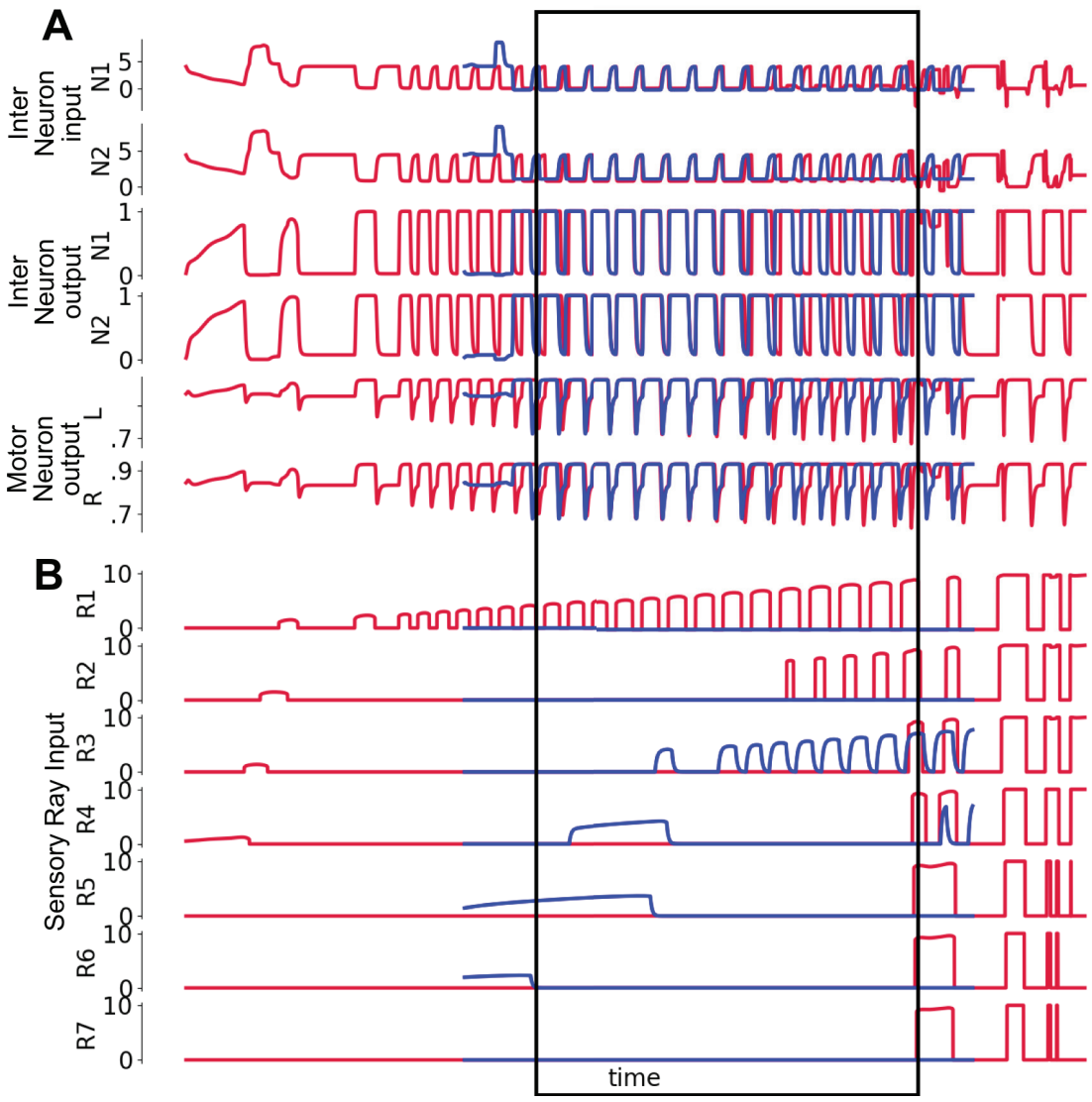

Figure 4.11: Transient/Driven dynamics reuse. [A] Activity of inter-neurons for the circle-catching (red) and pole-balancing (blue) tasks, time-shifted to show identical neural activity. [B] Sensory inputs to the 7 rays showing that although neural activity is indistinguishable, the agent tracks the circle along ray 1(red) but the pole is along rays 3 and 4 (blue).

Although the transient dynamics in the inter-neurons were identical, the behaviors were different in the circle-catching and pole-balancing tasks. In the former, the agent received oscillatory inputs only along ray 1, meaning that the agent oriented itself so as to track the circle along ray 1 before catching it (Fig. 4.11B). In contrast, the sensory inputs during the same transient dynamics in the pole-balancer shows that the agent maintains the pole oscillating around ray 3 bringing it across 4,5 and 6. This is an interesting outcome demonstrating that neural activity that is completely

indistinguishable can still produce behaviors starkly different from each other. The difference arises from the parts of the behavior leading up to this shared transient phase, where the agent has its own unique dynamics for each behavior through its interaction with the environment. Note that the weights to the interneurons from sensory rays 1 and 3 are not the same. Thus, transient reuse emerges purely from brain-body-environment interaction.

4.7 Discussion

In this chapter, I present results from the evolutionary optimization of embodied neural networks for cognitively-interesting behavior. Applying information-theoretic tools to extract task specific network representations based on neural dynamics in the interneuron layer showed the interesting characteristic of TE networks being clustered by task. We developed a novel metric, Cluster Specialization Coefficient (CSC), that quantifies the task-specificity of TE networks by dissecting the hierarchical clustering tree of trial-by-trial effective networks. Further analysis of the ensemble revealed a trend that shows positive correlation between CSC and performance. While we also noticed that there are some cases where agents performed well and had a low CSC, we found no agents that performed poorly with a high CSC. Further, we only claim correlation and not causation. However, in combination with neuroscience literature on the existence of task-specific effective networks in the macro scale (M. D. Fox and Raichle, 2005), we hypothesize that biological networks self-organize to have high within-task homogeneity and high across-task heterogeneity for the purposes of efficient categorization of stimuli.

It is widely known that biological neural networks perform multiple functions using the same underlying structural neural circuit (D. M. Blitz and Nusbaum, 1999; Briggman and Kristan, 2006b). Our work using transfer entropy (TE) provides a framework to study the task-specific effective networks that emerge from these otherwise structurally-similar networks. The same analysis can be performed to study the task-relevant variations in the neural dynamics and its robustness across variations of the task. Another perspective to look at this from, is that of neural reuse (Anderson, 2010b). Presenting new tasks in evolutionary time, would allow the networks to reuse behaviorally

appropriate functional components developed for previously evolved behaviors for the new behavior. While it is obvious that the same structural components are reused, functional reuse can be analyzed by comparing the task-specific effective networks. Furthermore, systematic studies of multiple behaviors can further explain the relationship between different types of behaviors and neural reuse. Our results have revealed a relationship between different categorization behaviors and the allocation of neural resources to perform them: acquiring category-specific effective networks yields high performance. This can be further tested for causal effect by including the CSC as component in the fitness estimation. This will help determine if encouraging this phenomenon makes it more likely to produce multifunctional circuits. Based on our results, we hypothesize that this will in fact be the case.

One of the limitations of this methodology is that TE does not scale elegantly with the size of the networks. While this is not an issue with neuroscientists where TE is used as a post-task analysis method, an online estimation of TE networks during evolution can be computationally expensive. Although small spiking networks, even as small as 2 neurons like we have shown here, have high computational power (Maass, 1997), this can be a problem in large networks for complex tasks. Therefore, in order to use CSC as a component of fitness to more efficiently develop artificial systems, a computationally more efficient metric might be required. This is one of the directions in which this study can be explored. Meanwhile, the limits of TE can be pushed to evolving and analyzing increasingly larger networks for a number of different behaviors and also for performing multiple behaviors. These behaviors can be systematically chosen to be (a) completely independent, (b) overlapping or (c) partially overlapping. Studying the relationship between CSC and fitness in each of these cases would provide interesting insights into neural reuse and how neural dynamics shapes behavior. Another limitation of this study is that the TE analysis was performed on a subset of the evolved parameters: the recurrent connections between interneuron spiking neurons; sensor-to-interneuron weights and interneuron-to-motorneuron weights were not included. Constraining the optimization space to only the recurrent connections and setting other weights constant, would make sure that the TE analysis is performed on the system that is entirely-responsible for controlling

the differences in behaviors. Finally, TE averages neural dynamics over the entire trial. Although this is useful for comparisons across trials, in order to get an in-depth understanding of how the network produces behavior, unrolling the information analysis over time will be required.

In this work, the temporal analyses were performed by adopting a dynamical systems theoretic approach. We discovered reuse of neural circuits at the structural level, followed by reuse of autonomous dynamics with qualitative sharing of phase-portraits, overlapping basins of attraction and reuse of attractors identical in the location of their fixed-points. Furthermore, we discovered partial reuse of transient dynamics in the best multifunctional agent. The two main contributions of our work are as follows: (1) the same neural circuit can perform multiple behaviors using the same sensory and motor systems in the absence of explicit task identifying signals or processes such as neuromodulation; (2) indistinguishable neural activity, displaying reuse to the level of transient dynamics, can still produce completely different behaviors.

The rationale behind transient reuse in this multifunctional agent can be explained by analyzing the environment-body relationship and transient dynamics. The similarity in dynamics arises out of the agent's ability to generalize between the two behaviors by learning to align an object along a single ray - pole along center ray and circle along corner ray (Fig. 4.4 and 4.11). Generalization requires learning to use only one ray because the pole only intersects one ray at a time. The circle had to be balanced along the corner ray because otherwise it would intersect multiple rays and the pole needs to be balanced along the center ray to maximize fitness. The difference in behavior, however, arises out of the unique transient dynamics prior to shared transient phase of the behaviors. The unique dynamics in circle-catching orients the falling object along the corner ray even if the circle starts from the center, whereas in pole-balancing it brings the pole to the center, thereby setting up the system to perform generalized object tracking along a single ray for both behaviors (Fig. 4.4B). This is possible because of the structure provided by the environment and the body. Objects intersect only one ray or multiple rays, yet the agent is required to align with the object in both cases. Multifunctionality in this agent is made possible by the closed-loop interaction between brain, body, and environment.

Due to experimental limitations, the study of multifunctionality has been mostly concerned with motor neuron circuits capable of generating multiple patterns of activity (Briggman and Kristan, 2008). Here we extend this framework to circuits that are *behaviorally* multifunctional: from sensory input, through interneurons, to motor neurons responsible for generating actions. We demonstrate that multifunctionality can result from the closed-loop interaction between brain, body, and environment in the absence of other mechanisms such as neuromodulation. Therefore, our results expand the list of mechanisms that can result in multifunctionality to include closed-loop interactions. Ultimately, this mechanism can coexist with previously described mechanisms, including neuromodulation and synaptic plasticity.

The work presented here opens up several avenues for further research of neural basis of multifunctionality. For instance, an ensemble of multifunctional networks can be studied to investigate the diversity in the extent of reuse in individuals that are behaviorally similar yet different in their neural dynamics. Also, analyzing larger networks that can perform a greater number of tasks would enable a comparison of how behaviors may be organized from a neurodynamical perspective versus how humans might organize them based on our descriptive understanding of the tasks. Furthermore, important theoretical advancements in this domain would be providing a confluence of the information theoretic and dynamical systems theoretic approaches presented here akin to Williams and Beer (2010a). Transfer entropy estimated in time should provide an alternative explanation of transient reuse (or reuse of ongoing neural dynamics). The dynamics of effective networks would reveal that effective networks at different points in time during each task are identical. This would provide a scalable alternative to dynamical systems theoretic analysis since it captures the temporal aspects of neural reuse.

We show that estimating effective networks using transfer entropy enables us to capture the task-specific time-averaged neural dynamics in a network representation thereby enabling us to compare the difference in dynamics arising from the same network in the context of different tasks. A complementary temporal analysis rooted in dynamical systems theory is presented. These approaches have the potential to explain neural reuse in any behavioral system, biological or

artificial. We intentionally focused on a small neural controller and a simple set of behavioral tasks. However, the possibilities uncovered in this system should be available to larger neural networks solving more complex tasks. Ultimately, taking an embedded and embodied approach to studying neural reuse offers a distinct perspective on several topics of interest to understanding cognition, including modularity in brain organization, localization of cognitive functions, and more generally the mapping between brain structure and function.

Chapter 5

Multiagent Interaction

In this chapter, I present our work demonstrating how integral social interaction is to our study of cognition. In the first section of this chapter, I introduce our work and motivate it. Following that, related work in the social cognition domain including experimental and computational methods are discussed. Next, details of the embodied neural network model and the analyses methodologies are explained. Finally, results from analyzing the evolved social agents are described and discussed.

5.1 Introduction

Behavior in living organisms emerges from the continuous closed-loop interaction in the integrated brain-body-environment system. Does the brain generate behavior on its own merely using the body and the environment as a medium for receiving information and manipulating it? Or do the body, the environment and the interaction between these components themselves contribute to the behavior in a way that cannot be understood by studying the brain alone? The two previous chapters demonstrated how the environment and interaction with it, can induce (predictive) information in a neural network as well as enable reuse of neural resources across tasks. In this chapter we study cognition in brain-body-environment systems where the environment comprises other agents. Understanding neural mechanisms of social cognition, under this paradigm, requires analysis of not just the brain but also the environment and interactions with it.

Not too long ago a radical individualism about collective phenomena was the only game in town, leading respected philosophers to conclude that ultimately the basis of our mental life does not

depend on others at all, such that it would make no difference if others were just a hallucination of a "brain in a vat" (Searle, 1990). Nowadays there is a growing consensus that this pessimistic view is inadequate, and that social interaction can make a difference to the mental and behavioral activity of individuals (Froese, 2018). For instance, evidence from neuroimaging, psycho-physiological studies, and related fields has revealed that the mechanisms of social cognition are different when we are in real-time interaction with others compared to when we are passive spectators (Schilbach et al., 2013).

Nevertheless, the extent and nature of the influence of social interaction on an individual is still contentious. Most researchers adopt a moderate individualism in which interaction with others can make a difference but only externally so, for example by serving as a source of additional information, by having a causal influence, or by providing an opportunity for adopting a more socially oriented mode of cognition (Gallotti and Frith, 2013). Other researchers adopt an enactive approach that questions the validity of this restriction, proposing instead that the interaction in itself can play a role in realizing an individual's cognition, thereby transforming and augmenting the individual's capacities (De Jaegher et al., 2010). On this latter view, social interaction could allow an individual to overcome the limitations of their individual capacities by incorporating the complex dynamics of the interaction process into their internal activity.

Agent-based modeling offers a suitable framework with which to start investigating this possibility in a systematic manner. In particular, by simulating pairs of mobile agents in highly simplified scenarios it becomes possible to systematically assess the relationship between individual complexity and social interaction (Froese et al., 2013b). For instance, in previous work one of us provided a proof of concept that evolving two agents to locate each other in an open-ended arena via acoustic coupling can result in activity in their neural controllers, which in principle would have been too complex for them to generate in isolation (Froese et al., 2013a). Here, we show that this is not an isolated finding: directly evolving pairs of agents to increase the complexity of their neural activity consistently results in behavioral strategies involving mutually coordinated interaction between them. Moreover, we show that there is a crucial difference between forms of

interaction in which the agents' behaviors are interdependent compared to independent from each other: neural complexity achieved during mutually coordinated interaction tends to be even higher than what can be achieved during one-way coordinated interaction.

5.2 Related work

The Interactive Brain Hypothesis (IBH) proposed by Di Paolo and De Jaegher (2012) states, "...interactive experience and skills play enabling roles in both the development and current function of social brain mechanisms, even in cases where social understanding happens in the absence of immediate interaction.". In other words, they claim that the fact that animals are embedded in an environment is not just a consequence of having to exist in a social world but is in fact part of the machinery required for social cognition. Specifically, the continuously unfolding series of interactions, that is beyond the full-control of the animal, is crucial to social cognition and its neural mechanisms cannot be understood by studying an individual alone. This is in contrast to the weaker, more universally accepted claim that environmental interaction is important to consider for social neuroscience because it simulates the naturalistic environments that animals are embedded in. IBH makes the stronger claim that social interaction enables neural processes that are not possible otherwise, and that the unit of analysis should not be limited to the brain but should include the environment and interactions with it.

5.2.1 Behavioral studies of social cognition

The IBH is a culmination of a number of computational modeling and experimental studies on the influence of social interaction and joint action on an individual (reviewed in Marsh et al. (2009); De Jaegher et al. (2010) and Riley et al. (2011)). Two landmark behavioral experiments that highlight this influence are the "perceptual crossing" experiment (Auvray et al., 2009) and the "double TV monitor" experiment (Murray, 1985; Nadel et al., 1999). The former involved a task where two blind-folded subjects were asked to identify each other's cursor on a shared screen amidst other distractors based on tactile feedback received upon encountering an object on the screen. This

study showed that subjects were only able to identify the other based on the self-organized mutual sensory stimuli they could elicit upon scanning each other's cursors that they cannot otherwise elicit while scanning other objects on the screen. Importantly, subjects were easily able to distinguish conditions where they were scanning a shadow of the other subject that was not responsive to their own movement. Thus, the ability to recognize the other individual was not entirely a function of an individual's ability to categorize tactile stimuli, but was in fact dependent on the ongoing interaction during the course of the task. In the "double TV monitor" experiment, facial cues such as smiling versus frowning, mouth closures and eye contact were measured in infants under two conditions, namely playback of their mother on video compared to a live video conference with their mother. It was shown that in ages as young as 12 - 16 months the children were able to distinguish the two conditions and demonstrated different behaviors in each case. These behavioral experiments suggest that animals are involved in joint sense-making in a way that cannot be performed by one individual on their own, and hence cannot be understood by studying one individual in isolation.

The idea that understanding of the neural basis of an individuals behavior requires a comprehensive study of brain-body-environment system has existed for a while (Varela et al., 1991) and has been demonstrated by analyzing computational models of "minimally cognitive" agents (Beer et al., 1996). Extending this idea to environments that include another agent, the behavior of the agent of interest then constitutes the dynamics of the other agent as well (Froese and Di Paolo, 2011). The perceptual crossing experiment was replicated using a computational model by Di Paolo et al. (2008). That study and further experiments using variations of this model in Froese and Di Paolo (2010) and Froese and Di Paolo (2009) demonstrated that success in this task very strongly linked to the stability of the mutual interaction between the two agents, further emphasizing that this task is a demonstration of cognition being offloaded into environmental interaction rather than being entirely within the individual.

5.2.2 Neuroscience of social cognition

What if we can replicate the stimuli from interaction in an isolated brain? One experiment that can support this idea is Ramirez et al.'s (2013) work on creating false memories in mice by optogenetically stimulating hippocampal neurons consistently under certain contexts to replicate behaviors similar to those developed from a classical conditioning paradigm in the same contexts. Specifically, by stimulating the memory engram–bearing cells in the hippocampus in contexts where the animal would have otherwise received an electric shock, the authors were able to elicit behaviors that were consistent with conditioning from actually receiving electric shock. While this study appears to provide support for the case that one can create simulations of environment and elicit actions precisely, as De Jaegher et al. (2016) rightly point out, this stimulation paradigm does not extend to general behaviors in a complex environment. To stimulate an interaction with another animal would mean that the entire temporal pattern of stimulus must be replicated. This is not only complex, but also highly-dependent on the experimental mouse's actions that are not deterministic. It would unfold into a complex causal web of influences between the two animals. Importantly, note that the other mouse is also a dynamical system on its own and can manifest its own stochastic behavioral patterns. Thus, real-time cognition cannot be practically captured by a series of input-output transformations happening in the brain alone.

Experimental work in support of the idea that interactions are key to cognition come from the neural reuse literature (Anderson, 2010b; Anderson et al., 2012) showing activation of sensory and/or motor areas when higher cognitive functions are performed. For instance, it has been shown that verb retrieval tasks activate motor areas and noun retrieval tasks activate visual processing areas (Damasio and Tranel, 1993; Damasio et al., 1996; Martin et al., 2000, 1996) in the brain. Perception of manipulable objects activates grasping related ares in the motor cortex (Chao and Martin, 2000). Further, abstract planning tasks that themselves do not require any actions have been shown to activate motor areas (Dagher et al., 1999) and finally, motor ares have been shown to be activated in number processing as well (Andres et al., 2007; Roux et al., 2003; Rusconi et al., 2005; Zago et al., 2001). These results suggest that sensory and motor capacities are co-activated

during higher cognitive functions in way that suggests these faculties are available in aiding the performance of the behavior. It is, however, not yet clear what the causal implications of these co-activations are to performance of the behavior.

One of the crucial technological requirements that is holding back experimental progress in this field is the ability reliably record neural activity under conditions that involve social interaction. In this regard, Redcay et al. (2010) setup a live video stream inside an fMRI scanner that enabled to setup conditions such as mutual interaction with the experimenter and playback of recorded video. They observed greater activation during the live condition when compared to the playback condition in brain regions such as right temporoparietal junction (rTPJ), anterior cingulate cortex (ACC), right superior temporal sulcus (rSTS), ventral striatum, and amygdala. While it is still unclear what increased activations in these brain regions mean, their tool opened up the possibilities of conducting further experiments that could incorporate joint sense-making during a task. More recently, Lorenz et al. (2016) presented a machine-learning based methodology for closed-loop stimulation in an fMRI experiment. While their work was not directly applied to social cognition, it can be adapted to study social cognition. Finally, with the advent of virtual reality (VR) simulators, scientists are performing fMRI experiments when the subject is in the scanner wearing VR goggles (Pine et al., 2002); another technology that could be potentially used in social neuroscience.

From a computational modeling perspective, previous studies of social cognition have mostly been limited to behavioral replication. Although Di Paolo et al. (2008) and the following modeling studies used Continuous-Time Recurrent Neural Networks (CTRNNs), their analysis was limited to the behavior of the optimized models and not the neural dynamics during social interaction. One modeling study that did analyze the implications of social interaction on neural dynamics by Froese et al. (2013a) demonstrated that agents controlled by 2-neuron CTRNNs exhibited chaotic dynamics when optimized in the presence of another interactive agent; dynamics that are beyond the capability of 2-neuron CTRNNs in isolation. The work presented in this chapters extends this work to systematically test if the richness in neural dynamics achievable in the presence of an interactive agent much higher than that achievable in isolation or in the presence of non-mutually-interactive

agent.

5.2.3 Measuring richness of neural dynamics

Typically, increased activation in brain regions observed from recording methods is interpreted as an increase in "processing" or "computation". Alternatively, the richness or complexity of neural dynamics can be measured from neural activity using information theoretic measures. Tononi et al. (1994) proposed a measure called "neural complexity", C_N with the intention of capturing the functional organization of brain networks. According to this metric, there is high neural complexity when small subsets of a system (think local brain regions) show high statistical independence between one another, and larger subsets show low statistical independence. This measure was intended to capture a specific idea that independent functional sub-components of a system that can all act in unison is required to generate behavior. Thus, although aptly named neural complexity, this measure was developed for that specific context and is not suitable for the work presented in this chapter.

In this work, we measure the complexity of neural dynamics as the entropy of neural activity (Strong and Bialek, 1998). This is based on the idea that entropy captures the information capacity of the neural dynamics; greater the entropy, greater its ability to encode information. Thus, by optimizing agents to maximize their neural entropy we in turn maximize their ability to encode information. We study the maximum achievable neural entropy under conditions of mutual social interaction, one-way interaction and isolation to demonstrate that interaction is not just a source of stimuli but in fact augments neural capacities to more than what can be achieved otherwise.

5.3 Methods

Experiments were conducted on pairs of simulated agents that interacted with one another in an empty 2-dimensional environment. Each agent emitted an acoustic signal, which could be sensed by the other via two sensors positioned at the perimeter of their circular bodies (Di Paolo, 2000).

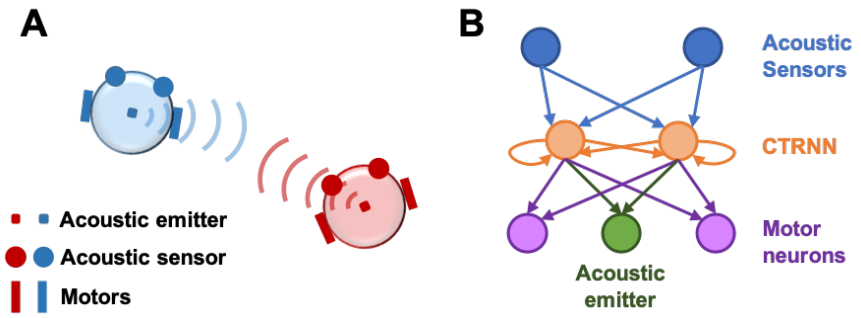

Figure 5.1: Setup of computational model and neural network architecture. [A] Illustration of socially interacting agents. Two agents, each consisting of an acoustic emitter that they are able to modulate, a pair of acoustic sensors to sense the other agent, and two motors to move in a 2-dimensional environment. The ability to modulate their own signal combined with their ability to listen to their counterpart, enables interaction in this model. Agents cannot sense themselves. [B] Neural architecture of the agents. The two acoustic sensors feed into a 2-neuron fully-connected continuous-time recurrent neural network (CTRNN) circuit which in turn feed into the two motors and the acoustic emitter. The movement of the agent is result of the net activation of the left and right motor neurons.

The strength of the emitted signal faded linearly with distance, and sensors were positioned to be 90 degrees apart from one another (Fig. 5.1). Thus, agents can gather information about their relative distance and orientation to one another. Neural controllers were modeled as dynamical recurrent neural networks (Beer, 1995c). Sensory input filtered through sensory neurons into an inner layer of two interconnected neurons, whose activity modulated the power of the emitted acoustic signal and controlled motor neurons that propelled the agent around its environment.

This section outlines the implementation details to recreate the simulated agents, and all experiments and analyses. We first describe the agent, their neural circuitry and their principles of movement and acoustic signaling, followed by an explanation of the evolutionary stochastic search algorithm used to optimize these agents, and then a description of the analyses is provided.

5.3.1 Agent design

The design of the simulated robots has been primarily inspired by Di Paolo's acoustically coupled agents (Di Paolo, 2000). The simulated agents were circular with a radius, R, of 4 units and were equipped with two acoustic sensors and an emitter. The two sensors were located symmetrically at

an angle of 45 degrees to the central axis. An acoustic emitter was located at the geometric center of the body, and hence equidistant to an agent's own sensor making them essentially deaf to their own signal. The emitted signal experiences linear loss in strength with distance. The strength will be maximum and equal to that of the emitted strength at a distance equal to the 2R between the center of the agents and linearly drops off with increasing distance. The sensors that pick up this signal hence provide an estimate of the distance of the source and the differential sensory stimuli in the two sensors give an estimate of the direction of the source relative to the sensing agent (Fig. 1A).

In addition to attenuation due to distance, the sensing agent also experiences attenuation when the signal passes through its own body – a self-shadowing mechanism that is a natural consequence of embodied agents. This attenuation is computed as a scaling factor over the sensory inputs which ranges from 0.1 to 1, depending on the proportion of the body's diameter that the signal travelled through i.e. the sensory signal is scaled by 0.1 if the sensors are diametrically opposite from the source and scaled by 1 if the agents is facing the source. This is computed as follows - for a distance D between the two agents, and a distance of D_{sen} between the source and the sensor, the shielded distance that the signal travels through the body, D_{sh}, is given by

$$\begin{aligned} D_{sh} &= D_{sen}(1 - A) \;\; 0 \leq A \leq 1 \\ A &= (D^2 - R^2)/D_{sen}^2 \end{aligned} \tag{5.1}$$

where when $A \geq 1$, there is an unobstructed line between the source and the sensor. The sensory input for a sensor is first calculated based on the distance between the sensor and the source, and it is then multiplied by a scaling factor between 0 and 0.1 by linearly mapping D_{sh} between 0 and 2R respectively. This process was then repeated for the other sensor.

Collisions were modeled as point elastic, meaning that upon colliding, the agents experienced no change in their angular velocity (no friction between bodies) and momentum of the entire system was conserved by having zero net effect on their velocity vectors. This was performed by simply

exchanging the velocity vectors of the agents thereby causing them to bounce off of each other without losing any energy in the process.

The internal neural circuitry is made up of three layers – sensory layer, interneuron layer and the motor control layer (Fig. 1B). The sensory neurons can be thought of input neurons with a sigmoidal activation function. Their output is given by

$$o_s = g_s\sigma(I_s + \theta_s)\ where\ \sigma(x) = 1/(1 + e^{-x}) \tag{5.2}$$

is the sigmoidal activation function, g_s is the sensory gain that is kept the same for both sensors, and θ_s is the bias that is also same for both sensors.

The interneuron layer is a continuous-time recurrent neural network (CTRNN) that is fully recurrently connected. This corresponds to a two-dimensional dynamical system with the activity in each neuron governed by the following state equation

$$\tau_i \frac{dy_i}{dt} = -y_i + \sum_{j=1}^{N} w_{ij}\sigma(y_j + \theta_j) + \sum_{s=1}^{2} w_{is}o_s \tag{5.3}$$

where dy_i/dt refers to the rate of change of internal state, y_i of neuron i based on a time constant τ_i. This rate of change depends on three values – the current state, the weighted sum of outputs from all N neurons in the network, and the total external input. The input from other neurons is calculated by weighting their output with weights from neuron j to i specified by w_{ij}. The output of each neuron based on its internal state is given by $\sigma(y_j + \theta_j)$ where θ_j refers to a bias term for that neuron. Finally, the state is also influenced by the total external input received by the neuron, in this case given by the weighted sum of the sensory input with weights w_is from sensory neuron s to interneuron i and o_s being the sensory output from two sensors.

The interneurons feed into the motor control layer, where the input to each motor neuron is a weighted sum of the outputs of the interneuron. The motor control layer contains three neurons, two corresponding to the left and right motors and one that corresponds to the acoustic signal emitter. All three of them are sigmoidal units with a gain and bias but no internal state such that

the output of unit i, m_i, is given by

$$m_i = g_m \sigma(\sum_{n=1}^{N} w_{ni} * o_n + \theta_i) \tag{5.4}$$

where o_n is the output of the interneuron, that are weighted by w_{ni} and θ_i is the bias term that is common across all motor units, and so is their gain g_m.

Locomotion is controlled by the effective control of the two motors, where net linear velocity is given by the average of their outputs and the angular velocity which rotates the agent and hence the direction of movement is given by their difference divided by the radius of the agent.

5.3.2 Measuring neural entropy

The neural activity in an agent is recorded during the course of behavior, and the neural complexity is measured as the entropy in the two-dimensional time series from the outputs of the two interneurons. Since these are outputs from a sigmoid function, they are bounded between 0 and 1. The output space is binned with 100 bins along each dimension, totaling 10000 bins in all. A 2-dimensional histogram is built from binning data points collected over all trials of the behavior and the probability of the neural activity in a bin [i,j], p_{ij}, is given by the number of points in that bin divided by the total number of points. From these probabilities, the entropy H of the neural time series is given by

$$H = \sum_{i=1}^{100} \sum_{j=1}^{100} -p_{ij} log(p_{ij}) \tag{5.5}$$

The maximum possible entropy that can be achieved is when all bins are uniformly populated giving a uniform distribution over the two-dimensional histogram. The entropy computed is normalized to be in the range [0,1] by dividing by this maximum entropy which is equal to log(100*100). Thus, normalized entropy is $\hat{H} = H/log(100 * 100))$.

5.3.3 Evolutionary optimization

Parameters of the neural controllers, such as weights and signs of the connections, biases, and time-constants, were optimized using an evolutionary algorithm. Each evolutionary run was initialized with a random population of 96 solutions, that was evolved over 500 generations. 100 such runs were executed and the best solution in the population from each run was collected to be analyzed. In order to evaluate the fitness of the individuals, we computed the entropy of the time series of neural activity taken from simulated trials. This measure allowed us to operationalize the complexity of internal neural dynamics exhibited by each agent in various interaction conditions. In particular, neural entropy was measured for each agent in trials where they were evolved and interacted in pairs (*interaction entropy*), as well as control conditions where agents were placed in the environment by themselves (*isolation entropy*). Our decision to use neural entropy as an index of internal complexity was motivated by its interpretability and computational tractability, as well as a range of previous studies that have associated elevated levels of neural entropy with improved cognitive performance, including therapeutic benefits (Carhart-Harris et al., 2014), increased levels of consciousness (Schartner et al., 2017), and improved generalization in motor learning tasks (Dotov and Froese, 2018).

A real-valued stochastic evolutionary search algorithm was used to optimize these agents to maximize their neural entropy. Each agent had 20 parameters that need to be tuned, and they were encoded as the genotype of the evolutionary search. For K agents, the genotype contained 20K values in the range [-1,1] that were scaled appropriately to construct the agent. The sensor and motor gains were scaled to be in [1,5] while their biases were scaled to be in [-3,3]. All weights were scaled to be in the range [-8,8] while the time-constants in the CTRNN were set in the range [1,2] and their biases were scaled in [-3.3]. Thus, one genotype representing one solution was mapped to construct the agent(s) to then be simulated and evaluated for performance.

One generation of evolutionary optimization involved evaluating a population of 96 solutions and generating a new population for the next generation based on their performance. Agent(s) built from each genotype were evaluated over 4 independent trials each lasting 200 simulation seconds

at a step size of 0.1. Their neural activity was recorded, and entropy was computed over the 4 trials and the normalized neural entropy was assigned as their fitness.

In the interaction scenarios, the agents were always placed at a distance of 20 units from each other, but their relative angle was varied as $[0, \pi/2, \pi, 3\pi/2]$ for each trial. The fitness of each genotype in the multi-agent runs was computed as the average neural entropy of all agents that the genotype encoded. In order to promote interaction, and avoid the case where agents moved too far from each other and not be able to sense the other, a distance threshold of 100 units was set. Any trial where the agents moved further than 100 units from each other was cut-off.

Upon evaluating their fitness, an elite fraction of the top 4% solutions were kept as is, and a new population was created by creating the remainder of the solutions by mutating and crossing over these elite solutions. Mutation involved adding a zero-mean Gaussian mutation noise with variance 0.1 to the solutions. Following that, they were crossed over with other solutions such that each parameter between a pair of solutions was swapped with a probability of 0.1. This process was repeated over 500 generations and the best solution from the population at the end of 500 generations was chosen as the representative solution from that run. 100 such independent runs were executed for each scenario discussed in the results section, and the analyses were carried out from the collection of the 100 best agents from all these runs.

There are two different scenarios under which agents are evolved - isolation and interaction. When evolved in isolation, the genotype encodes the parameters of a single agent, and the fitness of each agent is evaluated by the entropy of its neural activity during and evaluation period. On the other hand, when agents are evolved with interaction, they are evolved in pairs. In this case, each genotype encodes the parameters for two agents. The pair of agents built from the genotype can now potentially interact with each other to influence each other's neural activity. Fitness for each genotype is then evaluated as the average neural entropy of the two agents.

5.3.4 Measuring distance entropy

The distance between the two agents, unlike their neural entropy, is a one-dimensional time series. During the optimization trials, the distance between agents was set to a maximum limit of 100 units, beyond which the trial was cut off. In order to maximize fitness all evolved agents optimized their behavior to stay within this bound. This range of distance ([0,100]) was binned into 100 bins, which allowed the creation of one-dimensional histogram from the distance between the agents during the course of the behavior, over all trials. From this, the probability and the entropy were computed in the same way as neural entropy and normalized by the maximum entropy of log(100).

5.3.5 Analysis with "ghost" partner

In order to delineate the role played by interdependent interaction on internal complexity as opposed to independent interaction, agents that showed high levels of internal complexity in the presence of a partner were tested under a "ghost" condition. In this case, from the genotype of the best solution of a particular run, only one agent was constructed. This agent was placed in an environment in the presence of another agent. However, the other agent, not being constructed from the genotype, was not an active system that was behaving in this environment. Instead, that agent was replaying pre-recorded behavior from the trials that were conducted after optimization. This is referred to the "ghost" agent. In order for the active agent, whose entropy is being measured, to also not repeat its behavior from those trials, the two agents were started at different random initial angles from each other, while keeping the initial distance the same. Same as the evolutionary fitness evaluation, 4 trials were conducted, and the neural entropy of the active agent was measured based on its behavior in the presence of a "ghost". This setup allows neural entropy to be measured under similar conditions of sensory complexity that the agents experience during interdependent interaction, while at the same time removing any possibility for the same.

5.3.6 Curve fitting and statistical testing

IBM SPSS Statistics software was used for regression modeling and statistical testing. The distribution of evolved entropies between the different setting of isolation and interaction were statistically tested by comparing means using a two-tailed t-test with significance declared if $p \leq 0.05$. For the regression model, coefficients were fit based on 95% confidence bounds and an analysis of variance (ANOVA) over the regressors, and the residuals was conducted and tested using an F-statistic test to check against all coefficients being 0. Further, each co-efficient was individually tested against being 0, using a two-tailed t-test with significance declared for p-values below 0.05.

5.4 Interaction enhances internal complexity beyond what is possible alone

First, in order to study the effect of interaction on internal complexity, we artificially evolved pairs of agents to maximize their interaction entropy, without explicitly specifying any desired behavior. The resulting movement and neural traces from one trial of one of the best evolved pairs of agents from 100 runs is shown in fig. 5.2. During interactions, these agents exhibited normalized neural entropies of 0.7568 and 0.8763. Although behavioral interactions were not selected for, evolved agents exhibited a complex pattern of moving towards and away from each other in a coordinated manner. Qualitatively similar behaviors were observed in the rest of the evolutionary runs.

We expected that agents would evolve to make use of social interaction to enhance their internal complexity, if there was an opportunity to do so. In order to verify this prediction, we performed another set of experiments where we evolved isolated agents using the same fitness function. Comparing the neural entropy achieved by agents in 100 independent evolutionary runs in each condition revealed that internal complexity

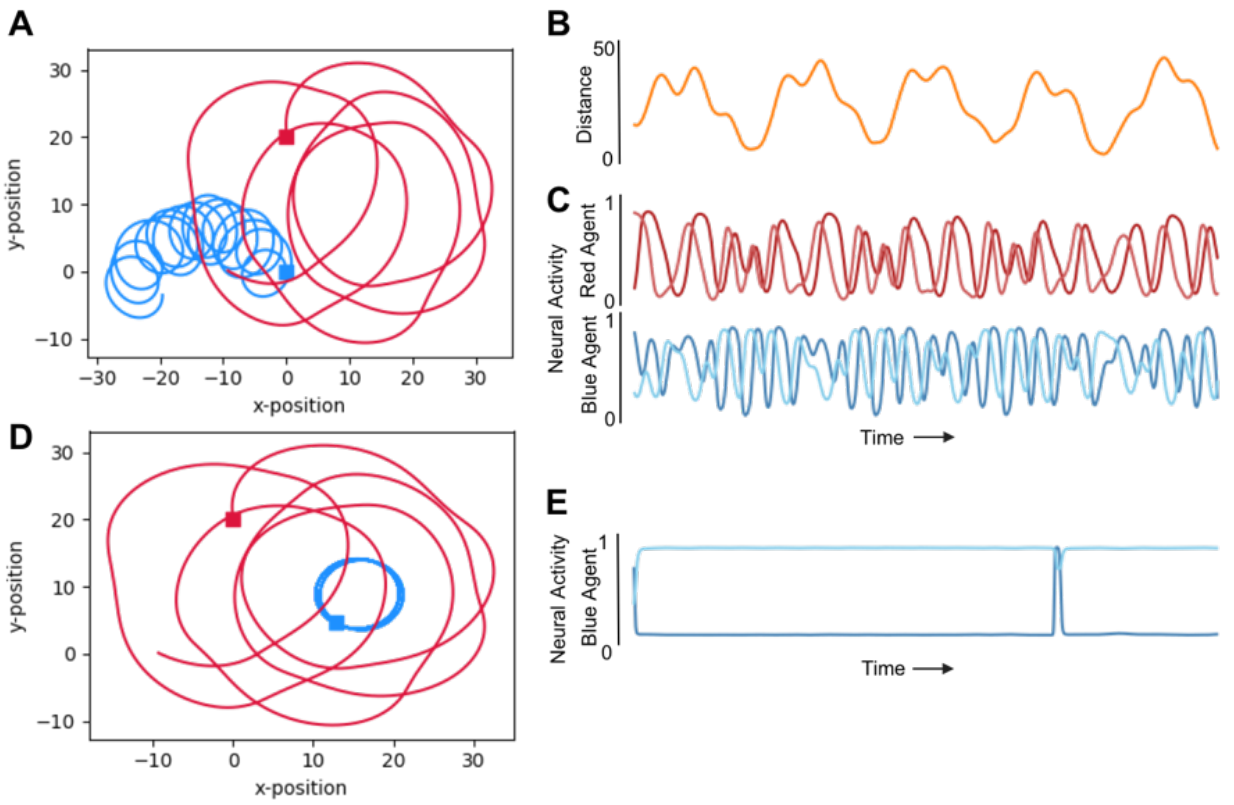

Figure 5.2: An agent's behavior, neural activity and distance traces in interaction and its performance and neural activity in the ghost condition. [A] An illustration of the 2-dimensional behavioral pattern of two agents evolved to interact demonstrating aperiodic oscillatory patterns that cannot be achieved by the 2-neuron systems of each agent in isolation. [B] Relative distance over time of the two agents shown in B, also demonstrating interesting complex patterns that cannot be achieved by passive 2-neuron CTRNNs. [C] The neural activity of the 2 interneurons of red and blue agents shown in B, demonstrating chaotic aperiodic activity that cannot be generated by 2-dimensional CTRNNs in isolation in the absence of interaction. [D] The same agents as in B, but in this case the red agent plays back the recorded behavior from the trial shown in B, while the blue agent is allowed to interact with it. Significantly reduced behavioral complexity is observed under this "ghost" condition where agents are unable to mutually interact with each other. [E] Neural activity in interneurons of blue agent under the ghost condition, showing significantly lower complexity compared to the same agent's neural activity in the interactive mode shown in D.

was significantly higher when agents had the ability to interact as opposed to when they existed in isolation (Fig. 5.3A; significance tests results in table 5.1). In other words, the interaction entropy of agents evolved in social contexts is consistently larger than the isolation entropy of agents evolved in isolation. Crucially, agents that were optimized to maximize to neural complexity in the presence of others, exploited their ability to interact to achieve it (Fig. 5.4).

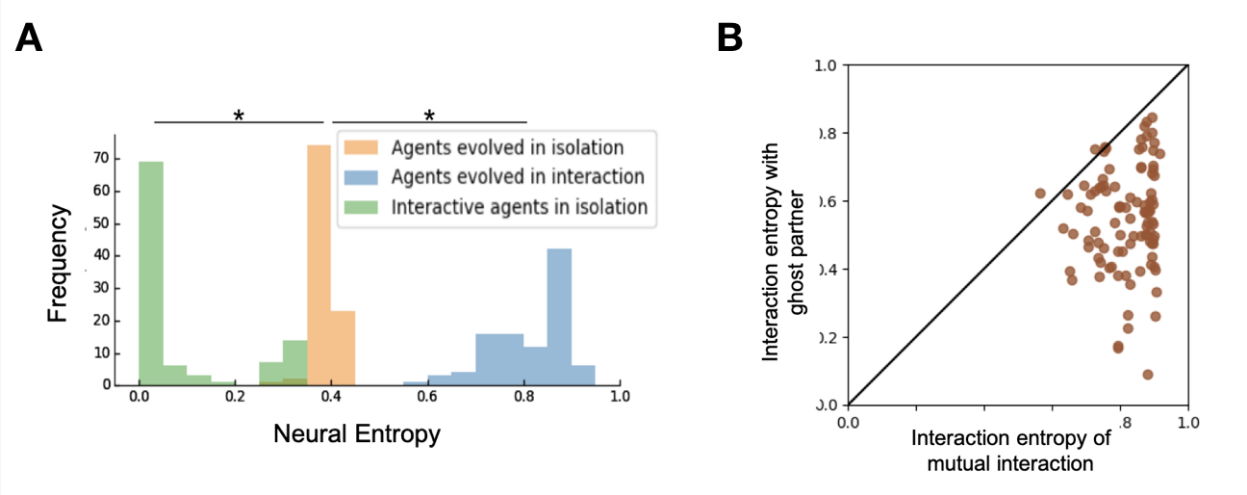

Figure 5.3: Results depicting effect of social interaction on neural complexity. [A] Fitness distributions of best agent in the population from 100 runs for each of the different levels of social interaction. Agents evolved with interaction (blue) showed highest neural complexity, however, when the same agents were evaluated in isolation (green) showed significantly lower neural complexity even compared to agents evolved in isolation (orange). [B] Neural entropy and behavior in the presence of an active partner versus ghost partner. All agents exhibit high values along the horizontal axis demonstrating high internal complexity in the presence of responding partners. However, as it can be seen from the spread along the vertical axis, below the diagonal, these agents lose internal complexity when their partner is a ghost. This loss tends to be more pronounced for higher levels of interaction entropy, which suggests that these higher levels are more readily achieved by interdependent rather than independent interaction.

5.5 Complex interactive behavior does not require high isolation entropy

From the previous results, it does not directly follow that agents that show high interaction entropy would also exhibit high entropy in isolation. This is an empirical question regarding the emergence of complex interactive behaviors from simple systems. In order to test this, we disabled the sensors in agents that were optimized in interactive environments and measured their neural entropy in isolation. These agents consistently showed lower levels of entropy than what they exhibited during interaction (Fig. 5.3A). Importantly, all of these agents also showed significantly lower levels of entropy than what was typically achieved by agents evolved in isolation to maximize isolation entropy (Fig. 5.3A; significance tests results in table 5.1). In other words, although these agents were more complex during interaction, they are not intrinsically more complex. This has implications for developmental psychology, since these results suggest that complex interactive

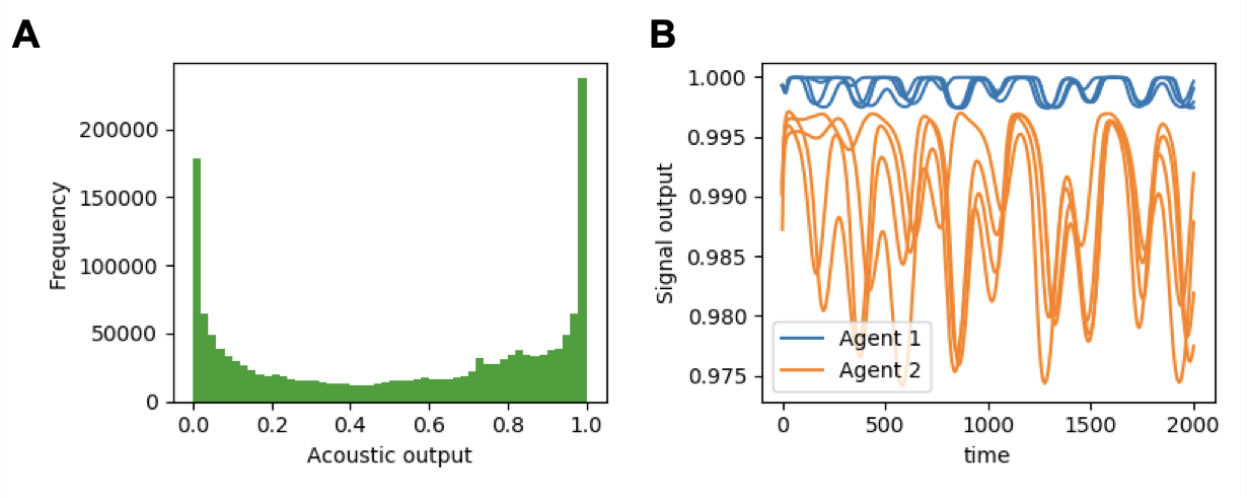

Figure 5.4: Emergent interaction upon evolving for neural complexity in the presence of interaction. [A] Distribution of strength of acoustic signal emitted by all 200 agents (one pair each of 100 different runs) that were evolved with interaction. While the acoustic signal show maximum frequency on either saturated ends, it can be seen that the acoustic signal bears a non-saturated value for a significant amount of time. [B] Illustration of acoustic signal for pair of agents shown in Fig. 5.2. Each trace corresponds to the acoustic signal emitted during one trial of the simulation by each agent in the presence of the other.

behaviors do not require high intrinsic internal complexity, as long as infants have the capacity to take advantage of the complexity provided by interaction.

5.6 Internal complexity is enhanced by interdependent interaction

This dependence on interaction with their partner to enhance neural complexity, and hence behavioral complexity, could be from two categorically different underlying interactive modes.

1. The partner could be a source of complex stimuli that drives the agent in question to perform behaviors through complexification of neural dynamics. In this case, the other agent becomes a passive component of a complex environment that the agent in question "uses" to realize complex neural dynamics. This mode of interaction is henceforth referred to as *independent interaction*.

2. The two agents could be engaged in mutually interdependent interactive behaviors, thereby bootstrapping neural complexity in each other through continuous interaction via acoustic

	Levene's test for Equality of Variances		t-test for equality of means						
	F	Sig	t	df	Sig. 2-tailed	mean difference	std. error difference	95% confidence interval	
								lower	upper
Equal variances assumed	195.013	.00	52.23	198	.000	.4233	.008	.407	.439
Equal variances not assumed			52.23	106	.000	.4233	.008	.4071	.4392

Table 5.1: Independent samples test between agents evolved in isolation and agents evolved in the presence of another agent

modulation and spatial navigation. In this case, the other agent is no longer passive but is an active responsive component that continuously influences and is influenced by the neural dynamics of the agent in question. This mode of interaction is henceforth referred to as *interdependent interaction*, which is a generic form of coordination.

In order to disambiguate the aforementioned two modes of independent and interdependent interaction, we measured interactive entropy in the presence of "ghost" partners. "Ghosts" were agents that were merely playing back their movements from a previous trial, without being responsive to the "live" agent whose neural entropy is being measured. The ghost condition preserves complexity of the signal that the live agent experiences, nevertheless, it does not present any opportunity for interdependent interaction or coordination. Under the ghost condition, live agents suffered a loss in internal complexity in most cases. This demonstrates that their neural complexities were enhanced by active interdependent interaction with the other agent, and not just because of the presence of complex driving signals (Fig. 5.3F).

The same pair of agents described in figure 5.3B were examined again in Figure 5.3E. This time, however, one of the agents was made into a "ghost" (same movement as before, but unresponsive to environmental feedback). As a result from this change, the live agent's behavior becomes starkly different and its entropy drops to 0.4712. This shows that the agents did not simply rely on the complex sensory stimuli from the behavior of the other agent. Instead, the two agents were

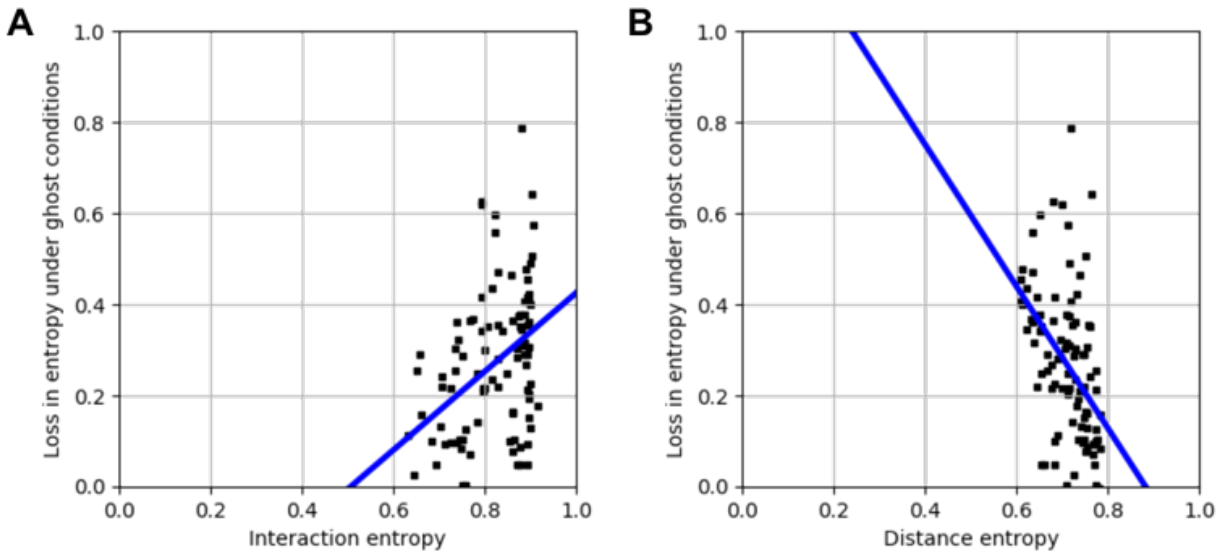

Figure 5.5: Relationship between interdependent interactions, behavioral complexity and internal complexity. [A] A positive correlation between interaction entropy and loss in entropy under ghost conditions suggests that interdependence increases with internal complexity. [B] A negative correlation between distance entropy (behavioral complexity) and loss in entropy under ghost conditions suggests that interdependence is hindered by greater behavioral complexity. Thus, interdependent interaction is achieved by trading-off internal complexity for behavioral complexity.

mutually interacting: they were coordinating their movements and were thereby enhancing each other's neural and behavioral complexity in a complementary manner. More generally, we found a statistically significant correlation between increasing internal complexity and interdependent interaction, and that this form of interaction tended to be more ordered, as would be expected from social coordination.

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% confidence interval for B	
	B	Std.error	Beta			Lower bound	Upper bound
(Constant)	-0.437	0.159		-2.753	.007	-0.751	-0.122
Interaction Entropy	0.862	0.193	0.411	4.462	.000	0.479	1.245

Table 5.2: Coefficients from the linear model fit with statistical test for the predictability of Loss in entropy under ghost conditions

R	R^2	Adjusted R^2	Std. error of estimate	Change Statistics				
				R^2 change	F change	df1	df2	Sig. F change
.411	.169	.160	.1529	.169	19.913	1	98	.000

Table 5.3: Model Summary for linear fit between interaction entropy and Loss in entropy under ghost condition. Predictors: (Constant), Interaction Entropy

5.7 Interdependent interaction occurs at a balance between behavioral complexity and internal interaction complexity

Greater loss in entropy between the active and the ghost conditions corresponds to greater levels of interdependent interaction. In order to study the relationship between an agent's complexity during interaction and the level of interdependence in it, we performed a linear fit between the two (Fig. 5.5A). A linear regression analysis showed that interaction entropy significantly predicted loss in entropy under the ghost condition (R2=0.16, F(1,98), p<.01) and beside the constant, interaction entropy was specifically tested using a t-statistic showing that it significantly predicted loss in entropy under ghost conditions (two-tailed t-test $p \ll 0.05$, see tables 5.2 and 5.3 for detailed statistical test results). This suggests that an increased level of internal complexity typically corresponds to greater levels of interdependent interaction.

On the other hand, when the entropy in distance between the agents, which is a proxy for behavioral complexity, was fit against loss in interaction entropy under ghost conditions, a negative correlation was revealed (Fig. 5.5B). The linear fit was found to be statistically significant (R2=0.208, F(1,98), $p \leq .01$) and distance entropy independently tested showed significant predictability of loss in entropy under ghost conditions (two-tailed t-test, $p \ll 0.05$, table 5.3). Thus, it can be said that greater levels of complexity in behavior hinders interdependent interaction. From the two trends we've shown in the factors that predict interdependent interaction, we posit that while mutually interdependent interaction enhances internal complexity, it is constrained by behavioral complexity that hinders the same. In other words, productive social behavior occurs at a fine balance between behavioral and internal complexity.

5.8 Agents exhibit higher-dimensional dynamics during interaction

From a dynamical systems perspective, in isolation, these simulated mobile robot systems are two-dimensional autonomous systems (2 neuronal states) that can at most have fixed-point or limit-cycle attractors (Beer, 1995c). During the course of interaction with another agent, these dynamical systems show aperiodic dynamics more complex than limit-cycles and that in principle require at least 3 dimensions (Fig. 5.2D). In the presence of another agent, the coupled system is of higher dimensionality involving both agents and their relative environmental states. In this case, measuring the entropy in one agent's neural activity, i.e. only analyzing the brain of one individual, is akin to measuring the entropy of the two-dimensional projection of a higher dimensional system. This explains the enhanced levels of internal complexity in agents that are in the presence of others – through their interaction the two embodied agents can become integrated into a larger, dynamically extended system (Froese and Fuchs, 2012).

5.9 Discussion

From a complex systems perspective we expected that placing embodied agents in an interactive context would transform their neural and behavioral dynamics, and that certain forms of interaction would lead to an increase in their complexity. Our modeling results confirmed this expectation by providing a proof of concept that the behavior of embodied agents in real-time dyadic interaction cannot be fully understood from studying their brains in isolation, nor even in the context of non-responsive social stimuli.

In our simulation model an agent's neural complexity could increase beyond its individual degrees of freedom when the agent is interacting with a complex environment, and especially so when it is coordinating its behavior with another responsive agent. Our analysis revealed that this increase is not just a matter of activating latent internal complexity: interaction allows an agent's neural activity to increase its complexity to such an extent that in principle it would be impossible for that activity to be generated in isolation. It remains to be seen to what extent this

increase in individual complexity scales with the number of individuals that are interacting as well as the size of the neural network. In the models presented in this work, the effect of social interaction was more pronounced because 2-dimensional agents that are not capable of producing dynamics more complex than oscillators in isolation, end up producing aperiodic activity when interacting with another agent. However, if each individual were composed of larger networks that are themselves capable of producing aperiodic dynamics, perhaps this effect might diminish. In any case, the dimensionality of an individual will still be only a fraction of the total dimensionality of the coupled multi-agent system and so it is expected that environmental interaction will need to be taken into consideration comprehensively study social cognition. Overall, this finding suggests that the enactive approach to social cognition is on the right track: the dynamical basis of an agent's behavior during real-time interaction with another agent becomes the whole brain-body-environment-body-brain system (Froese et al., 2013b), of which each agent's brain is just one important component (Gallagher et al., 2013) whose neural activity becomes a projection of the overarching interaction process. Future modeling work could analyze in more detail how this interactive expansion of individual complexity is dynamically realized, for example by analyzing the transformation of the state space of the overarching brain-body-environment-body-brain system as it goes from an uncoupled to a coupled mode.

Another avenue for future investigation is to verify these modeling findings in the context of actual human social interaction. The so-called "second-person" approach to social cognitive neuroscience has already revealed that the brain is activated differently when participants are engaged in real-time social interaction when compared to passive observer scenarios (Schilbach et al., 2013). The complex systems perspective adopted by the enactive approach could help to provide an explanation for this observed difference. More specifically, it would be interesting to verify our finding that an agent's neural activity tends to be transformed more substantially in scenarios involving interdependent compared to independent forms of interaction between agents. Importantly, our results reveal that interpersonal behavioral synchrony in itself is not sufficient to distinguish between interdependent and independent forms of interaction. Accordingly, future

experimental work could compare neural activity in a task requiring real-time coordination with neural activity in a non-responsive "playback" control condition, for instance by employing the human dynamic clamp paradigm (Dumas et al., 2014).

Chapter 6

Conclusion

The previous chapters have presented details of all the scientific contributions of this dissertation. In this final chapter, I present a summary of the learnings from our work, and provide some closing remarks.

6.1 Summary of contributions

Over the course of this dissertation, recurrent dynamical computational models of neural networks have been employed to study how including the environment in our analysis gives a more comprehensive picture of the neural basis of behavior. Specifically, such models were applied to three phenomena of adaptive behavior namely, predictive coding, multifunctionality and social cognition. Optimizing the models for specifically designed tasks, followed by information-theoretic analyses enabled us to make the following novel contributions:

1. Neural networks encode predictive information provided by stimuli as well as generate their own as needed. The source of predictive information dynamically changes over the course of a behavior. Mutual information cannot distinguish the source from which predictive information was encoded in a neural network. We demonstrated that estimating transfer of information about a future stimulus uniquely from the environmental stimuli and uniquely from the neural networks' past activity allows us to infer the dynamics of the source of predictive information. Further, predictive information provided by the environment is a consequence of its regularities and can vary within the same task. Therefore, encoding

predictive information alone is not a sufficient indicator of adaptation to the environment. Thus, we demonstrate that multivariate information-theoretic analysis that includes the environment enables a more comprehensive understanding predictive coding in dynamical neural networks.

2. An agent can perform multiple tasks using the same neural controller, same sensory and same motor systems in the absence of neuromodulation or plasticity. In such cases, estimating the effective network using transfer entropy for each task enables capturing the task-specific dynamics of the neural network that unfolds in the context of each task over the same fixed structural network. The relationship between behavioral performance and the task-specific effective network was as follows: all neural networks that possessed distinct task-specialized effective networks for each task, consistently performed the different behaviors with near-perfect accuracy; however, the converse was not true: different tasks could also be performed with effective networks that are not specialized for each task. When task-specific effective networks are not distinct for each task, neural dynamics are shared between the tasks at a finer level. Sharing of neural resources can be favorably exploited to perform multiple tasks. This can happen to the extent that indistinguishable neural dynamics produced behaviors that were starkly different upon observing the agent. Multifunctionality in this case can only be understood by also including the state of the environment in the analysis. Thus, we demonstrate that multifunctionality can emerge from fixed circuits when embedded and embodied, and that to comprehensively understand multifunctionality, it is essential to include the environment in the analysis.

3. Mutual interaction with other individuals in the environment enables richness in neural dynamics that cannot be achieved in isolation or in the presence of an unresponsive agent. Agents that are in continuous closed-loop interaction with another agent receive stimuli as a function of internal states of both agents. As a result, neural dynamics in an individual agent is part of the higher-dimensional system that includes the environment as well. We demonstrate

this by showing that individual agents exhibit neural dynamics that are theoretically not capable by an agent in isolation. Empirically, the neural complexity achievable by an agent during social interaction was significantly higher than what was achievable in one-way interaction or isolation. Thus, we demonstrate that a comprehensive study of neural network operation requires that the environment and on-going interaction with the environment is included in the analysis.

4. Theoretical insights from in each of the three domains studied have been discussed above. The common underlying observation across the three seemingly distinct phenomena of predictive coding, multifunctionality and social cognition are the benefits of taking environmental variables and agent-environment interaction seriously. With predictive coding, it revealed that predictive information is available for free from the environment and that any network can encode it. With multifunctionality, it provides an additional mechanisms that can enable multifunctionality: agent-environment interaction. Finally, with social interaction, it shows that real-time interaction enhances neural dynamics in a way that cannot be understood by merely studying the brain of an individual in isolation. Thus, as a general contribution to different areas within Neuroscience, this dissertation emphasizes the importance of including the environment in our study of the neural basis of adaptive behavior.

6.2 Concluding remarks

As embedded, embodied and dynamical systems, living organisms are in constant closed-loop interaction with their environment. Environmental regularities shape information available to the animals, which are further shaped by animals themselves based on their actions. As a result, behavior involves a continuous exchange of influence between the agent and the environment that are a coupled dynamical system. Since both the agent and the environment are dynamical systems in their own right, this closed-loop interaction cannot be simply replicated by presenting stimuli passively to animals. Technological advancements are continuously improving the ability of experimentalists to record neural activity from freely-behaving animals. To fully take advantage of

these advancements, appropriate analytical tools will be required to be developed. Work presented in this dissertation, provides theoretical insights as well as frameworks for analysis when data is obtained from closed-loop agent-environment interaction.

Potential for extensions of the work presented in this dissertation were discussed at the end of chapter. In addition to those, as a dissertation that has employed computational models, one natural extension is to apply these methods to experimental data, or devise experiments based on these models to test these results. Specifically, existing studies in predictive coding such as Palmer et al. (2015) employed experimental designs very similar to the ones that have been used in this dissertation and can already utilize the framework presented here. Similarly, with regards to the work presented here on multifunctionality, Lizier et al. (2011) employed multivariate analysis methods in fMRI data to study task-relevant changes in neural dynamics. While this study evaluated the task-specific changes, the same data could be used to evaluate the extent of overlap rather than the uniqueness for each task. Finally, with regards to the work presented on social interaction, Froese et. al. are currently working on replicating the isolation, one-way and two-way interaction conditions and measuring neural complexity to see if similar results can be observed in EEG data.

Toward the ultimate goal of discovering the general principles that underlie neural information processing as it pertains to behavior, this dissertation establishes the following: first, the benefits of using computational models to advance our understanding of the neural basis of behaviors; second, the ability of multivariate information-theoretic analysis and its application in-time to reveal information flow in agent-environment systems; and third, and most crucial, the importance of integrating the environment in our analyses to get a more comprehensive understanding of the neural basis of behaviors.

Bibliography

[1] Abbott, L. F. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain research bulletin*, 50(5-6):303–304.

[2] Abbott, L. F. (2008). Theoretical neuroscience rising. *Neuron*, 60(3):489–495.

[3] Ackley, D. and Littman, M. (1991). Interactions between learning and evolution. *Artificial life II*, 10:487–509.

[4] Agmon, E. and Beer, R. D. (2014). The evolution and analysis of action switching in embodied agents. *Adaptive Behavior*, 22(1):3–20.

[5] Alexander, W. H. and Brown, J. W. (2010). Computational models of performance monitoring and cognitive control. *Topics in cognitive science*, 2(4):658–677.

[6] Alexander, W. H. and Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature neuroscience*, 14(10):1338.

[7] Alexander, W. H. and Brown, J. W. (2015). Reciprocal interactions of computational modeling and empirical investigation. In *An introduction to model-based cognitive neuroscience*, pages 321–338. Springer.

[8] Alexander, W. H. and Brown, J. W. (2019). The role of the anterior cingulate cortex in prediction error and signaling surprise. *Topics in cognitive science*, 11(1):119–135.

[9] Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30(8):2960–2966.

[10] Anderson, M. L. (2010a). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4):245–266.

[11] Anderson, M. L. (2010b). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4):245–266.

[12] Anderson, M. L., Kinnison, J., and Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *Neuroimage*, 73:50–58.

[13] Anderson, M. L., Richardson, M. J., and Chemero, A. (2012). Eroding the boundaries of cognition: Implications of embodiment 1. *Topics in cognitive science*, 4(4):717–730.

[14] Andres, M., Seron, X., and Olivier, E. (2007). Contribution of hand motor circuits to counting. *Journal of Cognitive Neuroscience*, 19(4):563–576.

[15] Auvray, M., Lenay, C., and Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New ideas in psychology*, 27(1):32–47.

[16] Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B*, 63(3):329–339.

[17] Baltieri, M. and Buckley, C. L. (2017). An active inference implementation of phototaxis. In *Artificial Life Conference Proceedings 14*, pages 36–43. MIT Press.

[18] Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, 24(4):602–607.

[19] Barto, A., Sutton, R., and Anderson, C. (1984). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 5:834–846.

[20] Bassett, D. S. and Sporns, O. (2017). Network neuroscience. *Nature neuroscience*, 20(3):353.

[21] 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.

[22] Beer, R. D. (1995a). A dynamical systems perspective on agent-environment interaction. *Artificial intelligence*, 72(1-2):173–215.

[23] Beer, R. D. (1995b). On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509.

[24] Beer, R. D. (1995c). On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509.

[25] Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in cognitive sciences*, 4(3):91–99.

[26] Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4):209–243.

[27] Beer, R. D., Chiel, H. J., and Gallagher, J. C. (1999). Evolution and analysis of model cpgs for walking: Ii. general principles and individual variability. *Journal of computational neuroscience*, 7(2):119–147.

[28] Beer, R. D. et al. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. *From animals to animats*, 4:421–429.

[29] Beer, R. D. and Gallagher, J. C. (1992). Evolving dynamical neural networks for adaptive behavior. *Adaptive behavior*, 1(1):91–122.

[30] Beer, R. D. and Ritzmann, R. E. (1993). *Biological neural networks in invertebrate neuroethology and robotics*. Academic Pr.

[31] Beer, R. D. and Williams, P. L. (2009). Animals and animats: Why not both iguanas? *Adaptive Behavior*, 17(4):296–302.

[32] Beer, R. D. and Williams, P. L. (2015). Information processing and dynamics in minimally cognitive agents. *Cognitive science*, 39(1):1–38.

[33] Beggs, J. M. and Plenz., D. (2003). Neuronal avalanches in neocortical circuits. *Journal of neuroscience*, 23(35).

[34] Beheshti, Z. and Shamsuddin, S. M. H. (2013). A review of population-based meta-heuristic algorithms. *Int. J. Adv. Soft Comput. Appl*, 5(1):1–35.

[35] Bem, T., Cabelguen, J.-M., Ekeberg, Ö., and Grillner, S. (2003). From swimming to walking: a single basic network for two different behaviors. *Biological cybernetics*, 88(2):79–90.

[36] Berkowitz, A. (2002). Both shared and specialized spinal circuitry for scratching and swimming in turtles. *Journal of Comparative Physiology A*, 188(3):225–234.

[37] Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

[38] Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. (2014). Quantifying unique information. *Entropy*, 16(4):2161–2183.

[39] Besserve, M., Schölkopf, B., Logothetis, N. K., and Panzeri., S. (2010). Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *Journal of computational neuroscience*, 29(3):547–566.

[40] Beyer, H.-G. and Schwefel, H.-P. (2002). Evolution strategies–a comprehensive introduction. *Natural computing*, 1(1):3–52.

[41] Bialek, W. (2012). *Biophysics: searching for principles*. Princeton University Press.

[42] Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463.

[43] Bialek, W. and Tishby, N. (1999). Predictive information. *arXiv preprint cond-mat/9902341*.

[44] Bialek, W., Van Steveninck, R. R. D. R., and Tishby, N. (2006). Efficient representation as a design principle for neural coding and computation. In *2006 IEEE international symposium on information theory*, pages 659–663. IEEE.

[45] Blitz, D. M. and Nusbaum, M. P. (1997). Motor pattern selection via inhibition of parallel pathways. *Journal of Neuroscience*, 17(13):4965–4975.

[46] Blitz, D. M. and Nusbaum, M. P. (1999). Distinct functions for cotransmitters mediating motor pattern selection. *Journal of Neuroscience*, 19(16):6774–6783.

[47] Bolz, J. and Gilbert, C. D. (1986). Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature*, 320(6060):362–365.

[48] Borst, A. and Theunissen, F. E. (1999). Information theory and neural coding. *Nature neuroscience*, 2(11):947–957.

[49] Braver, T. S., Barch, D. M., and Cohen, J. D. (1999). Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biological psychiatry*, 46(3):312–328.

[50] Briggman, K. L. and Kristan, W. B. (2006a). Imaging dedicated and multifunctional neural circuits generating distinct behaviors. *Journal of Neuroscience*, 26(42):10925–10933.

[51] Briggman, K. L. and Kristan, W. B. (2006b). Imaging dedicated and multifunctional neural circuits generating distinct behaviors. *Journal of Neuroscience*, 26(42):10925–10933.

[52] Briggman, K. L. and Kristan, W. B. (2008). Multifunctional pattern-generating circuits. *Annual Review of Neuroscience*, 31:271–294.

[53] Brown, H., Friston, K. J., and Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in psychology*, 2:218.

[54] Brown, J. W. and Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307(5712):1118–1121.

[55] Brunel, N. (2000). Dynamics of networks of randomly connected excitatory and inhibitory spiking neurons. *Journal of Physiology-Paris.*, 94(5):445–463.

[56] Buckley, C. L., Fine, P., Bullock, S., and Paolo, E. D. (2008). Monostable controllers for adaptive behaviour. In *Int. Conf on Simulation of Adaptive Behavior*, pages 103–112.

[57] Butts, D. A. and Goldman, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS biology*, 4(4).

[58] Cahana-Amitay, D. and Albert, M. L. (2014). Brain and language: evidence for neural multifunctionality. *Behavioural neurology*, 2014.

[59] Candadai, M. and Izquierdo, E. (2018). Multifunctionality in embodied agents: Three levels of neural reuse. *arXiv preprint arXiv:1802.03891*.

[60] Candadai, M. and Izquierdo, E. (2020). infotheory: A c++/python package for multivariate information theoretic analysis. *Journal of Open Source Software*, 5(47):1609.

[61] Candadai, M. and Izquierdo, E. J. (2019a). infotheory: A c++/python package for multivariate information theoretic analysis. *arXiv preprint arXiv:1907.02339*.

[62] Candadai, M. and Izquierdo, E. J. (2019b). Sources of predictive information in dynamical neural networks. *bioRxiv*.

[63] Candadai, M., Setzler, M., Izquierdo, E. J., and Froese, T. (2019). Embodied dyadic interaction increases complexity of neural dynamics: A minimal agent-based simulation model. *Frontiers in psychology*, 10:540.

[64] Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., and Nutt, D. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, 8:20.

[65] Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science*, 15(366):92–96.

[66] Chao, L. L. and Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, 12(4):478–484.

[67] Chao, Z. C., Takaura, K., Wang, L., Fujii, N., and Dehaene, S. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. *Neuron*, 100(5):1252–1266.

[68] Chen, K. S., Chen, C.-C., and Chan, C. (2017). Characterization of predictive behavior of a retina by mutual information. *Frontiers in computational neuroscience*, 11:66.

[69] Chiappalone, M., Bove, M., Vato, A., Tedesco, M., and Martinoia., S. (2006). Dissociated cortical networks show spontaneously correlated activity patterns during in vitro development. *Brain research*, 1093(1).

[70] Chiel, H. J. and Beer, R. D. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in neurosciences*, 20(12):553–557.

[71] Chiel, H. J. and Beer., R. D. (1997). The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in neurosciences*, 20(12):553–557.

[72] Chiel, H. J. and Beer, R. D. (2008). Computational neuroethology. *Encyclopedia of the Neuroscience. Elsevier*, 54.

[73] Chiel, H. J., Beer, R. D., and Gallagher, J. C. (1999). Evolution and analysis of model cpgs for walking: I. dynamical modules. *Journal of computational neuroscience*, 7(2):99–118.

[74] Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

[75] Cliff, D., Husbands, P., and Harvey, I. (1993). Explorations in evolutionary robotics. *Adaptive behavior*, 2(1):73–110.

[76] Cohen, J. E. (2004). Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS biology*, 2(12).

[77] Combes, D., Meyrand, P., and Simmers, J. (1999a). Dynamic restructuring of a rhythmic motor program by a single mechanoreceptor neuron in lobster. *Journal of Neuroscience*, 19(9):3620–3628.

[78] Combes, D., Meyrand, P., and Simmers, J. (1999b). Motor pattern specification by dual descending pathways to a lobster rhythm-generating network. *Journal of Neuroscience*, 19(9):3610–3619.

[79] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[80] Crisp, K. M. and Mesce, K. A. (2004). A cephalic projection neuron involved in locomotion is dye coupled to the dopaminergic neural network in the medicinal leech. *Journal of Experimental Biology*, 207(26):4535–4542.

[81] D. M. Blitz, A. E. Christie, M. J. C. B. J. N. E. M. and Nusbaum, M. P. (1999). Different proctolin neurons elicit distinct motor patterns from a multifunctional neuronal network. *Journal of Neuroscience*, 19(13):5449–5463.

[82] D. Thalmeier, M. Uhlmann, H. K. and Memmesheimer, R. (2016). Learning universal computations with spikes. *PLOS Comput. Biol*, 12(6).

[83] Dagher, A., Owen, A. M., Boecker, H., and Brooks, D. J. (1999). Mapping the network for planning: a correlational pet activation study with the tower of london task. *Brain*, 122(10):1973–1987.

[84] Damasio, A. R. and Tranel, D. (1993). Nouns and verbs are retrieved with differently

distributed neural systems. *Proceedings of the National Academy of Sciences*, 90(11):4957–4960.

[85] Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., and Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, 380(6574):499–505.

[86] Datta, S. R., Anderson, D. J., Branson, K., Perona, P., and Leifer, A. (2019). Computational neuroethology: A call to action. *Neuron*, 104(1):11–24.

[87] Dawson, M. R. (1998). *Understanding cognitive science*. Blackwell Oxford.

[88] Dayan, P. and Abbott, L. F. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems.

[89] Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, 7(5):889–904.

[90] De Jaegher, H., Di Paolo, E., and Adolphs, R. (2016). What does the interactive brain hypothesis mean for social neuroscience? a dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1693):20150379.

[91] De Jaegher, H., Di Paolo, E., and Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in cognitive sciences*, 14(10):441–447.

[92] Den Ouden, H. E., Friston, K. J., Daw, N. D., McIntosh, A. R., and Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral cortex*, 19(5):1175–1185.

[93] Desimone, R. and Schein, S. J. (1987). Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of neurophysiology*, 57(3):835–868.

[94] DeWeese, M. R. and Meister., M. (1999). How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 10(4):325–340.

[95] Di Paolo, E. A. (2000). Behavioral coordination, structural congruence and entrainment in a simulation of acoustically coupled agents. *Adaptive Behavior*, 8(1):27–48.

[96] Di Paolo, E. A. and De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in human neuroscience*, 6:163.

[97] Di Paolo, E. A., Noble, J., and Bullock, S. (2000). Simulation models as opaque thought experiments.

[98] Di Paolo, E. A., Rohde, M., and Iizuka, H. (2008). Sensitivity to social contingency or stability of interaction? modelling the dynamics of perceptual crossing. *New ideas in psychology*, 26(2):278–294.

[99] Dimitrov, A. G., Lazar, A. A., and Victor, J. D. (2011). Information theory in neuroscience. *Journal of computational neuroscience*, 30(1):1–5.

[100] Dorigo, M. and Di Caro, G. (1999). Ant colony optimization: a new meta-heuristic. In *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, volume 2, pages 1470–1477. IEEE.

[101] Dotov, D. and Froese, T. (2018). Entraining chaotic dynamics: A novel movement sonification paradigm could promote generalization. *Human movement science*, 61:27–41.

[102] Doya, K., Ishii, S., Pouget, A., and Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

[103] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159.

[104] Dumas, G., de Guzman, G. C., Tognoli, E., and Kelso, J. S. (2014). The human dynamic clamp as a paradigm for social interaction. *Proceedings of the National Academy of Sciences*, 111(35):E3726–E3734.

[105] E. S. Schaffer, S. O. and Abbott, L. F. (2013). A complex-valued firing-rate model that approximates the dynamics of spiking networks. *PLoS Comput Biol*, 9(10).

[106] Egner, T., Monti, J. M., and Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, 30(49):16601–16608.

[107] Egner, T. and Summerfield, C. (2013). Grounding predictive coding models in empirical neuroscience research. *Behavioral and Brain Sciences*, 36(3):210–211.

[108] Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12.

[109] Eugene M. Izhikevich, Niraj S. Desai, E. C. W. and Hoppensteadt., F. C. (2003). Bursts as a unit of neural information: selective communication via resonance. *Trends in neurosciences*, 26(3):161–167.

[110] F. J. Varela, E. T. and Rosch, E. (1991). *The Embodied Mind*. MIT Press, Cambridge, Massachusetts.

[111] F. Rieke, D. Warland, R. d. R. v. S. and Bialek., W. (1996). *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, Massachusetts.

[112] Faber, S. P., Timme, N. M., Beggs, J. M., and Newman, E. L. (2019). Computation is concentrated in rich clubs of local cortical networks. *Network Neuroscience*, 3(2):384–404.

[113] Fernando, T., Maier, H., and Dandy, G. (2009). Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology*, 367(3-4):165–176.

[114] Ferster., D. and Lindstrom, S. (1983). An intracellular analysis of geniculo-cortical connectivity in area 17 of the cat. *Journal of physiology*, 342(181).

[115] FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445.

[116] Floreano, D., Dürr, P., and Mattiussi, C. (2008). Neuroevolution: from architectures to learning. *Evolutionary intelligence*, 1(1):47–62.

[117] Fouad, A. D., Teng, S., Mark, J. R., Liu, A., Alvarez-Illera, P., Ji, H., Du, A., Bhirgoo, P. D., Cornblath, E., Guan, S. A., et al. (2018). Distributed rhythm generators underlie caenorhabditis elegans forward locomotion. *Elife*, 7:e29913.

[118] Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

[119] Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301.

[120] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127.

[121] Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221.

[122] Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78.

[123] Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS one*, 4(7).

[124] Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302.

[125] Froese, T. (2018). Searching for the conditions of genuine intersubjectivity: From agent-based models to perceptual crossing experiments. In *The Oxford Handbook of 4E Cognition*.

[126] Froese, T. and Di Paolo, E. A. (2009). Toward minimally social behavior: social psychology meets evolutionary robotics. In *European Conference on Artificial Life*, pages 426–433. Springer.

[127] Froese, T. and Di Paolo, E. A. (2010). Modelling social interaction as perceptual crossing: an investigation into the dynamics of the interaction process. *Connection Science*, 22(1):43–68.

[128] Froese, T. and Di Paolo, E. A. (2011). The enactive approach: Theoretical sketches from cell to society. *Pragmatics & Cognition*, 19(1):1–36.

[129] Froese, T. and Fuchs, T. (2012). The extended body: a case study in the neurophenomenology of social interaction. *Phenomenology and the Cognitive Sciences*, 11(2):205–235.

[130] Froese, T., Gershenson, C., and Rosenblueth, D. A. (2013a). The dynamically extended mind. In *Evolutionary computation (CEC), 2013 IEEE Congress on*, pages 1419–1426. IEEE.

[131] Froese, T., Iizuka, H., and Ikegami, T. (2013b). From synthetic modeling of social interaction to dynamic theories of brain–body–environment–body–brain systems. *Behavioral and Brain Sciences*, 36(4):420–421.

[132] Funahashi, K.-i. and Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806.

[133] Gallagher, S., Hutto, D. D., Slaby, J., and Cole, J. (2013). The brain as part of an enactive system. *Behavioral and Brain Sciences*, 36(4):421–422.

[134] Gallotti, M. and Frith, C. D. (2013). Social cognition in the we-mode. *Trends in cognitive sciences*, 17(4):160–165.

[135] Gao, S., Guan, S. A., Fouad, A. D., Meng, J., Kawano, T., Huang, Y.-C., Li, Y., Alcaire, S., Hung, W., Lu, Y., et al. (2018). Excitatory motor neurons are local oscillators for backward locomotion. *Elife*, 7:e29915.

[136] Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychological science*, 4(6):385–390.

[137] Gemba, H., Sasaki, K., and Brooks, V. (1986). 'error'potentials in limbic cortex (anterior cingulate area 24) of monkeys during motor learning. *Neuroscience letters*, 70(2):223–227.

[138] Gentner, D. and Kurtz, K. J. (2005). Relational categories.

[139] Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., and Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, 2(11):1527–1537.

[140] Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419.

[141] Gerhard, F., Pipa, G., Lima, B., Neuenschwander, S., and Gerstner, W. (2011). Extraction of network topology from multi-electrode recordings: is there a small-world effect? *Frontiers in computational neuroscience*, 5:4.

[142] Getting, P. A. (1989). Emerging principles governing the operation of neural networks. *Annual Review of Neuroscience*, 12:185–204.

[143] Getting, P. A. and Dekin, M. S. (1985). Tritonia swimming. In *Model neural networks and behavior*, pages 3–20. Springer.

[144] Giurfa, M., Zhang, S., Jenett, A., Menzel, R., and Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference'in an insect. *Nature*, 410(6831):930.

[145] Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning.

[146] Gourévitch, B. and Eggermont., J. J. (2007a). Evaluating information transfer between auditory cortical neurons. *Journal of Neurophysiology*, 97:2533–2543.

[147] Gourévitch, B. and Eggermont., J. J. (2007b). Evaluating information transfer between auditory cortical neurons. *Journal of Neurophysiology*, 97(3):2533–2543.

[148] Graham, D. and Field, D. (2007). Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. *Spatial vision*, 21(1-2):149–164.

[149] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

[150] Griffith, V., Chong, E., James, R., Ellison, C., and Crutchfield, J. (2014). Intersection information based on common randomness. *Entropy*, 16(4):1985–2000.

[151] Griffith, V. and Koch, C. (2014). Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pages 159–190. Springer.

[152] Harder, M., Salge, C., and Polani, D. (2013). Bivariate measure of redundant information. *Physical Review E*, 87(1):012130.

[153] Harvey, I. (1997). Cognition is not computation; evolution is not optimisation. In *International Conference on Artificial Neural Networks*, pages 685–690. Springer.

[154] Harvey, I. (2000). Robotics: Philosophy of mind using a screwdriver. *Evolutionary robotics: From intelligent robots to artificial life*, 3:207–230.

[155] Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. (1996). Evolutionary robotics at sussex. In *Proc. Intl. Symposium on Robotics and Manufacturing*.

[156] Harvey, I., Paolo, E. D., Wood, R., Quinn, M., and Tuci, E. (2005). Evolutionary robotics: A new scientific tool for studying cognition. *Artificial life*, 11(1-2):79–98.

[157] Helmholtz, H. (1860). 1962 handbuch der physiologischen optik (ed. jpc southall), vol. 3.

[158] Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161.

[159] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., and Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46.

[160] Hobert, O. (2003). Behavioral plasticity in *C. elegans*: paradigms, circuits, genes. *Developmental Neurobiology*, 54(1):203–223.

[161] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[162] Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544.

[163] Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

[164] Holland, J. H. et al. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

[165] Honey, C. J., Kötter, R., Breakspear, M., and Sporns., O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences*, 104(24):10240–10245.

[166] Hooper, S. L. and DiCaprio, R. A. (2004). Crustacean motor pattern generator networks. *Neurosignals*, 13(1-2):50–69.

[167] Huang, Y. and Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593.

[168] Hubel, D. H. and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of neurophysiology*, 28(2):229–289.

[169] Ijspeert, A. J., Crespi, A., Ryczko, D., and Cabelguen, J.-M. (2007). From swimming to walking with a salamander robot driven by a spinal cord model. *science*, 315(5817):1416–1420.

[170] Ikeda, S. and Manton, J. H. (2009). Capacity of a single spiking neuron channel. *Neural Computation*, 21(6):1714–1748.

[171] Ince, R. (2017). Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318.

[172] Ince, R., Bartolozzi, C., and Panzeri, S. (2009). An information-theoretic library for the analysis of neural codes.

[173] Ince, R. A., Senatore, R., Arabzadeh, E., Montani, F., Diamond, M. E., and Panzeri, S. (2010). Information-theoretic methods for studying population codes. *Neural Networks*, 23(6):713–727.

[174] Ito, S., Hansen, M. E., Heiland, R., Lumsdaine, A., Litke, A. M., and Beggs, J. M. (2011a). Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. *PloS one*, 6(11).

[175] Ito, S., Hansen, M. E., Heiland, R., Lumsdaine, A., Litke, A. M., and Beggs., J. M. (2011b). Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. *PLoS One*, 6(11).

[176] Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6).

[177] Izhikevich, E. M. and Moehlis, J. (2008). Dynamical systems in neuroscience: The geometry of excitability and bursting. *SIAM review*, 50(2):397.

[178] Izquierdo, E. and Beer., R. D. (2016). The whole worm: brain,body,environment models of c. elegans. *Current Opinion in Neurobiology*, 40:23–30.

[179] Izquierdo, E. and Bührmann, T. (2008). Analysis of a dynamical recurrent neural network evolved for two qualitatively different tasks: Walking and chemotaxis. In *ALIFE*, pages 257–264.

[180] Izquierdo, E. J. (2019). Role of simulation models in understanding the generation of behavior in c. elegans. *Current Opinion in Systems Biology*, 13:93–101.

[181] Izquierdo, E. J. and Buhrmann, T. (2008). Analysis of a dynamical recurrent neural network evolved for two qualitatively different tasks: Walking and chemotaxis. In *Proc. of the 11th Int. Conf. on Artificial Life*, pages 257–264.

[182] Izquierdo, E. J., Williams, P. L., and Beer, R. D. (2015a). Information flow through a model of the c. elegans klinotaxis circuit. *PloS one*, 10(10).

[183] Izquierdo, E. J., Williams, P. L., and Beer, R. D. (2015b). Information flow through a model of the c. elegans klinotaxis circuit. *PloS one*, 10(10):e0140397.

[184] Izquierdo-Torres, E. and Harvey, I. (2005). Learning to discriminate between multiple possible environments: an imprinting scenario. In *Memory and Learning Mechanisms in Autonomous Robots Workshop (ECAL 2005)*.

[185] James, R. and Crutchfield, J. (2017). Multivariate dependence beyond shannon information. *Entropy*, 19(10):531.

[186] James, R. G., Barnett, N., and Crutchfield, J. P. (2016). Information flows? a critique of transfer entropies. *Physical review letters*, 116(23):238701.

[187] James, R. G., Ellison, C. J., and Crutchfield, J. P. (2011). Anatomy of a bit: Information in a time series observation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037109.

[188] James, R. G., Ellison, C. J., and Crutchfield, J. P. (2018). dit: a python package for discrete information theory. *Journal of Open Source Software*, 3(25):738.

[189] Jehee, J. F. and Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS computational biology*, 5(5).

[190] Jessup, R. K., Busemeyer, J. R., and Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *Journal of Neuroscience*, 30(9):3467–3472.

[191] Joseph T. Lizier, Jakob Heinzle, A. H. J.-D. H. and Prokopenko., M. (2011). Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of computational neuroscience*, 30(1):85–107.

[192] Juergens, E. and Eckhorn., R. (1997). Parallel processing by a homogeneous group of coupled model neurons can enhance, reduce and generate signal correlations. *Biological Cybernetics*, 76(3).

[193] Jung, T. I., Vogiatzian, F., Har-Shemesh, O., Fitzsimons, C. P., and Quax, R. (2014). Applying information theory to neuronal networks: from theory to experiments. *Entropy*, 16(11):5721–5737.

[194] Kaiser, A. and Schreiber, T. (2002). Information transfer in continuous processes. *Physica D*, 166:43–62.

[195] Katz, P. S. and Harris-Warrick, R. M. (1991). Recruitment of crab gastric mill neurons into the pyloric motor pattern by mechanosensory afferent stimulation. *Journal of neurophysiology*, 65(6):1442–1451.

[196] Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE.

[197] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[198] Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490.

[199] Krishnan, G. P., González, O. C., and Bazhenov, M. (2018). Origin of slow spontaneous resting-state neuronal fluctuations in brain networks. *Proceedings of the National Academy of Sciences*, 115(26):6858–6863.

[200] Kristan, W., Wittenberg, G., Nusbaum, M., and Stern-Tomlinson, W. (1988). Multifunctional interneurons in behavioral circuits of the medicinal leech. *Experientia*, 44(5):383–389.

[201] Kurtz, K. J. and Boukrina, O. (2004). Learning relational categories by comparison of paired examples. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

[202] L. F. Abbott, B. D. and Memmesheimer, R. (2016). Building functional networks of spiking model neurons. *Nature*, 201(6).

[203] Lee, C., Rohrer, W. H., and Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, 332(6162):357–360.

[204] Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448.

[205] Levine, A. B., Schlosser, C., Grewal, J., Coope, R., Jones, S. J., and Yip, S. (2019). Rise of the machines: Advances in deep learning for cancer diagnosis. *Trends in cancer*.

[206] Lieske, S., Thoby-Brisson, M., Telgkamp, P., and Ramirez, J. (2000). Reconfiguration of the neural network controlling multiple breathing patterns: eupnea, sighs and gasps. *Nature neuroscience*, 3(6):600–607.

[207] Lindner, M., Vicente, R., Priesemann, V., and Wibral., M. (2011). Trentool: A matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC neuroscience*, 12(1).

[208] Lisman, J. E. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in neurosciences*, 20(1):38–43.

[209] Lizier, J., Bertschinger, N., Jost, J., and Wibral, M. (2018). Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work.

[210] Lizier, J. T. (2014). Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11.

[211] Lizier, J. T., Heinzle, J., Horstmann, A., Haynes, J.-D., and Prokopenko, M. (2011). Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of computational neuroscience*, 30(1):85–107.

[212] Lockery, S. R. and Goodman, M. B. (2009). The quest for action potentials in c. elegans neurons hits a plateau. *Nature neuroscience*, 12(4):377–378.

[213] London, M. and Häusser, M. (2005). Dendritic computation. *Annu. Rev. Neurosci.*, 28:503–532.

[214] Lorenz, R., Monti, R. P., Violante, I. R., Anagnostopoulos, C., Faisal, A. A., Montana, G., and Leech, R. (2016). The automatic neuroscientist: A framework for optimizing experimental design with closed-loop real-time fmri. *NeuroImage*, 129:320–334.

[215] Loshchilov, I. and Hutter, F. (2016). Cma-es for hyperparameter optimization of deep neural networks. *arXiv preprint arXiv:1604.07269*.

[216] Lurie, D. J., Kessler, D., Bassett, D. S., Betzel, R. F., Breakspear, M., Keilholz, S., Kucyi, A., Liégeois, R., Lindquist, M. A., McIntosh, A. R., et al. (2019). Questions and controversies in the study of time-varying functional connectivity in resting fmri. *Network Neuroscience*, pages 1–40.

[217] M. D. Fox, A. Z. Snyder, J. L. V. M. C. D. C. V. E. and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9673–9678.

[218] M. Garofalo, T. Nieus, P. M. and Martinoia., S. (2009). Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks. *PLoS one*, 4(8).

[219] M. Shimono, J. M. B. (2015). Functional clusters, hubs, and communities in the cortical microconnectome. *Cerebral Cortex*, 25(10):3743–3757.

[220] Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9):1659–1671.

[221] MacKay, D. M. and McCulloch, W. S. (1952). The limiting information capacity of a neuronal link. *The bulletin of mathematical biophysics*, 14(2):127–135.

[222] Mainen, Z. F. and Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506.

[223] Marder, E. (1994). Invertebrate neurobiology: Polymorphic neural networks. *Current Biology*, 4(8):752–754.

[224] Marder, E. and Bucher, D. (2001). Central pattern generators and the control of rhythmic movements. *Current biology*, 11(23):R986–R996.

[225] Marder, E. and Bucher, D. (2007). Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Annu. Rev. Physiol.*, 69:291–316.

[226] Markman, A. B. and Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(4):329–358.

[227] Marr, D. (1980). Visual information processing: The structure and creation of visual representations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):199–218.

[228] Marsh, K. L., Richardson, M. J., and Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, 1(2):320–339.

[229] Martin, A., Ungerleider, L. G., and Haxby, J. V. (2000). Category specificity and the brain: The sensory/motor model of semantic representations of objects. *The new cognitive neurosciences*, 2:1023–1036.

[230] Martin, A., Wiggs, C. L., Ungerleider, L. G., and Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, 379(6566):649–652.

[231] Martius, G., Der, R., and Ay, N. (2013). Information driven self-organization of complex robotic behaviors. *PloS one*, 8(5).

[232] McDonnell, M. D., Ikeda, S., and Manton, J. H. (2011). An introductory review of information theory in the context of computational neuroscience. *Biological cybernetics*, 105(1):55.

[233] Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

[234] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[235] Moioli, R. C. and Husbands., P. (2013). Neuronal assembly dynamics in supervised and unsupervised learning scenarios. *Neural Computation*, 25(11):2934–2975.

[236] Möller, R., Lambrinos, D., Pfeifer, R., Labhart, T., and Wehner, R. (1998). Modeling ant navigation with an autonomous agent. *From animals to animats*, 5:185–194.

[237] Montague, P. R. and Sejnowski, T. J. (1994). The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. *Learning & Memory*, 1(1):1–33.

[238] Morton, D. W. and Chiel, H. J. (1994). Neural architectures for adaptive behavior. *Trends in Neurosciences*, 17(10):413–420.

[239] Murray, L. (1985). Emotional regulations of interactions between two-month-oldsand their mothers. *Social perception in infants*, pages 177–197.

[240] N. Kopell, G. B. Ermentrout, M. A. W. and Traub, R. D. (2000). Gamma rhythms and beta rhythms have different synchronization properties. *Proceedings of the National Academy of Sciences*, 97(4):1867–1872.

[241] N. Kriegeskorte, R. G. and Bandettini., P. (2006). Information-based functional brain mapping. *Proceedings of the National academy of Sciences of the United States of America*, 103(10):3863–3868.

[242] N. Timme, S. Ito, M. M. F.-C. Y. E. H. P. H. and Beggs., J. M. (2014). Multiplex networks of cortical and hippocampal neurons revealed at different timescales. *PLoS one*, 9(12).

[243] Nadel, J., Carchon, I., Kervella, C., Marcelli, D., and Réserbat-Plantey, D. (1999). Expectancies for social contingency in 2-month-olds. *Developmental science*, 2(2):164–173.

[244] Naeem, M., McDaid, L. J., Harkin, J., Wade, J. J., and Marsland, J. (2015). On the role of astroglial syncytia in self-repairing spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2370–2380.

[245] Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070.

[246] Nersessian, N., Chandrasekharan, S., and Subramanian, V. (2012). Computational modeling: Is this the end of thought experimenting in science. *Thought experiments in philosophy, science and the arts*, pages 239–260.

[247] Nigam, S., Shimono, M., Ito, S., Yeh, F.-C., Timme, N., Myroshnychenko, M., Lapish, C. C., Tosi, Z., Hottowy, P., Smith, W. C., et al. (2016). Rich-club organization in effective connectivity among cortical neurons. *Journal of Neuroscience*, 36(3):670–684.

[248] Nusbaum, M. P. and Kristan, W. (1986). Swim initiation in the leech by serotonin-containing interneurones, cells 21 and 61. *Journal of Experimental Biology*, 122(1):277–302.

[249] Nusbaum, M. P. and Marder, E. (1989a). A modulatory proctolin-containing neuron (mpn). i. identification and characterization. *Journal of Neuroscience*, 9(5):1591–1599.

[250] Nusbaum, M. P. and Marder, E. (1989b). A modulatory proctolin-containing neuron (mpn). ii. state-dependent modulation of rhythmic motor activity. *Journal of Neuroscience*, 9(5):1600–1607.

[251] Okatan, M., Wilson, M. A., and Brown, E. N. (2005). Analyzing functional connectivity

using a network likelihood model of ensemble neural spiking activity. *Neural computation*, 17(9):1927–1961.

[252] O'keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

[253] Olivares, E. O., Izquierdo, E. J., and Beer, R. D. (2018). Potential role of a ventral nerve cord central pattern generator in forward and backward locomotion in caenorhabditis elegans. *Network Neuroscience*, 2(3):323–343.

[254] Osterhout, L. and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806.

[255] Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913.

[256] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

[257] Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262.

[258] Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999.

[259] Pine, D. S., Grun, J., Maguire, E. A., Burgess, N., Zarahn, E., Koda, V., Fyer, A., Szeszko, P. R., and Bilder, R. M. (2002). Neurodevelopmental aspects of spatial navigation: a virtual reality fmri study. *Neuroimage*, 15(2):396–406.

[260] Popescu, I. R. and Frost, W. N. (2002). Highly dissimilar behaviors mediated by a multifunctional network in the marine mollusk tritonia diomedea. *Journal of Neuroscience*, 22(5):1985–1993.

[261] Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

[262] Prescott, T. J., González, F. M. M., Gurney, K., Humphries, M. D., and Redgrave, P. (2006). A robot model of the basal ganglia: behavior and intrinsic processing. *Neural networks*, 19(1):31–61.

[263] R. Pfeifer, M. L. and Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science*, 318(5853):1088–1093.

[264] R. Vicente, M. Wibral, M. L. and Pipa., G. (2011). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67.

[265] Ramirez, S., Liu, X., Lin, P.-A., Suh, J., Pignatelli, M., Redondo, R. L., Ryan, T. J., and Tonegawa, S. (2013). Creating a false memory in the hippocampus. *Science*, 341(6144):387–391.

[266] Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79.

[267] Rapoport, A. and Horvath, W. J. (1960). The theoretical channel capacity of a single neuron as determined by various coding systems. *Information and control*, 3(4):335–350.

[268] Real, E., Asari, H., Gollisch, T., and Meister, M. (2017). Neural circuit inference from function to structure. *Current Biology*, 27(2):189–198.

[269] Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D., and Saxe, R. (2010). Live face-to-face interaction during fmri: a new tool for social cognitive neuroscience. *Neuroimage*, 50(4):1639–1647.

[270] Rieke, F., Warland, D., Van Steveninck, R. d. R., Bialek, W. S., et al. (1999). *Spikes: exploring the neural code*, volume 7. MIT press Cambridge.

[271] Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nature neuroscience*, 3(11):1199–1204.

[272] Riley, M. A., Richardson, M., Shockley, K., and Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in psychology*, 2:38.

[273] Ritter, D. A., Bhatt, D. H., and Fetcho, J. R. (2001). In vivo imaging of zebrafish reveals differences in the spinal networks for escape and swimming movements. *Journal of Neuroscience*, 21(22):8956–8965.

[274] Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

[275] Roux, F.-E., Boetto, S., Sacko, O., Chollet, F., and Trémoulet, M. (2003). Writing, calculating, and finger recognition in the region of the angular gyrus: a cortical stimulation study of gerstmann syndrome. *Journal of neurosurgery*, 99(4):716–727.

[276] Rubinov, M. and Sporns., O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3).

[277] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[278] Rusconi, E., Walsh, V., and Butterworth, B. (2005). Dexterity with numbers: rtms over left angular gyrus disrupts finger gnosis and number processing. *Neuropsychologia*, 43(11):1609–1624.

[279] Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach.

[280] S. Ito, M. E. Hansen, R. H.-A. L. A. M. L. and Beggs, J. M. (2011). Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. *Plos one*, 6(11).

[281] Saldanha, E. L. and Bitterman, M. E. (1951). Relational learning in the rat. *The American Journal of Psychology*, 64(1):37–53.

[282] Sayood, K. (2018). Information theory and cognition: A review. *Entropy*, 20(9):706.

[283] Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K., and Muthukumaraswamy, S. D. (2017). Increased spontaneous meg signal diversity for psychoactive doses of ketamine, lsd and psilocybin. *Scientific reports*, 7:46421.

[284] Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., and Vogeley, K. (2013). Toward a second-person neuroscience 1. *Behavioral and brain sciences*, 36(4):393–414.

[285] Schreiber, T. (2000a). Measuring information transfer. *Physical review letters*, 85(2):461.

[286] Schreiber, T. (2000b). Measuring information transfer. *Physical review letters*, 85(2):461.

[287] Schreiber., T. (2000). Measuring information transfer. *Physical review letters*, 85(2).

[288] Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

[289] Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

[290] Scott, D. W. (1985a). Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics*, pages 1024–1040.

[291] Scott, D. W. (1985b). Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics*, pages 1024–1040.

[292] Scott, D. W. (2012). Multivariate density estimation and visualization. In *Handbook of computational statistics*, pages 549–569. Springer.

[293] Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the american Statistical association*, 82(400):1131–1146.

[294] Searle, J. R. (1990). Collective intentions and actions. *Intentions in communication*, 401:401.

[295] Sederberg, A. J., MacLean, J. N., and Palmer, S. E. (2018). Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. *Proceedings of the National Academy of Sciences*, 115(5):1105–1110.

[296] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

[297] Shannon, C. E. and W., W. (1949). *The mathematical theory of communication*. University of Illinois Press, Urbana, Illinois.

[298] Shimono, M. and Beggs., J. M. (2014). Functional clusters, hubs, and communities in the cortical microconnectome. *Cerebral Cortex*, 25(10).

[299] Siegler, M. and Burrows, M. (1979). The morphology of local non-spiking interneurones in the metathoracic ganglion of the locust. *Journal of Comparative Neurology*, 183(1):121–147.

[300] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

[301] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

[302] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

[303] Simmers, J. and Moulins, M. (1988). A disynaptic sensorimotor pathway in the lobster stomatogastric system. *Journal of neurophysiology*, 59(3):740–756.

[304] Soffe, S. (1997). The pattern of sensory discharge can determine the motor response in young xenopus tadpoles. *Journal of Comparative Physiology A*, 180(6):711–715.

[305] Sparks, D. L., Holland, R., and Guthrie, B. L. (1976). Size and distribution of movement fields in the monkey superior colliculus. *Brain research*, 113(1):21–34.

[306] Sporns, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nature neuroscience*, 17(5):652.

[307] Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4).

[308] Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459.

[309] Stein, R. B. (1967). The information capacity of nerve cells using a frequency code. *Biophysical journal*, 7(6):797.

[310] Still, S. (2009). Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85(2):28005.

[311] Still, S. (2014). Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989.

[312] Still, S. and Crutchfield, J. P. (2007). Structure or noise? *arXiv preprint arXiv:0708.0654*.

[313] Still, S., Crutchfield, J. P., and Ellison, C. J. (2010). Optimal causal inference: Estimating stored information and approximating causal architecture. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(3):037111.

[314] Still, S. and Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148.

[315] Strong, Steven P., R. K. R. R. D. R. V. S. and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical review letters*, 80(1).

[316] Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M., and Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature neuroscience*, 11(9):1004.

[317] Summerfield, C., Wyart, V., Mareike Johnen, V., and De Gardelle, V. (2011). Human scalp electroencephalography reveals that repetition suppression varies with expectation. *Frontiers in Human Neuroscience*, 5:67.

[318] Swadlow., H. A. (1985). Physiological properties of individual cerebral axons studied in vivo for as long as one year. *Journal of neurophysiology*, 54(5).

[319] Swadlow., H. A. (1994). Efferent neurons and suspected interneurons in motor cortex of the awake rabbit: axonal properties, sensory receptive fields, and subthreshold synaptic inputs. *Journal of neurophysiology*, 7(2).

[320] Swensen, A. M. and Marder, E. (2000). Multiple peptides converge to activate the same voltage-dependent current in a central pattern-generating circuit. *Journal of Neuroscience*, 20(18):6752–6759.

[321] Takahata, M., Nagayama, T., and Hisada, M. (1981). Physiological and morphological characterization of anaxonic non-spiking interneurons in the crayfish motor control system. *Brain research*, 226(1-2):309–314.

[322] Tateno, T. and Jimbo., Y. (1999). Activity-dependent enhancement in the reliability of correlated spike timings in cultured cortical neurons. *Biological Cybernetics*, 80(1).

[323] Theunissen, F. E. and Miller, J. P. (1991). Representation of sensory information in the cricket cercal sensory system. ii. information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *Journal of neurophysiology*, 66(5):1690–1703.

[324] Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

[325] Timme, N., Alford, W., Flecker, B., and Beggs, J. M. (2014). Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *Journal of computational neuroscience*, 36(2):119–140.

[326] Timme, N. M., Ito, S., Myroshnychenko, M., Nigam, S., Shimono, M., Yeh, F.-C., Hottowy, P., Litke, A. M., and Beggs, J. M. (2016). High-degree neurons feed cortical computations. *PLoS computational biology*, 12(5).

[327] Timme, N. M. and Lapish, C. (2018). A tutorial for information theory in neuroscience. *ENeuro*, 5(3).

[328] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

[329] Todorovic, A., van Ede, F., Maris, E., and de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an meg study. *Journal of Neuroscience*, 31(25):9118–9123.

[330] Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037.

[331] van Vreeswijk, C. and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724.

[332] Varela, F. J., Thompson, E., and Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT press.

[333] Vasu, M. C. and Izquierdo, E. J. (2017a). Evolution and analysis of embodied spiking neural networks reveals task-specific clusters of effective networks. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 75–82.

[334] Vasu, M. C. and Izquierdo, E. J. (2017b). Information bottleneck in control tasks with recurrent spiking neural networks. In *International Conference on Artificial Neural Networks*, pages 236–244. Springer.

[335] Vasu, M. C. and Izquierdo, E. J. (2017c). Information bottleneck in control tasks with recurrent spiking neural networks. In *Int. Conf. on Artificial Neural Networks*, pages 236–244.

[336] Vazquez, R. (2010). Izhikevich neuron model and its application in pattern recognition. *Australian Journal of Intelligent Information Processing Systems*, 11(1):35–40.

[337] Vicente, R., Wibral, M., Lindner, M., and Pipa., G. (2011). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67.

[338] Villegas, R., Castillo, C., and Villegas, G. M. (2000). The origin of the neuron: The first neuron in the phylogenetic tree of life. In *Astrobiology*, pages 195–211. Springer.

[339] Viol, K., Aas, B., Kastinger, A., Kronbichler, M., Schöller, H. J., Reiter, E.-M., Said-Yürekli, S., Kronbichler, L., Kravanja-Spannberger, B., Stöger-Schmidinger, B., et al. (2019). Erroneously disgusted: fmri study supports disgust-related neural reuse in obsessive-compulsive disorder (ocd). *Frontiers in behavioral neuroscience*, 13:81.

[340] Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

[341] Wang, J., Korayem, M., and Crandall, D. (2013). Observing the natural world with flickr. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 452–459.

[342] Wang, X.-J. and Krystal, J. H. (2014). Computational psychiatry. *Neuron*, 84(3):638–654.

[343] Webb, B. (2001). Can robots make good models of biological behaviour? *Behavioral and brain sciences*, 24(6):1033–1050.

[344] Webb, B. (2002). Robots in invertebrate neuroscience. *Nature*, 417(6886):359–363.

[345] Weimann, J. M., Meyrand, P., and Marder, E. (1991). Neurons that form multiple pattern generators: identification and multiple activity patterns of gastric/pyloric neurons in the crab stomatogastric system. *Journal of Neurophysiology*, 65(1):111–122.

[346] Wessberg, J., Stambaugh, C. R., Kralik, J. D., Beck, P. D., Laubach, M., Chapin, J. K., Kim, J., Biggs, S. J., Srinivasan, M. A., and Nicolelis, M. A. (2000). Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408(6810):361–365.

[347] Wibral, M., Lizier, J. T., and Priesemann, V. (2015). Bits from brains for biologically inspired computing. *Frontiers in Robotics and AI*, 2:5.

[348] Wibral, M., Priesemann, V., Kay, J. W., Lizier, J. T., and Phillips, W. A. (2017). Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain and cognition*, 112:25–38.

[349] Wibral, M., Rahm, B., Rieder, M., Lindner, M., Vicente, R., and Kaiser., J. (2011). Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks. *Progress in biophysics and molecular biology*, 105(1):80–97.

[350] Wibral, M., Vicente, R., and Lindner, M. (2014a). Transfer entropy in neuroscience. In *Directed information measures in neuroscience*, pages 3–36. Springer.

[351] Wibral, M., Vicente, R., and Lindner, M. (2014b). Transfer entropy in neuroscience. In *Directed information measures in neuroscience*, pages 3–36. Springer.

[352] Williams, P. and Beer, R. D. (2013). Environmental feedback drives multiple behaviors from the same neural circuit. In *Advances in Artificial Life*, pages 268–275.

[353] Williams, P. L. (2011). *Information dynamics: Its theory and application to embodied cognitive systems*. PhD thesis, PhD thesis, Indiana University.

[354] Williams, P. L. and Beer, R. D. (2010a). Information dynamics of evolved agents. In *International Conference on Simulation of Adaptive Behavior*, pages 38–49. Springer.

[355] Williams, P. L. and Beer, R. D. (2010b). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

[356] Williams, P. L. and Beer, R. D. (2011). Generalized measures of information transfer. *arXiv preprint arXiv:1102.1507*.

[357] Williams, P. L., Beer, R. D., and Gasser, M. (2008). An embodied dynamical approach to relational categorization. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

[358] Wills, S. (1999). Relational learning in pigeons? *The Quarterly Journal of Experimental Psychology Section B*, 52(1b):31–52.

[359] Wollstadt, P., Lizier, J. T., Vicente, R., Finn, C., Martınez-Zarzuela, M., Mediano, P., Novelli, L., and Wibral, M. (2019). Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software*, 4(34):1081.

[360] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

[361] Wonders, C. P. and Anderson, S. A. (2006). The origin and specification of cortical interneurons. *Nature Reviews Neuroscience*, 7(9):687–696.

[362] Xu, T., Huo, J., Shao, S., Po, M., Kawano, T., Lu, Y., Wu, M., Zhen, M., and Wen, Q. (2018). Descending pathway facilitates undulatory wave propagation in caenorhabditis elegans through gap junctions. *Proceedings of the National Academy of Sciences*, 115(19):E4493–E4502.

[363] Young, S. R., Rose, D. C., Karnowski, T. P., Lim, S.-H., and Patton, R. M. (2015). Optimizing deep learning hyper-parameters through an evolutionary algorithm. In *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments*, pages 1–5.

[364] Zago, L., Pesenti, M., Mellet, E., Crivello, F., Mazoyer, B., and Tzourio-Mazoyer, N. (2001). Neural correlates of simple and complex mental calculation. *Neuroimage*, 13(2):314–327.

CURRICULUM VITAE

Madhavun Candadai

EDUCATION

Indiana University, Bloomington, IN, U.S.A. 08/2015 to 05/2020

Doctor of Philosophy in Cognitive Science

University of Cincinnati, Cincinnati, OH, U.S.A. 08/2012 to 05/2015

Master of Science in Electrical Engineering

Amrita Vishwa Vidyapeetham, TN, India 06/2007 to 04/2011

Bachelor of Technology in Electronics and Communication Engineering

CAREER

Intel A.I. Labs

Research Intern 05/2018 to 08/2018

Cincinnati Children's Hospital Medical Center

Student Researcher 08/2014 to 03/2015

IBM

Associate System Engineer 07/2011 to 07/2012

TEACHING

Instructor

Q260: Introduction to Python for Cognitive Scientists SP 2018, FA 2019

Q320: Computational Modeling for Cognitive Scientists SP 2018

Associate Instructor

Q350: Math and Logic for Cognitive Scientists FA 2016, FA 2017

Q355: Neural Networks and the Brain SP 2019

Q320: Computational Modeling for Cognitive Scientists SP 2020

PUBLICATIONS

1. **Candadai, M.**, & Izquierdo, E. J. (2019). Sources of predictive information in dynamical neural networks. *bioRxiv*. (Submitted)

2. **Candadai, M.**, & Izquierdo, E. J. (2019). infotheory: A C++/Python package for multivariate information theoretic analysis. *Journal of Open Source Software*, 5(47), 1609.

3. **Candadai, M.**, Setzler, M., Izquierdo, E. J., & Froese, T. (2019). Embodied dyadic interaction increases complexity of neural dynamics: A minimal agent-based simulation model. *Frontiers in psychology*, 10, 540.

4. Dwiel, Z., **Candadai, M.**, Phielipp, M. J., & Bansal, A. K. (2019). Hierarchical Policy Learning is Sensitive to Goal Space Design. *Task-Agnostic Reinforcement Learning Workshop, ICLR*.

5. Dwiel, Z., **Candadai, M.**, & Phielipp, M. (2019). On Training Flexible Robots using Deep Reinforcement Learning. *Intelligent Robots and Systems (IROS) conference*.

6. **Candadai, M.**, & Izquierdo, E. (2018). Multifunctionality in embodied agents: Three levels of neural reuse. 40th Cognitive Science Conference, Madison, Wisconsin.

7. **Vasu, M. C.**, & Izquierdo, E. J. (2017). Information bottleneck in control tasks with recurrent spiking neural networks. *In International Conference on Artificial Neural Networks* (pp. 236-244). Springer, Cham.

8. **Vasu, M. C.**, & Izquierdo, E. J. (2017). Evolution and analysis of embodied spiking neural networks reveals task-specific clusters of effective networks. In *Proceedings of the Genetic and Evolutionary Computation Conference* (pp. 75-82).

Nominated for Best Student Paper, 2017 by International Society of Artificial Life – Student chapter.

9. **Candadai, M.**, Vanarase, A., Mei, M., & Minai, A. A. (2015). ANSWER: An unsupervised attractor network method for detecting salient words in text corpora. In 2015 *International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

10. **Candadai Vasu**, M. (2015). ANSWER: A Cognitively-Inspired System for the Unsupervised Detection of Semantically Salient Words in Texts Master's dissertation, University of Cincinnati.

ABSTRACTS

1. **Candadai, M.**, & Izquierdo, E. J. (2019) Information dynamics in embodied multifunctional recurrent neural networks. Society for Neuroscience Annual Meeting (SfN)

2. **Candadai, M.**, & Izquierdo, E. J. (2019) On the Role of Predictive Coding in Adaptive Behavior. Greater Indiana, Society for Neuroscience Meeting.

3. **Vasu, M. C.**, & Izquierdo, E. J. (2018) Multifunctionality Can Emerge from Brain-Body-Environment Interaction: An Information Theoretic and Dynamical Systems Theoretic Account. Greater Indiana, Society for Neuroscience Meeting.

TALKS

1. Disentangling sources of predictive coding in embodied agents (2019, May). Midwestern Cognitive Science Conference, Cognitive Science Society.

2. Information theoretic exploration of the neural basis of behavior (2018, April). Intelligent and Interactive Systems Seminar, School of Informatics, Computing and Engineering, Indiana University, Bloomington.

AWARDS

NSF Research Traineeship affiliate, Complex Networks and Systems	2019
Supplemental Research Fellowship, Indiana University	2017 & 2018
Outstanding Graduate Teaching Award	2017 – 2018
ACM Graduate student travel grant to present at GECCO'17	2017
Graduate Fellowship, Indiana University	2015 – 2016
University Graduate Scholarship, University of Cincinnati	2012 – 2014
RevolutionUC Hackathon – 2nd place	2014