Cognitive Information Transformation in Functional Brain Networks

by

Takuya Ito

A Dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Behavioral and Neural Sciences

written under the direction of

Dr. Michael W. Cole

and approved by

Newark, New Jersey

January 2021

ABSTRACT OF THE DISSERTATION

Cognitive Information Transformation in Functional Brain Networks

By Takuya Ito

Dissertation Director:
Dr. Michael W. Cole

The human brain is a flexible information processing system. Across a range of simple and complex tasks, such as walking across the street to playing basketball, the brain transforms sensory information from the environment into corresponding motor actions. This sensory input to motor output transformation likely requires a sequence of complex neural computations implemented by brain networks. Though decades of cognitive neuroscience have made great progress in characterizing the functions of individual brain areas, less progress has been made in understanding exactly how these brain regions work in concert to implement the diverse cognitive computations underlying complex behaviors. In this thesis, I provide an account of how the brain's distributed functional networks implement neurocognitive functions and computations. First, I demonstrate how local cognitive task activations can be computed from the activity of other brain areas through distributed brain network connectivity patterns. This illustrates how intrinsic functional connectivity enables the transfer of task-relevant activations between brain regions. Second, I demonstrate how local cognitive information, such as sensory stimulus activations in sensory cortices, is transformed into motor activations in motor cortex through a sequence of computations governed by intrinsic functional connectivity during cognitive tasks in both humans and non-human primates. This demonstrates that the intrinsic brain network organization can provide insight into how the brain implements neurocognitive computations and transformations. Finally, I investigate the relationship between task activations and functional network connectivity from a dynamical systems perspective. Specifically, I demonstrate that task-state activity quenches ongoing functional correlations and variability, and that this quenching occurs due to a sigmoidal transfer function that describes local mean-field neural activations. This suggests that task-state functional network changes are meaningful, and reflect nonlinear relationships between brain regions. This provides a way forward to improve current models of neural computations and communication by leveraging nonlinear models of neural dynamics. Together, the results presented in this thesis provide a novel understanding of how functional brain network organization shapes cognitive computations.

Acknowledgements

As a 2nd year undergraduate at WashU, I began searching through neuroscience and psychology laboratories for potential research opportunities. Disenchanted with the endless thought experiments proposed in many of my philosophy classes, I wound up e-mailing Todd Braver in the Psychology department to see if I could get experience with "real" data. Todd put me in touch with one of his promising and talented post-docs, and through a bit of chance and luck, Mike took me on as an undergraduate research assistant. Soon after he showed me some images of FreeSurfer brains (which in hindsight may have been a cheap trick!), I switched my secondary major from philosophy to cognitive neuroscience. I was 19 years old, and did not even know what a "PhD" was.

I am grateful to Mike for his patience and mentorship through my undergraduate and graduate career. Looking back at that first year, I cringe with embarrassment. It pains me to think back on how long it took me to perform simple statistical tests in AFNI (a couple weeks). Fortunately for both of us, I can now perform statistical tests much more quickly.

I also recall the early days of moving to Rutgers with Mike as a research assistant and helping to set up the lab. It was certainly a unique experience, and I was able to learn a lot with Mike since there was so much freedom. I am extremely grateful to Bart during that time, as he graciously hosted me in his lab space while Mike's was under construction. Without my experience in Bart's lab and the lively interactions with Bart, Kohitij, Mar, Jon, Jeroen, and Jasmine, I doubt I would have considered pursuing a PhD.

I am thankful to the many lab friends that have made life at Rutgers so enjoyable: Doug, Levi, Ravi, Richard, Marjolein, KK, Carrisa, Julia, Emily, Nicole, Brian, Katie, Luke, Ruben, Kirsten, and Ethan. I am especially thankful to the first two post-docs in the lab, Doug and Ravi, who early on convinced me to pursue and stay with the PhD, and who made attending conferences fun, weird, and memorable. I am grateful to my BNS classmates and the greater CMBN community. I am amazed at the perseverance of my classmates, especially those that perform the heroic animal surgeries and experiments that can take months or years. They are cut from a different cloth.

I am especially thankful to my thesis committee members and the many mentors I've had, too. Of notable importance is Horacio, who has encouraged and supported me both as a mentor and friend throughout the academic and social challenges that come with the PhD process. He has also taught me everything I know about dynamical systems and numerical methods. When I first asked if he would teach me how to build and model dynamical systems from scratch, he sat with me and watched over my shoulder as he commented on my coding. I remember sweating and thinking, "This is not what I signed up for." But, thanks to those intense sessions, I will now never forget how to manually integrate differential equations or estimate a fixed point. I am also thankful to Drew for his collegiality and for teaching me to think critically, Mauricio for his friendliness and whose work I drew inspiration from, and of course Olaf, who has been my academic icon, from when I first learned what a "connectome" was as an undergraduate. I was so nervous when I first e-mailed Olaf asking him if he would serve on my committee. I must have read over that e-mail at least 10 times.

Outside of CMBN, there are many people without whom the PhD process may have been impossible. Of special note are my roommates through the years who have kept me sane: Zaid, Travis, Tyler, Jeremy, Neo, and Dyllon. Without them and many others, I would no longer have any social skills.

And of course to Michele, who has kept me grounded and motivated with her rational yet kind wit. Thank you for telling me when my ramblings are nonsense and also for proofreading this, too.

Finally, I have a lifetime of gratitude for my parents (Pat and Kenji) and sister (Maya). Though their passions in video production (mom), interior design (dad), architecture (sister), and "honk" (all) may seem distant to neuroscience, I like to think their unorthodox creativity serves as my strongest inspiration.

To my parents, Pat and Kanobies.

Table of Contents

List of Abbreviations

R^2	Coefficient of determination
ANN	Artificial neural network
AUD	Auditory network
BGC	Between-network global connectivity
BOLD	Blood-oxygen-level-dependent
C-PRO	Concrete-Permuted Rule Operations
CCN	Cognitive control network
CON	Cingulo-opercular network
DAN	Dorsal attention network
DMN	Default mode network
E-I	Excitatory-inhibitory
ENN	Empirically-estimated neural network
FDR	False discovery rate
FEF	Frontal eye fields
FIR	Finite impulse response
fMRI	Functional magnetic resonance imaging
FPN	Frontoparietal network
FWE	Family-wise error
GBC	Global brain connectivity
GLM	General linear model
GSR	Global signal regression
HCP	Human Connectome Project
Hz	Hertz

ICA	Independent component analysis
IT	Inferior temporal cortex
ITI	Inter-trial interval
LAN	Language network
LIP	Lateral intraparietal cortex
LPFC	Lateral prefrontal cortex
MT	Medial temporal area
MVPA	Multivariate pattern analysis
NHP	Non-human primate
OFC	Orbitofrontal cortex
ORA	Orbito-affective network
PCA	Principal component analysis
PFC	Prefrontal cortex
PMM	Posterior multimodal network
ReLU	Rectified linear unit
RSA	Representational similarity analysis
RSM	Representational similarity matrix
SMN	Somatomotor network
SVM	Support vector machine
V1	Visual area 1
V2	Visual area 2
V4	Visual area 4
VIS1	Primary visual network
VIS2	Secondary visual network
VMM	Ventral multimodal network

Chapter 0
Preface

This thesis is the culmination of a huge team effort. The work presented here would not have been possible without the essential contributions of my co-authors and collaborators.

Chapter 1 was inspired by and includes excerpts from a collaborative opinion piece with Luke J. Hearne, Ravi D. Mill, Carrisa V. Cocuzza, and Michael W. Cole [Ito et al., 2020b].

Chapter 2 was previously published with co-authors Kaustubh R. Kulkarni, Douglas H. Schultz, Ravi D. Mill, Levi Solomyak, and Michael W. Cole [Ito et al., 2017].

Chapter 3 is a study that is currently under review. This was joint work with co-authors Guangyu Robert Yang, Douglas H. Schultz, Patryk Laurent, and Michal W. Cole.

Chapter 4 is work that is currently in progress. Currently, this project includes collaborators Scott L. Brincat, Markus Siegel, Earl K. Miller, and Michael W. Cole.

Chapter 5 was previously published with co-authors Scott L. Brincat, Markus Siegel, Biyu J. He, Ravi D. Mill, Earl K. Miller, Horacio G. Rotstein, and Michael W. Cole [Ito et al., 2020a].

Chapter 1
Introduction

Portions of this section include excerpts (sometimes paraphrased) from [Ito et al., 2020b].

Imagine it's close to midnight, and you're strolling down 14th street along Union Square in New York City. Shops and restaurants are beginning to close down. Pedestrians are trickling out of the streets, descending into the subway or up into their apartments. Turning the corner, you hear rustling noises around the trash bin. Your face squeezes with distaste, and without seeing anything, you know: it's a rat. Sure enough, moments later, the rat darts back into sewer, dinner in mouth.

Several months later you're walking back to your hotel room in Singapore after a long day of meetings. Again, you hear a familiar rustle near the trash bin. But something is different – you're in Singapore. Instead of disgust, you pause, wondering about the possibilities of the source of that sound. It couldn't be a rat. Unlike New York City, the hygiene of Singaporean streets is impeccable. Baffled, you inspect the trash bin. You realize the sound was nothing, just tumbling leaves swirling around the sides of an empty trash bin.

Despite the similarity of auditory stimuli in these two situations, you actively respond differently. What matters here, rather than what you hear, is context. The contextual knowledge about the cleanliness of Singapore and New York alter how you respond to similar sensory information. How do context and sensory information interact to inform our actions? We integrate complex sensory and contextual information in daily actions, from crossing the street to driving a car. Nonetheless, though we may take for granted the ability to convert these complex environmental signals into appropriate actions, exactly how our brains achieve this information processing capability is not fully understood.

Scientists and philosophers have long speculated that the brain implements input-output relationships like a computer [Turing, 1948, Von Neumann and Kurzweil, 2012]. Despite this widespread acknowledgement, the exact neural implementation of the diverse cognitive computations we are capable of are not fully known. Here I refer to cognitive computations as the computational (or formal) description of a cognitive process, such as performing math, reading words, or perceiving colors. While there may be infinite formal descriptions of a cognitive computation, a central aim of cognitive neuroscience is to understand the principles by which the brain implements such computations.

In recent decades, advances in functional neuroimaging have enabled cognitive brain mapping. In particular, imaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography have facilitated progress in understanding how mental functions map onto brain areas. However, the modern application of functional neuroimaging to understanding cognitive processes has focused on brain mapping localization – the practice of identifying what functions each brain structure performs. This approach primarily maps function to structure by establishing relationships between external stimuli or tasks onto the activity of neurons and cortical areas [Henson, 2005]. Some examples of this general strategy include analyzing event-related spike rate changes in single- or multi-unit recordings [Wallis, 2018], general linear modeling with fMRI [Poldrack et al., 2011], and event-related potentials with electroencephalography [Luck, 2014]. This strategy has undoubtedly been tremendously useful for characterizing the functions of spatially localized neurons and cortical areas

[Genon et al., 2018]. However, while the approach of mapping function to structure in a piecemeal manner may provide a comprehensive cognitive cartography of the brain, this approach would fall short of explaining how the brain actually implements cognitive computations through the interaction among its components.

1.1 Cognition and its relation to brain network organization

So how might one study how the brain implements cognitive computations using modern functional neuroimaging techniques? Theoretical work suggests that the brain may implement cognitive computations from the interaction among the brain's components [Craver, 2007, Mill et al., 2017, Medaglia et al., 2015]. However, this requires a mechanistic understanding of the relationship between the system's components [Reid et al., 2019, Bassett and Sporns, 2017]. Fortunately, efforts to understand the brain mechanistically are becoming more prominent. Recent advances in network neuroscience are providing ever more detailed descriptions of brain network architecture, facilitating a functional understanding of the brain by revealing the relations among its neural components [Bassett and Sporns, 2017, Sporns et al., 2005]. Coined the "connectome", recent advances in data acquisition and analytic techniques have improved our understanding of the brain's large-scale network organization [Sporns et al., 2005]. This includes a large-scale structural description of physical connections between brain regions [Hagmann et al., 2008, van den Heuvel and Sporns, 2011] and large-scale functional network organization estimated through mapping statistical dependencies of the living brain [Power et al., 2011, Yeo et al., 2011, Biswal et al., 1995]. However, while descriptions of brain network connectivity provide the foundations from which function likely emerges, brain connectivity mapping alone would provide limited insights into the relations among neural entities. This is because brain network connectivity describes relations among neural entities without reference to the cognitive processes targeted by experimental task manipulations.

Nevertheless, the discovery of resting-state functional connectivity (FC; i.e., correlated brain activity between sets of brain areas) using modern neuroimaging techniques [Biswal et al., 1995] has led to a wealth of important findings. For example, by carefully mapping out the intrinsic FC structure of spontaneous brain activity from hundreds of cortical areas, scientists have discovered that the brain is organized into modular functional networks [Fox et al., 2005, Yeo et al., 2011, Power et al., 2011]. These functional networks were also found to be related to the large-scale structural network organization of the brain, as estimated with diffusion weighted imaging [Honey et al., 2009, Deco et al., 2013a]. Moreover, properties of brain networks have been shown to be associated with cognitive ability and behavior [Cole et al., 2012, Schultz and Cole, 2016, Kong et al., 2019, Finn et al., 2015]. While these findings provide important evidence that brain network organization is important for cognitive processes, they offer limited mechanistic insight into how the brain implements cognitive computations.

However, several recent studies have revealed a closer mechanistic link between network organization and cognitive functions [Saygin et al., 2012, Saygin et al., 2016, Tavor et al., 2016, Cole et al., 2016a]. These studies illustrate the power of intrinsic network organization in predicting the task-evoked activations thought to reflect cognitive processes [Wallis, 2018]. For example, Saygin and colleagues demonstrated that anatomical connectivity precedes the functional specialization in both the fusiform gyrus (for face selectivity) and visual word form area [Saygin et al., 2012, Saygin et al., 2016]. Building on those findings, Tavor and colleagues revealed that resting-state fMRI can predict individualized cognitive task activations. And finally, we recently illustrated that "activity flow" between brain regions can predict regional task activation patterns [Cole et al., 2016a]. Together, these findings suggest that distributed network interactions can explain the emergence of the task activations associated with cognitive processes.

1.2 Local functions and distributed computations

Neuroscientific studies have revealed both local functional specialization and distributed functional organization. From the localist perspective, Horace Barlow championed the study of brain function at the level of single neurons, suggesting that distributed function and "mass action" were misguided views of the brain [Barlow, 1992]. In contrast, others such as Karl Lashley have advocated that the "notion of decentralization or of cerebral function without absolute anatomical localization need not involve an abandonment of recognized physiological principles or a denial of known facts of localization" [Lashley, 1931].

How might these two views of brain function be reconciled? Careful observations of connectivity and local function have led to a coherent hypothesis that allows for both hypotheses to be true: That the functional localization (of a neuron or cortical area) depends on its intrinsic connectivity [Passingham et al., 2002, Jbabdi et al., 2013]. In neuroscience, grid and place cell selectivity have been proposed to depend on their intrinsic connectivity from entorhinal cortex [McNaughton et al., 2006, Hafting et al., 2005]. In the visual system of a developing human infant, it has been shown that structural connectivity precedes the functional localization of the visual word form area [Saygin et al., 2012]. More generally, a recent review suggested that identifying a neural unit's connectivity fingerprint likely determined its function and therefore its representational capacity [Mars et al., 2018], providing a basis for the concept of connectivity-based receptive fields. These findings and hypotheses suggest that functional localization and specificity may naturally emerge through distributed connectivity.

1.3 Connectionist architectures and cognitive computations

In the last two decades, modern cognitive neuroimaging has focused primarily on either mapping brain network organization or mapping cognitive function to brain areas. How can these two approaches be merged? Originating in the 1980s, connectionist (or neural network) theory has produced successful models that can perform complex cognitive tasks [Rumelhart et al., 1986, Cohen et al., 1990]. Loosely designed from the brain [McCulloch and Pitts, 1943, Rosenblatt, 1958], connectionist models are a class of computational models defined by a network of interconnected units that are optimized for a specific task [Rumelhart et al., 1986]. This includes recently developed deep neural networks that improve model performance by including additional neural units with structured connectivity as 'hidden' layers between input and output [Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014, Wen et al., 2018]. Interestingly, when these networks included biological constraints (e.g., number of layers and number of units per layer to match the number of regions within the ventral visual system), networks can exhibit similar neuronal responses and selectivity (e.g., the emergence of face-selective units) to empirical neural data [Wen et al., 2018, Yamins et al., 2014]. Thus, like the brain, these connectionist networks also naturally exhibit functional specialization across different layers and units, providing additional evidence that specialization can emerge through distributed connectivity.

Other connectionist modeling studies have illustrated that very simple (or so-called "vanilla") recurrent neural network models can perform many different complex cognitive tasks while exhibiting brain-like network dynamics [Song et al., 2016, Mante et al., 2013]. Recurrent neural networks are connectionist models where connections recurrently feedback on units, rather than being strictly feedforward [Yang and Wang, 2020]. One such study illustrated that a single recurrent network could learn up to 20 different tasks, suggesting that despite having a set of fixed connections, a single set of connections is capable of implementing a diversity of cognitive functions [Yang et al., 2019]. In particular, Yang and colleagues found that a mixture of simple- and mixed-selectivity naturally emerges among neural units by learning many different tasks. (Simple and mixed-selectivity refer to the ability of a neural unit to be useful (or responsive) during few or many different task components, respectively [Fusi et al., 2016, Rigotti et al., 2013].) Importantly, the selectivity profiles of these networks were entirely determined by task training and its connectivity organization. Together, these findings suggest the importance of connectivity architectures in guiding cognitive computations.

1.4 Searching for network computations in the brain

Both connectionist modeling and empirical connectivity-behavior studies support the long-standing hypothesis that the brain's network organization constrains its functionality. This leads to a more targeted question: how much function (in terms of both functional selectivity and task-evoked neural activity) can connectivity patterns explain? One way to evaluate the relative contribution of connectivity in determining the functionality of neural populations is to test the plausibility of each with empirical data. We recently proposed the notion of "activity flow modeling" with empirical data that tests the ability of estimated intrinsic connectivity patterns in predicting task-evoked brain activations

[Cole et al., 2016a]. The approach of activity flow was formulated in an effort to model the propagation of neural activity (e.g., spike rates or fMRI signal amplitudes in biological contexts) between neural entities using empirically estimated connectivity. Though current approaches are limited in their causal and mechanistic inferences due to data and methodological limitations [Reid et al., 2019], questions that address the relationship between functional network organization and task-evoked activations can still be addressed.

Quantitatively, the activity flow approach involves simulating a single neural network computation (e.g., a forward iteration in a feedforward neural network): $a_h = f(\sum_{i \in I} a_i w_{hi})$, where a unit a_h's activity is a linear combination of all other units' activity ($\sum_{i \in I} a_i w_{hi}$) weighted by their connectivity (w_{hi}) to a_h before passing through a transfer function f, such as a sigmoid. Within the connectionism framework, this formalization is referred to as a combination of the propagation and activation rules [Rumelhart et al., 1986]. Other names given to more specific instantiations of this algorithm (e.g., with a particular form of nonlinearity) include the McCulloch-Pitts neuron [McCulloch and Pitts, 1943], the perceptron [Rosenblatt, 1958], the divisive normalization model of neural activity [Carandini and Heeger, 2012], adaptive linear element [Widrow and Lehr, 1990], and spreading activation [Collins and Loftus, 1975]. Critically, [Cole et al., 2016a] adapted this algorithm for use with empirical data (e.g., fMRI activity and functional connectivity estimates) to parameterize empirically derived models that make quantitative predictions of the spread of estimated activity over brain network connections [Cole et al., 2016a].

Previously, in neural network simulations we found that activity flow mapping was only effective in network architectures with strong inter-area coupling (relative to recurrent/local) coupling [Cole et al., 2016a]. This is consistent with previous findings, where large effects of inter-regional synaptic coupling (relative to local coupling) were important for predicting FC from structural connectivity [Deco et al., 2013a]. Thus, we concluded that distributed connectivity plays a substantial role in determining local activity. This is highly compatible with the notion that each localized population has a 'connectivity fingerprint' that largely determines its functionality [Passingham et al., 2002, Mars et al., 2018]. These results are also in line with the observation that large-scale propagation of neural activity in animal models tends to conform to large-scale anatomical connectivity patterns [Kotter and Sommer, 2000, Honey et al., 2009].

These results demonstrate the feasibility of identifying network computations in the brain by identifying estimated activity flow patterns in modern neuroimaging data. By leveraging empirical estimates of intrinsic brain network connectivity, we can begin to probe the contribution of these connections in producing the task-related activity that reflect cognitive processes. The primary objective of this thesis is to build on these previously established ideas to investigate how network organization contributes to cognitive computation across a range of tasks using modern functional imaging and electrophysiology techniques.

1.5 Overview of chapters

In this thesis, I combine functional network mapping with cognitive task manipulations to identify how cognitive information is transferred and transformed within the brain [Ito et al., 2017, Ito et al., 2020b, Ito et al., 2020a]. Cognitive brain mapping studies focus on investigating where in the brain information is located. In contrast, network neuroscience studies describe the functional and physical organization of brain networks. Inspired by connectionist theory where connections determine the computations performed by neural network models during tasks, I combine cognitive brain mapping with functional network mapping to ask how brain networks compute cognitive information through their connectivity patterns. This would provide insight into how cognitive computations are supported (and constrained) by the network organization of the brain.

I present three scientific aims. Each aim is a self-contained research study (with the exception of Aim 2, which contains two independent yet related studies).

In **Aim 1** (Chapter 2), I investigate how cognitive information is transferred within the brain through intrinsic functional connectivity patterns [Ito et al., 2017]. (In general, and unless otherwise clearly stated, "cognitive information" or "cognitive representations" refer to task-evoked *activation* patterns that are decodable, such as with a linear classifier. In other words, task-relevant information can be extracted from neural or fMRI *activations.*) In this study, I show that decodable patterns of task-evoked activity in one brain region can be projected (or transferred) to another brain region (while preserving decodable information). This projection (or information transfer) is estimated from inter-region intrinsic FC estimates. This provides evidence for the hypothesis that activity flow over resting-state functional connections transfers information between brain regions.

Building on findings from Aim 1, in **Aim 2**, I address how information is transformed within the brain during context-dependent tasks. Aim 2 contains two self-contained studies (Chapter 3 and 4). In a context-dependent task, sensory stimulus information is transformed into a motor response according to task contexts. This typically involves a nonlinear mapping from stimulus to response. How do stimulus representations in sensory cortices get transformed into appropriate motor signals in motor cortex? Thus, rather than asking where cognitive information is in the brain (i.e., a traditional cognitive mapping approach), in Aim 2, I ask *how cognitive information is used in the brain.* We show that multi-step activity flow computations provide a potential mechanism for cognitive information transformation. This involves identifying and mapping the conjunction of both task context and sensory stimulus representations to generate the appropriate motor signals.

Aim 2 comprises of Chapters 3 and 4, which are two separate yet conceptually related studies. Chapter 3 uses a combination connectionist modeling and fMRI data analysis. We show that we can predict stimulus-response transformations in a 64-context cognitive control paradigm. Chapter 4 contains a shorter study (currently in progress), and uses multi-unit neuronal spiking data from six cortical sites during a context-dependent sensorimotor task [Siegel et al., 2015]. Thus, in Chapter 4, I demonstrate that the network estimation and activity flow techniques that were originally developed for fMRI data analysis extend to neural firing rate data. Critically, we show that we can predict the firing rate patterns in frontal eye fields (FEF) that correspond to behavioral outputs (saccades) using spiking activity from other cortical areas.

Aims 1 and 2 (Chapters 2-4) focus on revealing the plausibility of using empirically-estimated network connections to predict how cognitive information is computed between brain regions. The network connectivity estimates are obtained by measuring the statistical dependencies of spontaneous time series between brain regions (e.g., correlations and/or regression-based techniques). Previous work has shown that the statistical dependencies between brain regions are generally preserved across intrinsic and task-evoked states [Cole et al., 2014a, Krienen et al., 2014]. Despite this, there are small yet robust changes that occur across rest and task states [Cole et al., 2014a, Gratton et al., 2016].

In **Aim 3** (Chapter 5), I investigate the biophysical basis of these changes with the goal of illuminating the precise mechanisms that govern statistical dependencies between brain regions [Ito et al., 2020a]. This was achieved by linking fMRI FC results with non-human primate noise correlation analyses (using multi-unit activity). Furthermore, I propose a dynamical systems approach to understanding the biophysical basis of large-scale neural correlations. Though Chapter 5 does not directly address cognitive information transfer and transformation, the characterization of the mechanisms that underlie shared neural dynamics (e.g., FC) facilitate understanding of how neural systems communicate during intrinsic and task-evoked states. For example, I show evidence that neural populations activate in a sigmoid-like fashion, suggesting that future models can take into account more detailed approaches to modeling ‘activity flow’ by taking into account nonlinear relationships among brain regions (e.g., see [Cole et al., 2020]).

Collectively, this thesis provides evidence for how cognitive information is used and communicated within the brain using cognitive neuroimaging techniques.

Chapter 2

Cognitive task information is transferred between brain regions via resting-state network topology

This chapter has been published in Nature Communications [Ito et al., 2017]. The contents have been reformatted for this thesis.

2.1 Abstract

Resting-state network connectivity has been associated with a variety of cognitive abilities, yet it remains unclear how these connectivity properties might contribute to the neurocognitive computations underlying these abilities. We developed a new approach – information transfer mapping – to test the hypothesis that resting-state functional network topology describes the computational mappings between brain regions that carry cognitive task information. Here we report that the transfer of diverse, task-rule information in distributed brain regions can be predicted based on estimated activity flow through resting-state network connections. Further, we find that these task-rule information transfers are coordinated by global hub regions within cognitive control networks. Activity flow over resting-state connections thus provides a large-scale network mechanism for cognitive task information transfer and global information coordination in the human brain, demonstrating the cognitive relevance of resting-state network topology.

2.2 Introduction

The human brain is thought to be a distributed information-processing device, its routes of information transfer constituting a core feature that determines its computational architecture. Many studies have used correlations among resting-state functional MRI (fMRI) time series to study functional connectivity (FC) in the human brain [Raichle et al., 2001]. It remains unclear, however, if these resting-state FC routes are related to the brain's routes of cognitive information transfer. Evidence that group and individual differences in resting-state FC correlate with cognitive differences [Cole et al., 2011a, Shannon et al., 2011, Smith et al., 2015] suggests that there is a systematic relationship between resting-state FC and cognitive information processing. However, without linking FC to information transfer, it remains unclear whether or how resting-state FC might mechanistically contribute to neurocognitive computations. Additionally, while a number of studies have shown that task information representations are distributed throughout the brain [Haxby et al., 2006, Muhle-Karbe et al., 2016, Poldrack et al., 2009, Zhang et al., 2013], such studies have yet to reveal how these distributed representations are coordinated, and how information in any one brain region is used by other brain regions to produce cognitive computations [De-Wit et al., 2016]. Other studies investigating interdependence of brain regions during tasks (rather than during rest) have typically emphasized statistical dependencies between regional time series [Cole et al., 2013, Gratton et al., 2016, Sadaghiani et al., 2015], rather than the mechanistic transfer of task-relevant information content (reflected in task activation patterns [Norman et al., 2006]) between those regions. Thus, it remains unclear whether or how the network topology described by either resting-state or task-evoked FC is relevant to the neurocognitive computations underlying task performance.

Here, we provide evidence for a network mechanism underlying the transfer and coordination of distributed cognitive information during performance of a variety of complex multi-rule tasks. Based on recent evidence that resting-state FC describes the routes of task-evoked activity flow [Cole et al., 2016a] (Figure 2.1a) – the movement of task activations between brain regions – we hypothesized that resting-state network topology describes the mappings underlying task information (task-evoked activation pattern) transfer between brain regions. If true, this hypothesis implicates a network mechanism for an information-preserving mapping across brain regions involving communication channels [Shannon, 1948, De-Wit et al., 2016] described by resting-state network topology. Identifying such a mechanism would provide an important new window into the large-scale information processing architecture of the human brain.

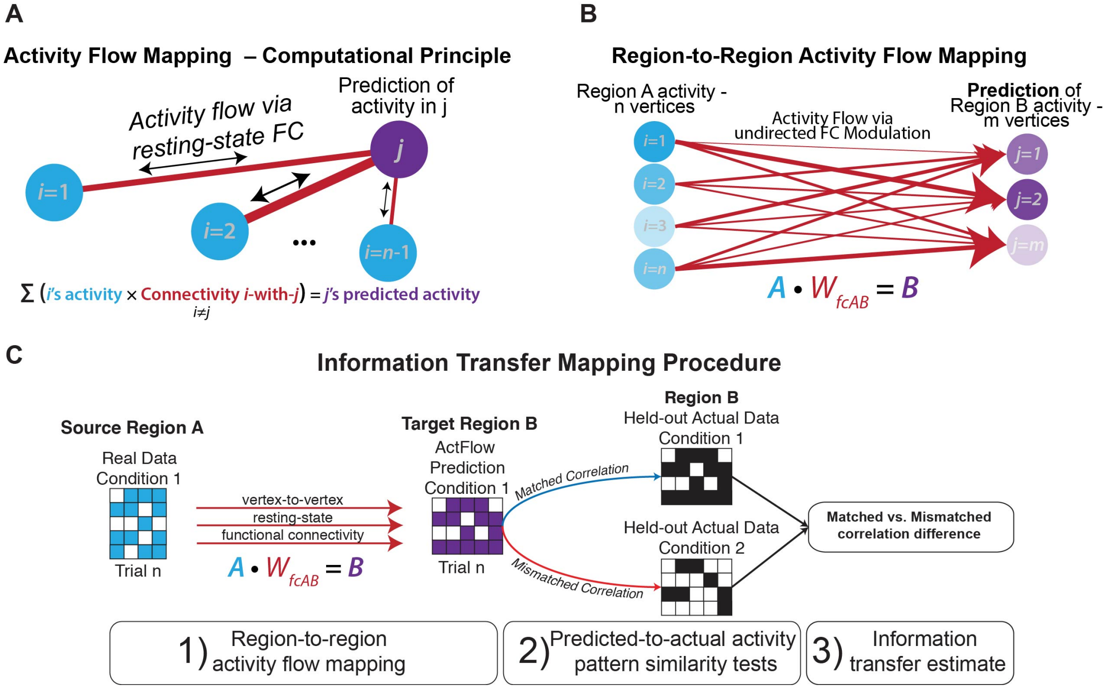

A
Activity Flow Mapping – Computational Principle
Activity flow via resting-state FC
Prediction of activity in j
j
i=1
i=2
i=n-1
Σ (i's activity × Connectivity i-with-j) = j's predicted activity
i≠j
B
Region-to-Region Activity Flow Mapping
Region A activity - n vertices
Activity Flow via undirected FC Modulation
Prediction of Region B activity - m vertices
i=1
i=2
i=3
i=n
j=1
j=2
j=m
A • W_fcAB = B
C
Information Transfer Mapping Procedure
Source Region A
Real Data Condition 1
Trial n
vertex-to-vertex resting-state functional connectivity
A • W_fcAB = B
Target Region B
ActFlow Prediction Condition 1
Trial n
Matched Correlation
Mismatched Correlation
Region B
Held-out Actual Data Condition 1
Held-out Actual Data Condition 2
Matched vs. Mismatched correlation difference
1) Region-to-region activity flow mapping
2) Predicted-to-actual activity pattern similarity tests
3) Information transfer estimate

Figure 2.1: Measuring information transfer through activity flow mapping and cognitive task information decoding. A) Computational principle of activity flow mapping [Cole et al., 2016a]. Activity in a held-out region is predicted by computing the linear weighted sum of all other regions' activity weighted by those regions' resting-state FC estimates with the held-out region. (The held-out region's activity is not included when computing the predicted activity of that region, thus avoiding a circular prediction.) B) Region-to-region activity flow mapping between vertices/voxels of isolated regions ("many-to-many" rather than "all-to-one" mapping of regions). Mathematically, we predict the activation pattern in Region B by computing the dot product of Region A's activation pattern vector with the vertex-to-vertex resting-state FC matrix between Region A and B. C) Information transfer mapping, which involves region-to-region activity flow mapping and representational similarity analysis (information decoding/classification) on held-out data. To test the transfer of task information from Region A to Region B, we compare the predicted activation pattern of Region B (mapped using Region A's activation pattern) to the actual task activation pattern of Region B for all task conditions using a spatial Spearman's rank correlation. For every prediction, spatial correlations to the task prototypes are computed and the information transfer estimate is measured by taking the difference of the correctly matched spatial correlation to the average of the incorrectly matched (mismatched) spatial correlations. Here we depict the approach for only two task conditions.

The current study focuses on fine-grained activation and FC topology, allowing us to infer the role of resting-state FC in carrying task-related information (represented by activation patterns [Norman et al., 2006, Poldrack et al., 2009, Zhang et al., 2013, Muhle-Karbe et al., 2016]). This is, in turn, critical for testing a novel network mechanism in which resting-state FC topologies of cognitive control networks globally coordinate task-related information. Further, correspondence between resting-state FC topology and information-representing activation patterns would demonstrate the general mechanistic relevance of resting-state FC for information processing in the human brain.

Recent evidence suggests that resting-state FC reflects the human brain's invariant global routing architecture [van den Heuvel et al., 2009, Marrelec et al., 2016]. Supporting this, it has been demonstrated that most of the functional network topology variance present during task performance (80%) is already present during rest [Cole et al., 2014a, Krienen et al., 2014]. Thus, resting-state FC primarily reflects an intrinsic functional network architecture that is present regardless of cognitive context, given that there are only moderate changes to functional network organization across tasks [Cole et al., 2014a, Krienen et al., 2014]. We built upon these findings to test the hypothesis that intrinsic network topology describes the baseline network state upon which distributed cognitive information processing occurs.

Our hypothesis required an approach to empirically derive the mapping between information representations of pairs of brain regions, similar to identifying the transformation weights between layers in a neural network model [Yamins et al., 2014]. The approach developed here contrasts with two previous approaches that describe the coordination of task-relevant information between brain regions. One of the previous approaches measures small shifts in task-evoked FC according to task-relevant content [Cole et al., 2013, Sadaghiani et al., 2015]. Another previous approach measures the correlation of moment-to-moment fluctuations in information content between regions [Coutanche and Thompson-Schill, 2013]. Critically, these prior approaches primarily describe time-dependent statistical dependencies rather than suggest a large-scale mechanism by which task representations are mapped between brain regions. Thus, neither of these earlier approaches were appropriate for characterizing a network mechanism by which cognitive information is mapped between regions. Nonetheless, these past approaches were important for demonstrating the basic phenomenon of large-scale task information coordination, which we sought to better understand via the recently developed activity flow mapping approach [Cole et al., 2016a].

The hypothesis that fine-grained resting-state FC describes the representational mappings between brain regions during tasks is compatible with several recent findings. First, resting-state FC topology was recently shown to be highly structured and reproducible, forming clusters of networks consistent with known functional systems [Power et al., 2011, Yeo et al., 2011, Gordon et al., 2014]. Second, as already mentioned, these resting-state networks are likely task-relevant given recent demonstrations that the network architecture estimated by resting-state FC is highly similar to FC architectures present during a variety of tasks [Cole et al., 2014a, Krienen et al., 2014]. Third, in addition to reflecting large-scale connectivity patterns, resting-state FC has been shown to reflect local topological mappings between retinotopic field maps in visual cortex, highlighting the specificity with which resting-state FC conserves functionally tuned connections [Heinzle et al., 2011, Haak et al., 2013]. Finally, resting-state FC has been shown to systematically relate to task-evoked activations, allowing prediction of an individual's task-evoked activations across a variety of tasks using that individual's resting-state FC [Cole et al., 2016a, Tavor et al., 2016]. This suggests a strong role for resting-state FC in shaping task activations – a core feature of our hypothesis that resting-state FC carries the fine-grained activation patterns that represent task-relevant information.

Traditional brain information mapping approaches localize task-related brain activity patterns. Because the experimenter is doing the information decoding, it is unclear whether (or how) that information is used for downstream processing by other brain regions. Thus, such approaches embody an experimenter-as-receiver framework, rather than a cortex-as-receiver framework, which estimates how brain regions send/receive information to/from other regions [De-Wit et al., 2016]. The proposed method – information transfer mapping – advances this perspective by analogizing resting-state connections with information channels. This allowed us to characterize whether distributed brain regions receive and decode task information from other brain regions via resting-state network connections, thus ascribing an information-theoretic description to resting-state network topology. Further, above-chance information transfers between two regions would indicate that the cognitive information in those brain regions is likely supported by the intrinsic network connectivity between them. Thus, information transfer mapping implicitly tests the cognitive relevance of resting-state FC topology.

Going beyond our general hypothesis, we additionally focus on the contribution of particular features of resting-state network topology in contributing to task-related information transfer. Recent studies have identified domain-general flexible hub networks that exhibit widespread resting-state FC and high activity during cognitive control tasks [Cole et al., 2010b, Cole et al., 2013, Power et al., 2013]. The strong involvement of these cognitive control networks - the frontoparietal network (FPN, which likely implements task sets [Power and Petersen, 2013]), cingulo-opercular network (CON, which likely implements task set maintenance [Power and Petersen, 2013]), and dorsal attention network (DAN, which likely implements top-down attentional processes [Corbetta and Shulman, 2002]) - in cognitively-demanding processes suggests a role for flexibly transferring task information across regions and networks.

We sought to isolate cognitive representations that would likely involve cognitive control networks by using a cognitive paradigm that involves multiple features thought to be central to cognitive control. We used the Concrete Permuted Rule Operations (C-PRO) [Cole et al., 2010a] paradigm (Figure 2.2), which permutes rules in three different cognitive domains to produce dozens of unique task-sets. We predicted that cognitive control networks would flexibly represent task-rule information and transfer that information to other regions through their widespread intrinsic connections. The combination of experimental design and analytical framework allowed us to isolate cognitive operations and relate them to the neurobiological processes underlying activity flow mapping, thus targeting cognitive information transfer.

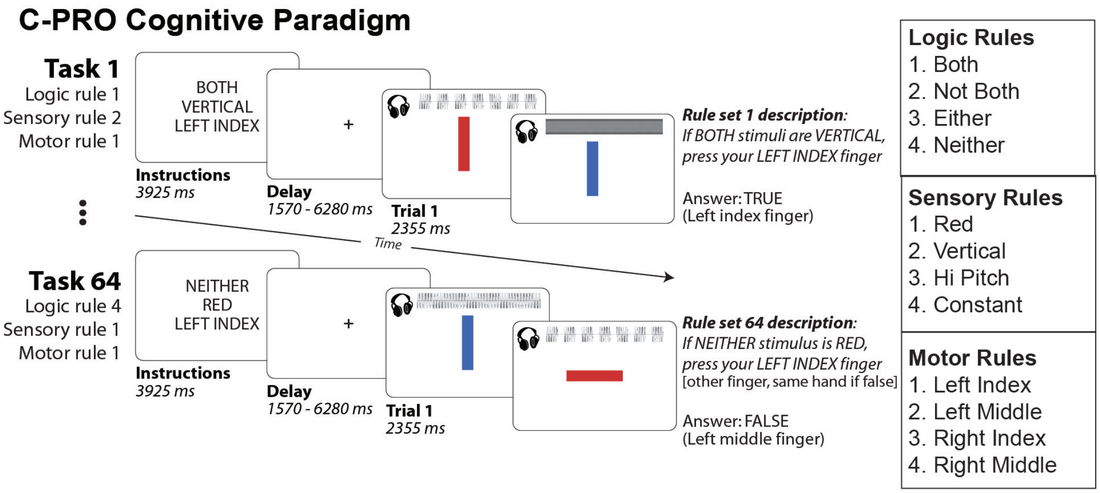


Figure 2.2: Concrete Permuted Rule Operations experimental paradigm. For a given task, subjects were presented with an instruction set (i.e., a task-rule set), in which they were presented with three rules each from a different rule domain (logic, sensory, and motor rule domains). Subjects were then asked to apply the presented rule set to two consecutively presented stimulus screens and respond accordingly. Auditory and visual stimuli were presented simultaneously for each stimulus screen. The auditory waveforms are depicted visually but were not presented visually to participants. A mini-block design was used, in which for a given set of instructions three trials were presented consecutively. The inter-trial interval was set to a constant 1570ms (2 TRs), with a jittered delay following the three trials prior to the subsequent task block (see Methods for more details). Task blocks lasted 28.26 seconds (36 TRs) each.

We began by replicating previously established properties of cognitive control networks, such as widespread resting-state FC [Cole et al., 2010b, Power et al., 2011]. We then used this replication to motivate a computational model that validates the effectiveness of the information transfer mapping procedure for estimating the role of resting-state network topology in transferring task information. Finally, we applied this framework to empirical fMRI data, allowing us to test our hypotheses that (1) resting-state FC describes channels of inter-region/network task information transfer and (2) cognitive control networks play a role in transferring task information to other regions based on their intrinsic functional network properties. Our results show that the transfer of cognitive information could be reliably predicted using resting-state network topology, and cognitive control networks were especially involved in transferring information across multiple cognitive rule domains. Based on these results and a series of control analyses that confirmed that cognitive information transfer depends on precise resting-state network topology, we conclude that cognitive information used for task performance is transferred between brain regions via the functional network topology already present during resting state.

2.3 Methods

2.3.1 Participants

35 human participants (17 females) were recruited from the Rutgers University-Newark community and neighboring communities. We excluded three subjects, leaving a total of 32 subjects for our analyses; two subjects were excluded due to exiting the scanner early, and one subject was excluded due to excessive movement. Excessive movement was defined as 3 standard deviations from the mean, in terms of framewise displacement [Power et al., 2012]. All participants gave informed consent according to the protocol approved by the Rutgers University Institutional Review Board. The average age of the participants was 20, with an age range of 18 to 29.

2.3.2 Behavioral paradigm

We used the Concrete Permuted Rule Operations (C-PRO) paradigm (Figure 2.2), which is a modified version of the original PRO paradigm introduced in Cole et al., (2010) [Cole et al., 2010a]. Briefly, the C-PRO cognitive paradigm permutes specific task rules from three different rule domains (logical decision, sensory semantic, and motor response) to generate dozens of novel and unique task sets. This creates a condition-rich dataset in the task configuration domain akin in some ways to movies and other condition-rich datasets used to investigate visual and auditory domains [Nishimoto et al., 2011, Huth et al., 2012, Simony et al., 2016]. The primary modification of the C-PRO paradigm from the PRO paradigm was to use concrete, sensory (simultaneously presented visual and auditory) stimuli, as opposed to the abstract, linguistic stimuli in the original paradigm. Visual stimuli included either horizontal or vertical oriented bars with either blue or red coloring. Simultaneously presented auditory stimuli included continuous (constant) or non-continuous (non-constant, i.e., “beeping”) tones presented at high (3000Hz) or low (300Hz) frequencies. Figure 2.2 demonstrates two example task-rule sets for “Task 1” and “Task 64”. The paradigm was presented using E-Prime software version 2.0.10.353 [Schneider et al., 2002].

Each rule domain (logic, sensory, and motor) consisted of four specific rules, while each task set was a combination of one rule from each rule domain (Figure 2.2). A total of 64 unique task sets (4 logic rules x 4 sensory rules x 4 motor rules) were possible, and each unique task set was presented twice for a total of 128 task miniblocks. Identical task sets were not presented in consecutive blocks. Each task miniblock included three trials, each consisting of two sequentially presented instances of simultaneous audiovisual stimuli. A task block began with a 3925 ms instruction screen (5 TRs), followed by a jittered delay ranging from 1570 ms to 6280 ms (2 – 8 TRs; randomly selected). Following the jittered delay, three trials were presented for 2355 ms (3 TRs), each with an inter-trial interval of 1570 ms (2 TRs). A second jittered delay followed the third trial, lasting 7850 ms to 12560 ms (10-16 TRs; randomly selected). A task block lasted a total of 28260 ms (36 TRs). Subjects were trained on four of the 64 task-rule sets for 30 minutes prior to the fMRI session. The four practiced rule sets were selected such that all 12 rules were equally practiced. There were 16 such groups of four task sets possible, and the task sets chosen to be practiced were counterbalanced across subjects. Subjects' mean performance across all trials performed in the scanner was 85% (median=86%) with a standard deviation of 8% (min=66%; max=96%). All subjects performed statistically above chance (25%).

2.3.3 fMRI Acquisition

Data were collected at the Rutgers University Brain Imaging Center (RUBIC). 35 human participants (17 females) were recruited from the Rutgers University-Newark community and neighboring communities. We excluded three subjects, leaving a total of 32 subjects for our analyses; two subjects were excluded due to exiting the scanner early, and one subject was excluded due to excessive movement. Excessive movement was defined as 3 standard deviations from the mean, in terms of framewise displacement [Power et al., 2012]. All participants gave informed consent according to the protocol approved by the Rutgers University Institutional Review Board. The average age of the participants was 20, with an age range of 18 to 29. Whole-brain multiband echo-planar imaging (EPI) acquisitions were collected with a 32-channel head coil on a 3T Siemens Trio MRI scanner with TR=785 ms, TE=34.8 ms, flip angle=55°, Bandwidth 1924/Hz/Px, in-plane FoV read=208 mm, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. Whole-brain high-resolution T1-weighted and T2-weighted anatomical scans were also collected with 0.8 mm isotropic voxels. Spin echo field maps were collected in both the anterior to posterior direction and the posterior to anterior direction in accordance with the Human Connectome Project preprocessing pipeline [Glasser et al., 2013]. A resting-state scan was collected for 14 minutes (1070 TRs), prior to the task scans. Eight task scans were subsequently collected, each spanning 7 minutes and 36 seconds (581 TRs). Each of the eight task runs (in addition to all other MRI data) were collected consecutively with short breaks in between (subjects did not leave the scanner).

2.3.4 fMRI Preprocessing

Imaging data were minimally preprocessed using the publicly available Human Connectome Project minimal preprocessing pipeline version 3.5.0, which included anatomical reconstruction and segmentation, EPI reconstruction, segmentation, spatial normalization to standard template, intensity normalization, and motion correction [Glasser et al., 2013]. All subsequent preprocessing steps and analyses were conducted on CIFTI 64k grayordinate standard space for vertex-wise analyses and parcellated time series for region-wise analyses using the Glasser et al. (2016) [Glasser et al., 2016a] atlas (i.e., one time series for each of the 360 cortical regions). We performed nuisance regression on the minimally preprocessed resting-state data using 12 motion parameters (6 motion parameter estimates plus their derivatives) and ventricle and white matter time series (extracted volumetrically), along with the first derivatives of those time series.

Task time series for task activation analyses were preprocessed in an identical manner to resting-state data. Task time series were additionally processed as follows: A standard fMRI general linear model (GLM) was fit to task-evoked activity convolved with the SPM canonical hemodynamic response function and the same 16 nuisance regressors as above. Block-by-block activity beta estimates were used for representational similarity analyses and information transfer mapping analyses. Task activity GLMs were performed at both the region-wise level and vertex-wise level for subsequent network-to-network and region-to-region information transfer mapping, respectively.

2.3.5 FC estimation

Given the success of FC estimation using multiple linear regression in our previous study [Cole et al., 2016a], we employed multiple linear regression to estimate FC. To estimate FC to a given node, we used standard linear regression to fit the time series of all other nodes as predictors (i.e., regressors) of the target nodes. Using ordinary least squares regression, we calculated whole-brain FC estimates by obtaining the regression coefficients from the equation

$$\overrightarrow{x_i} = \beta_0 + \sum_{j \neq i}^{N} \beta_{ji} \overrightarrow{x_j} + \epsilon \tag{2.1}$$

for all regions x_i. We define $\overrightarrow{x_i}$ as the time series in region x_i, β_0 as the y-intercept of the regression model, β_{ji} as the FC coefficient for the jth regressor/region (which we use as the element in the jth row and the ith column in the FC adjacency matrix), and ϵ as the residual error of the regression model. N is the total number of regressors included in the model, which corresponds to the number of all other regions. This provided an estimate of the contribution of each source region in explaining unique variance in the target region's time series. This approach was used for region-to-region FC estimation, where the time series for each parcel was averaged across a given parcel's vertices prior to FC calculation. For this model N=360, corresponding to the number of parcels in the Glasser et al. 2016 atlas [Glasser et al., 2016a]. Multiple linear regression FC is conceptually similar to partial correlation, but is actually semipartial correlation, as the estimates retain information about scaling a source time series (i.e., regressor time series) into the units of the to-be-predicted time series (i.e., predicted variable/target region).

For vertex-to-vertex FC estimation, due to computational intractability (i.e., more source vertices/regressors than time points), we used principal components regression with 500 principal components. This is the same form of regularized regression used in a previous study [Cole et al., 2016a] for voxel-to-voxel FC estimation. This approach involved reducing all source time series into 500 principal components and using the components as regressors to the target vertex. To reduce the possibility of spatial autocorrelation when estimating vertex-to-vertex FC, we excluded all vertices belonging to the same brain region/parcel as well as any vertices within 10mm of the border of that parcel in the principal components/regressors of the target vertex. (All vertices that fell within this criterion were given FC values of 0, preventing any vertices close to the target region from contaminating FC estimates.) Beta values obtained from the principal component regressors were then transformed back into the original 64k vertex space.

2.3.6 Replication of network topological properties

We sought to replicate a key property of resting-state network topology using our novel network assignments of the Glasser et al. (2016) parcels – high global connectivity of cognitive control networks. We included only functional networks which coincided with the seven most replicable functional networks found in three previously published network atlases [Power et al., 2011, Yeo et al., 2011, Gordon et al., 2014]: the frontoparietal network (FPN), the dorsal attention network (DAN), the cingulo-opercular network (CON), the default mode network (DMN), the visual network (VIS), the auditory network (AUD), and the somatomotor network (SMN). We measured the average between-network global connectivity (BGC) during resting-state FC, which was estimated using multiple linear regression (Figure 2.3d). BGC connections were defined as all connections from the source region to target regions outside the source region's network. Mathematically, we defined each region's BGC as

$$BGC_i = \frac{\sum_{j \notin C} W_{ij}}{N_{total} - N_C} \tag{2.2}$$

where BGC_i corresponds to the BGC of region i in network C, $j \notin C$ corresponds to all regions not in network C, W_{ij} corresponds to the FC estimate between regions i and j, N_{total} corresponds to the total number of regions, and N_C corresponds to the total number of regions in network C. To compute the average BGC for a network C, we averaged across all BGC_i for $i \in C$. To statistically test whether the average BGC was different for a pair of networks, we performed a cross-subject paired t-test for every pair of networks. We corrected for multiple comparisons across pairs of networks using FWE permutation testing [Nichols and Holmes, 2001].

2.3.7 Neural network model

To validate our information transfer estimation approach we constructed a simple dynamical neural network model with similar network topological properties identified in our empirical fMRI data. We constructed a neural network with 250 regions, each of which were clustered into one of five network communities (50 regions per community). Regions within the same community had a 35% probability of connecting to another region (i.e., 35% connectivity density), and regions not assigned to the same community were assigned a connectivity probability of 5% (i.e., 5% out-of-network connectivity density). We selected one community to act as a "network hub", and increased the out-of-network connectivity density of those regions to 20% density. We then applied Gaussian weights on top of the underlying structural connectivity to simulate mean-field synaptic excitation between regions. These mean-field synaptic weights were set with a mean of $1.0/\sqrt{K}$ with a standard deviation of $0.2/\sqrt{K}$, where K is the number of synaptic inputs into a region such that synaptic input scales proportionally with the number of inputs. This approach was recently shown to be a plausible rule in real-world neural systems based on in vitro estimation of between-neuron synaptic-weight-setting rules [Barral and Reyes, 2016].

To simulate network-level firing rate dynamics, as similar to Stern et al. (2014), region x_i's dynamics for $i = 1, 2, ..., 250$ obeyed the equation

$$\frac{dx_i}{dt}\tau_i = -x_i(t) + s\phi(x_i(t)) + g\big(\sum_{j \neq i}^{N} W_{ij}\phi(x_j(t))\big) + I_i(t) \tag{2.3}$$

We define the transfer function ϕ as the hyperbolic tangent, x_j the dynamics of region $j = 1, 2, ..., 250$ for $i \neq j$, $I_i(t)$ the input function (e.g., external spontaneous activity alone or both spontaneous activity and task stimulation) for $i \in [1, 250]$, W the underlying synaptic weight matrix, s the local coupling (i.e., recurrent) parameter, g the global coupling parameter, and τ_i the region's time constant. For simplicity, we set $s = g = 1$ and $\tau_i = 10ms$, though we show in a previous study [Cole et al., 2016a] that the activity flow mapping breaks down for parameter regimes $s >> g$.

We first simulated spontaneous activity in our model by injecting Gaussian noise (parameter $I_i(t)$; mean of 0.0, standard deviation 1.0). Numerical simulations were computed using a Runge-Kutta second order method with a time step of $dt = 10ms$. We ran our simulation for 600 seconds (10 minutes). To simulate resting-state fMRI, we then convolved our time series with the SPM canonical hemodynamic response function and down sampled to a 1 second TR, resulting in 600 time points. We then computed resting-state FC using multiple linear regression. To replicate the empirical data, we computed the BGC of the resting-state data (as in the empirical data; see equation 2) to validate that widespread out-of-network connectivity was preserved from synaptic to FC. To model task-evoked activity, we simulated four distinct task conditions by injecting stimulation into four randomly selected but distinct sets of twelve regions in the hub network. Stimulation to the hub network was chosen to mimic four distinct top-down, cognitive control task rules. (See Supplemental Methods for further details.) We simulated 30 subjects worth of data, and generated figures using group t-tests and controlled for multiple comparisons using FWE-correction permutation tests [Nichols and Holmes, 2001].

To perform network-to-network information transfer mapping in the model, we used the task-evoked activity (estimated by standard GLM beta estimates), and performed the information transfer mapping procedure between networks of regions using the resting-state FC matrix obtained via multiple linear regression. Network-to-network information transfer mapping is computationally identical to region-to-region information transfer mapping, and is described below.

2.3.8 Computing information estimates for regions and networks

To compute the baseline (i.e., unrelated to FC) information content at the region level (Figure 2.5), we performed a within-subject, cross-validated multivariate pattern analysis using representational similarity analysis for every Glasser et al. (2016) parcel (using the vertex-level multivariate activation pattern within each parcel). We estimated task-activation beta coefficients separately for each vertex within a region, and separately for each miniblock. Note that each miniblock was associated with a specific task-rule condition for each rule domain. Mathematically, we defined IE_B, the information estimate of region B, as

$$IE_B = Match_B - Mismatch_B \tag{2.4}$$

where $Match_B$ and $Mismatch_B$ correspond to the averaged Spearman rank correlation for matched and mismatched conditions, respectively. Specifically, we define $Match_B$ and $Mismatch_B$ as

$$Match_B = \frac{\sum_{k=1}^{K} scorr(B_k, B_{match})}{K} \tag{2.5}$$

$$Mismatch_B = \frac{\sum_{k=1}^{K}[\sum_{n=1}^{N}(scorr(B_k, B_{mismatch_n})/N]}{K} \tag{2.6}$$

where K corresponds to the total number of miniblocks (in this paradigm, 128 miniblocks), $scorr$ corresponds to a Fisher z-transformed Spearman's rank correlation between two activation vectors, B_k is the activation pattern in region B during block k, B_{match} is the task-rule condition prototype (obtained by averaging across blocks of the same condition, holding out block k) of region B's activation pattern for which block k's condition matches the condition prototype, and $B_{mismatch_n}$ as the task-rule condition prototypes for which block k's condition does not match. (In the present study $N = 3$, since each rule dimension has four task-rule conditions, and for a given miniblock there's one match and three mismatched conditions.) To avoid circularity, we performed a leave-four-out cross-validation scheme, holding out a miniblock of each task-rule. This ensured that miniblock B_k was not included in constructing the condition prototype B_{match} and that condition prototypes were each constructed using the same number of miniblocks. Prior to running the representational similarity analysis, all blocks were spatially demeaned to increase the likelihood that the representations we were identifying was a multivariate regional pattern (rather than a change in region-level mean activity). Use of Spearman's rank correlation also reduced the likelihood that the identified multivariate representation patterns were driven by mean activity changes or a small number of outlier values.

Statistical significance was assessed by taking a one-sided group t-test against

0 for each region's information estimate across subjects, since a greater than 0 difference of matches versus mismatches indicated significant representation of specific task rules. All p-values were corrected for multiple comparisons across the 360 parcels using FWE-correction with permutation tests [Nichols and Holmes, 2001], and significance was assessed using an FWE-corrected threshold of $p < 0.05$.

(See Supplementary Methods for details on estimating network-level information estimates for Supplementary Figure A.1b.)

2.3.9 Region-to-region information transfer mapping

We extended the original activity flow mapping procedure as defined in Cole et al. (2016) [Cole et al., 2016a] (Figure 2.1a) to investigate transfer of task-related information between pairs of brain regions using vertex-wise activation patterns (i.e., region-to-region activity flow mapping; Figure 2.1b). This involved predicting the activity of the vertices of a held-out target region based on the vertices within a source region. Mathematically, we define region-to-region activity flow mapping between regions A and B as

$$\bar{B}_k = A_k \cdot W_{RSFC} \tag{2.7}$$

where $\bar{B}_k$ corresponds to the predicted activation pattern vector for the target region B, A_k corresponds to region A's activation pattern vector (i.e., the source region), W_{RSFC} corresponds to the vertex-to-vertex resting-state FC between regions A and B, and the operator $\cdot$ refers to the dot product. This formulation allowed us to map activation patterns in one region's spatial dimension to the spatial dimension of another region.

To test the extent that task representations are preserved in the region-to-region multivariate predictions, we quantified how much information transfer occurred between the two regions. Briefly, information transfer mapping comprises three steps, illustrated in Figure 2.1c: (1) Region-to-region (or network-to-network) activity flow mapping; (2) A cross-validated representational similarity analysis between predicted activation patterns and actual, held-out activation patterns; (3) Information classification/decoding by computing the difference between matched condition similarities and mismatched condition similarities. This final step produces an information transfer estimate. Mathematically, our information transfer estimate was derived using the exact formulation (equations 5 and 6) as our information estimate formula, but we substituted the target region's actual activation pattern B_k for the target region's predicted activation pattern $\bar{B}_k$ based on a connectivity-based transformation of source region A's activation pattern. (See Supplementary Methods materials for more details.) Information transfer mapping was performed within subject between every pair of regions in the Glasser et al. (2016) [Glasser et al., 2016a] atlas (360 regions in total). Statistical tests were performed using a group one-sided t-test (t > 0) for every pair-wise mapping. Our use of mismatched correlations as a baseline ensured that any positive information transfer estimates was a result of a task-rule-specific representation, rather than a task-general effect. Any information estimate that was not significantly greater than 0 indicated that the predicted-to-actual similarity was at chance (akin to chance decoding using classifiers). We tested for multiple comparisons using permutation testing [Nichols and Holmes, 2001] for every region-to-region mapping, and significance was assessed using FWE-corrected p-values with $p < 0.05$. Note that to avoid circularity for region-to-region information transfer mapping, any vertices in a source region that fell within a 10mm radius of the to-be-predicted target region (e.g., an adjacent region) would not contribute any activity flow to the to-be-predicted target region (see FC estimation Methods section for details). (See Supplementary Methods for further details.)

2.3.10 Network-to-network information transfer mapping

Network-to-network information transfer mapping in both the computational model (Figure 2.4e) and empirical data (Supplementary Figure A.1c-e) was performed in the same computational framework as above, though instead of predicting region-level activation patterns using vertex-level activation patterns, network-level activation patterns were predicted using region-level activations (averaging across vertices within a given region). (See Supplementary Methods for more details.)

2.3.11 Behavioral relevance of information transfers

To characterize the behavioral relevance of information transfers, we performed a within-subject analysis to decode task performance using miniblock-by-miniblock information transfer estimates. We first sought to ensure that baseline miniblock information estimates could decode miniblock task performance within subjects prior to the information transfer mapping procedure. To perform a given task, knowledge of all three rule domains (i.e., logic, sensory, and motor rule domains) is required. Thus, we constructed a decoding model with logistic regression, training the model to decode the task performance of a given miniblock using the information estimates of a given brain region across all three rule domains. The model was tested using cross-validation in MATLAB using the glmfit function (with the logit link function). Miniblocks with over 50% of trials performed correctly were predicted as a 1, and 0 otherwise. However, to account for the imbalanced training data (on average, subjects performed 85% of trials correctly), we removed the intercept term β_0 to center our predictions (as computed by a logistic function) at 0.5 (see Supplementary Methods for further details).

We applied our decoding model to all regions within the FPN and CON across subjects. For each region, we applied one-sided t-tests against chance (50%), and corrected for multiple comparisons using FWE-correction permutation tests [Nichols and Holmes, 2001]. We identified a single FPN region in the LPFC (left hemisphere region 80 in the Glasser et al. atlas; Supplementary Figure A.5) whose baseline information estimates predicted miniblock task performance.

We subsequently tested whether information transfer estimates from the LPFC region could predict task performance. We applied the decoding model to information transfer estimates across all rule domains for all information transfers from the LPFC region to all other FPN and CON regions. We performed one-sided t-tests against chance (50%) for each information transfer, and corrected for multiple comparisons using FWE-correction permutation tests [Nichols and Holmes, 2001]. We identified a single information transfer from the LPFC to the OFC (left hemisphere region 91; both FPN regions) that survived multiple comparisons with an FWE-corrected $p < 0.05$. Surface visualizations for Supplementary Figure A.5 were made using Connectome Workbench software (version 1.2.3) [Glasser et al., 2016b].

2.3.12 Computational resources

Region-to-region information transfer mapping, vertex-to-vertex FC estimation, task-rule information estimation, and model simulations were performed on the Rutgers University-Newark supercomputer cluster (Newark Massive Memory Machine, NM3) using Python and MATLAB code.

2.3.13 Code and data availability

We have included code demos with accompanying tutorial data for both our computational model and the empirical network-to-network information transfer mapping. We have also provided a GitHub repository with both MATLAB and Python code to run FWE-correction using permutation tests using the approach described in [Nichols and Holmes, 2001]. Lastly, we have published all master scripts/jupyter notebooks used to generate results and figures in the manuscript. All other data presented in this study are available upon request.

Demo code for the information transfer mapping procedure is publicly available here: `https://github.com/ColeLab/informationtransfermapping`

Code for the FWE-correction via permutation testing is available here: `https://github.com/ColeLab/MultipleComparisonsPermutationTesting`

2.4 Results

2.4.1 Network organization of cognitive control networks

We began by establishing a strong basis for testing subsequent hypotheses regarding information transfer via cognitive control networks. Given the recent interest in reproducibility in neuroscience and other fields [Button et al., 2013, Szucs and Ioannidis, 2016], we replicated the hub-like characteristic of cognitive control networks [Cole et al., 2010b, Power et al., 2011, Power et al., 2013, van den Heuvel and Sporns, 2013] before moving forward with analyses that build on these previous findings.

Using a recently developed set of functionally defined cortical regions [Glasser et al., 2016a] (Figure 2.3a), we tested whether cognitive control networks are global (connector) hubs. We quantified global hubs as having high between-network global connectivity (BGC) (see Methods) estimated during resting-state fMRI using FC estimated with multiple regression (Figure 2.3c). Standard Pearson correlations (Figure 2.3b) were not used to compute BGC, given that Pearson correlations likely inflate the overall number of connections. We constrained our analyses to seven networks (Figure 2.3a), identified by being replicated across multiple previously published functional network atlases [Power et al., 2011, Yeo et al., 2011, Gordon et al., 2014]. We focused on

BGC to reduce the bias toward larger mean connectivity (i.e., weighted degree centrality, or global brain connectivity [Cole et al., 2010b]) for larger networks simply because they are larger [Power et al., 2011, Power et al., 2013]. We found that the top three networks with highest BGC estimated at rest were the three cognitive control networks: FPN, CON, and DAN (Figure 2.3d; FPN greater than all non-cognitive control networks, with an averaged t(31)=9.52; CON greater than all non-cognitive control networks, with an averaged t(31)=12.33; DAN greater than all non-cognitive control networks, with an averaged t(31)=11.56; all family-wise error (FWE) corrected $p < 0.0001$). These results replicated previous results suggesting cognitive control networks are global hubs [Power et al., 2011, Power et al., 2013, Cole et al., 2010b, van den Heuvel and Sporns, 2013], strengthening the basis for our hypothesis that cognitive control networks play a disproportionate role in shaping information transfer between regions throughout the brain. We test this hypothesis in a subsequent section, after establishing the validity of the newly-developed information transfer mapping procedure.

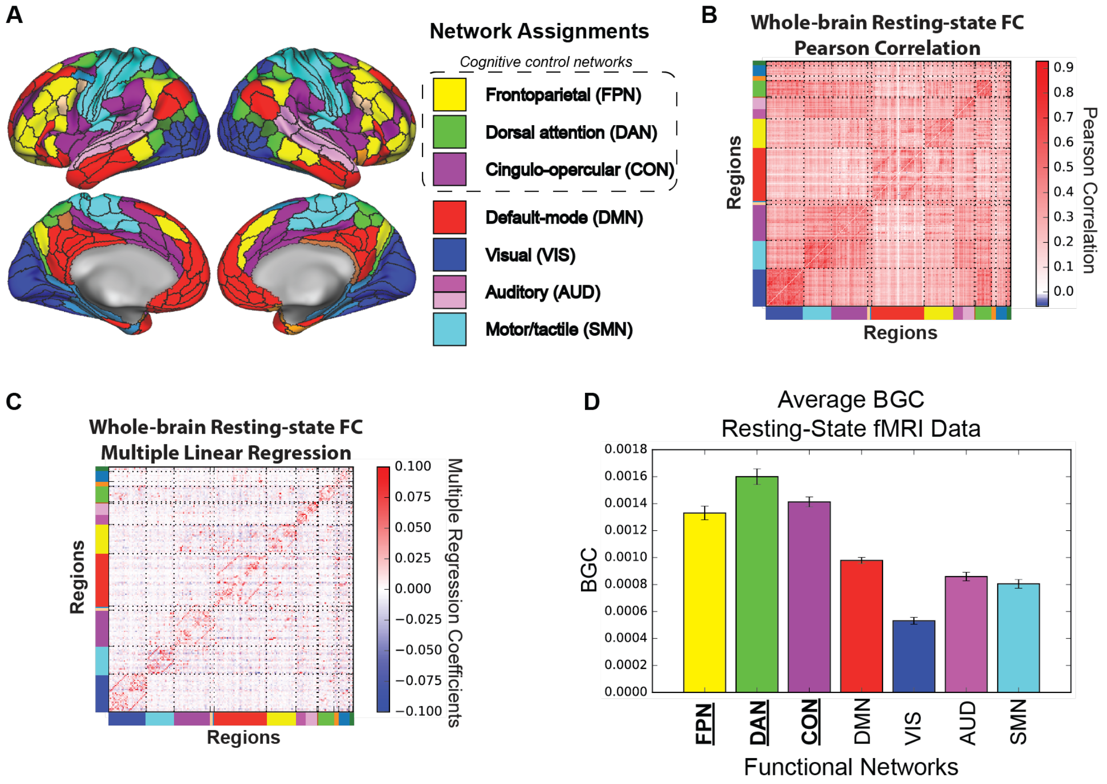


Figure 2.3: Large-scale network organization during rest. A) Using a recently released, multimodal parcellation of the human cerebral cortex [Glasser et al., 2016a], we assigned each region to a functional network using the Generalized Louvain method for community detection with resting-state fMRI data. We designated functional labels to seven networks that were replicated with other network assignments [Power et al., 2011, Yeo et al., 2011, Gordon et al., 2014]. B) Whole-brain resting-state FC matrix computed using Pearson correlation between regions in panel A. Colors along the rows and columns denote network assignments from panel A. C) Whole-brain resting-state FC matrix computed using multiple linear regression. For every region's time series, we fitted a multiple linear regression model using the time series of all other regions as regressors of the target region. Multiple regression FC strongly reduced the chance that a connection was indirect, since FC estimates are based on unique shared variance. We used multiple regression FC for information transfer mapping, suggesting the estimated information transfers were likely direct rather than indirect. D) Averaged BGC of resting-state fMRI for each defined functional network. Cognitive control networks (underlined and in bold) had higher average BGC estimates relative to non-cognitive control networks (i.e., DMN and sensorimotor networks; FWE-corrected p<0.05). Error bars reflect across-subject standard error.

2.4.2 Computational validation of information transfer mapping

We previously established that whole-brain activation patterns can be predicted based on activity flow over resting-state networks [Cole et al., 2016a]. However, it remains unclear whether one region's cognitive information – coded as fine-grained activation patterns – can by predicted based on activity flow over resting-state FC. Such a demonstration would indicate that resting-state FC carries cognitive task information between brain regions (and networks). We tested this possibility by shifting from an "all-to-one" activity flow approach (i.e., predicting the activity level of a single brain region using the activity flow from all other brain regions; Figure 2.1a) to modeling activity flow between a pair of regions (i.e., using the fine-grained activation pattern within one brain region to predict the fine-grained activation pattern within another region; Figure 2.1b).

Testing our hypothesis required developing a new approach – information transfer mapping – which quantifies the amount of information transferred between pairs of brain regions over resting-state FC (Figure 2.1b,c). Broadly, information transfer mapping tests the ability of resting-state FC topology (fine-grained connectivity patterns) to describe the mappings between cognitive-task-related activity patterns between pairs of brain regions. Specifically, each mapping (described by resting-state FC topology) must preserve the representational space between two regions, such that task-evoked information is decodable after the connectivity-based mapping. Beyond improving empirical understanding, this approach may have important theoretical implications given that it bridges biophysical (intrinsic FC) and computational (transformations between information-carrying activity patterns) properties into a convergent framework.

This approach (Figure 2.1c) predicts the activation pattern in a target region based on a source region's activation pattern. This predicted activation pattern is then compared to the target region's actual activation pattern during the current task condition. The matched condition predicted-to-actual similarity is then compared to the mismatched condition predicted-to-actual similarity, with the difference in similarity quantifying the amount of task-specific information present in the prediction. Since the prediction was based on estimated activity flow over resting-state FC patterns, this allowed us to infer the amount of task-relevant information transferred via resting-state FC. Note that it was important to compare the predicted with the actual activation pattern in the target region to ensure that our prediction preserved the same representational geometry [Diedrichsen and Kriegeskorte, 2017] as the actual activation pattern.

We validated this approach using a simple abstract neural network model with one hub network and four non-hub networks (see Methods; Figure 2.4a). This network organization was the basis for simulating fMRI dynamics during rest and task states, which allowed us to establish a “ground truth” to test the efficacy of the information transfer mapping procedure. This validation-via-modeling method was highly similar to the simple neural network model we previously used to validate the original activity flow mapping approach [Cole et al., 2016a]. Using Wilson-Cowan type firing rate dynamics [Stern et al., 2014, Cowan et al., 2016], we simulated resting state and four distinct task states, simulated the transformation of the simulated neural signals to fMRI data (see Methods), and estimated resting-state FC (Figure 2.4b) and task-evoked fMRI runs for each of the four task conditions (Figure 2.4c). Note that we focused on network-to-network information transfer for our model validation (see schematic in Supplementary Figure A.1a), but later extended the approach to region-to-region information transfer.

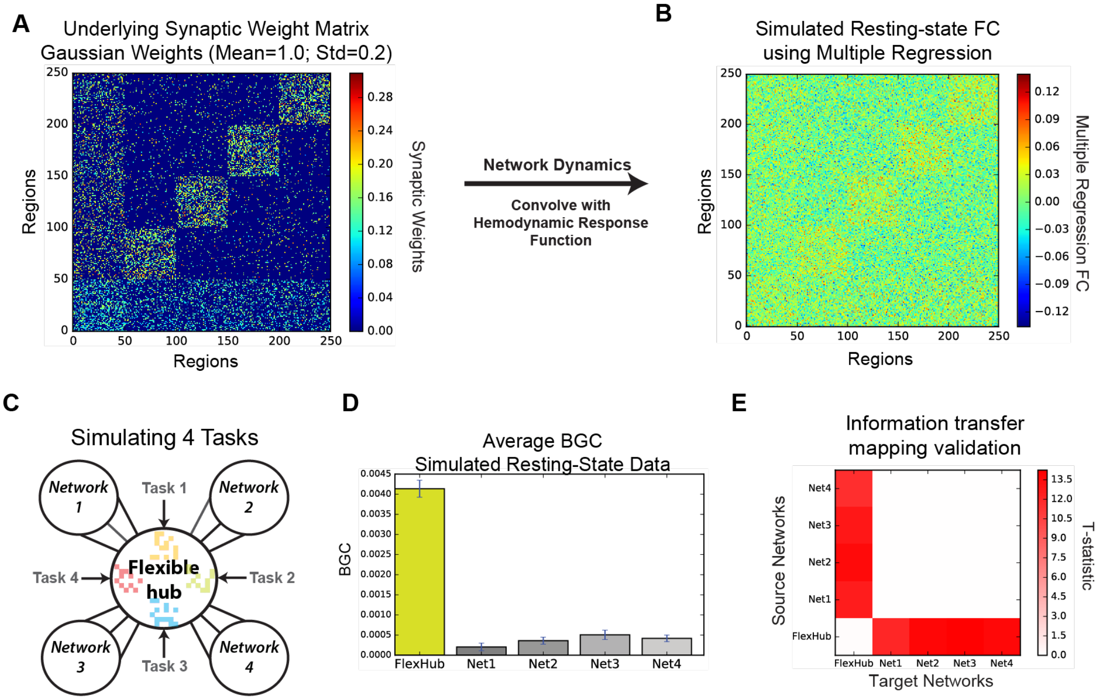


Figure 2.4: Computational validation of network-to-network information transfer mapping. A) Underlying synaptic weight matrix with four local networks and one hub network. We constructed an abstract neural network with a single hub network to see the relative effect of information transfer from the hub network to downstream local networks, similar to the hypothesized computational function of the cognitive control networks during task. B) Recovering large-scale synaptic organization via multiple regression FC estimates on a simulated resting-state time series. C) We simulated four 'cognitive control tasks' by stimulating four distinct ensembles of regions within the hub network. D) Increased BGC estimated at rest reflects underlying synaptic organization. Error bars represent across-subject standard error. E) Thresholded information transfer estimates between pairs of networks in a neural network model. Each row in the matrix corresponds to a source network from which we mapped activation patterns to other target networks using the information transfer mapping procedure (Figure 2.1c). Each column in the matrix corresponds to a target network to which we compared the predicted-to-actual activation patterns. FWE-corrected thresholded T-statistic map with $p < 0.05$.

We found that simulated resting-state FC accurately reflected high BGC for the hub network (BGC statistically greater for the hub-network versus all other networks; averaged t(29)=21.14; FWE-corrected $p < 0.0001$; Figure 2.4d). Further, given the underlying synaptic connectivity structure (Figure 2.4a) and the estimated intrinsic topology via resting-state FC (Figure 2.4b,d), we hypothesized that information transfer to and from the hub network would reliably preserve task-specific information. Using the information transfer mapping approach (Figure 2.1c; see Methods), we quantified the amount of information transfer via activity flow between every pair of networks (Figure 2.4e). We found that information transfers to/from the flexible hub network and non-hub networks preserved task-specific representations (averaged information transfer estimate=0.13; averaged t(29)=11.86; FWE-corrected $p < 0.0001$), while transfers between pairs of non-hub networks did not preserve statistically significant representations (averaged information transfer estimate=-0.0002; averaged t(29)=-0.02; averaged FWE-corrected $p = 0.91$). We also found that these results were consistent with simulations where both top-down (hub network) and bottom-up (local network) stimulation occurred simultaneously (Supplementary Figure A.3; see Supplementary Methods). These results suggest that FC estimates obtained during simulated resting-state fMRI dynamics reflected underlying synaptic organization enough to describe the task-information-carrying mappings that govern activity flow between functional networks – a key assumption underlying our new approach.

These model simulations validated the plausibility of two hypotheses critical to the proposed information transfer mechanism: (1) Resting-state FC estimates characterize intrinsic FC (potentially reflecting aggregate synaptic connectivity) effectively enough to reflect underlying communication channel capacities; (2) Intrinsic FC describes the information-preserving mappings necessary to predict task-relevant activation patterns transferred from one region or network to another. Thus, these results validated the analytical basis of estimating information transfer via activity flow, which is applied to network-to-network and region-to-region information transfer mapping with empirical fMRI data below.

2.4.3 Information transfer via resting-state network topology

We next applied the information transfer mapping procedure to real fMRI data, testing its ability to infer cognitive information transfer in the human brain. To test the hypothesis that cognitive control networks might widely distribute cognitive information via their resting-state network topology, we used an experimental paradigm with several features central to cognitive control to engage cognitive control networks. First, we used novel tasks given the need for control to specify behavior in such under-practiced scenarios [Rabbitt, 1997, Cole et al., 2013]. Second, we used complex tasks given the need to deploy additional cognitive control resources when working memory is taxed [Miller and Buschman, 2015]. Finally, we used a variety of abstract rules given that such rules are thought to be represented within cognitive control networks [Cole et al., 2011b, Cole et al., 2015, Muhle-Karbe et al., 2016]. Using many fully-counterbalanced rules also allowed us to test our hypotheses across a variety of task conditions (while controlling for differences in sensory stimuli during trials). These features converged in the Concrete Permuted Rule Operations paradigm (C-PRO; Figure 2.2). This paradigm was developed as part of this study, and is a modified version of the PRO paradigm [Cole et al., 2010a]. We predicted that cognitive control networks would flexibly represent C-PRO rule information and transfer that information to other regions through their widespread intrinsic connections. For simplicity, we began with large-scale network-to-network information transfers. This involved quantifying information in large-scale functional networks based on patterns of region-level task activations (Supplementary Figure A.1; see Methods). In subsequent analyses we focused on region-to-region information transfers (based on patterns of voxel/vertex-level task activations).

As a prerequisite to running the network-to-network information transfer tests, we sought to first establish that task-rule information from the C-PRO paradigm (Figure 2.2) was widely distributed across entire functional networks (Supplementary Figure A.1b). Logic rule information was significantly decodable in 6 out of 7 of the functional networks (averaged information estimate of significant effects=0.03; averaged significant t(31)=4.89; FWE-corrected $p < 0.01$), with the somatomotor network (SMN) being the single network that did not contain decodable logic rule information (information estimate=0.007; t(31)=1.22; FWE-corrected $p = 0.58$). Sensory rule information was significantly decodable in the FPN, DAN, CON, and visual network (VIS) (averaged information estimate=0.03; averaged t(31)=5.14; FWE-corrected $p < 0.001$), and not decodable in the default mode network (DMN), auditory network (AUD), and SMN (averaged information estimate=0.003; averaged t(31)=0.83; averaged FWE-corrected $p > 0.11$). Motor rule information was significantly decodable in the DAN, CON, and SMN (averaged information estimate=0.08; averaged t(31)=7.26; FWE-corrected $p < 0.0001$), and not decodable in the FPN, DMN, VIS, and AUD (averaged information estimate=0.006; averaged t(31)=1.93; averaged FWE-corrected $p > 0.05$). This allowed us to then evaluate whether significantly decodable representations of information were transferred to other functional networks.

In the logic rule domain, we identified information transfers between the FPN, CON, DMN, and AUD networks (Supplementary Figure A.1c; averaged information transfer estimate=0.009; averaged t(31)=4.73; FWE-corrected $p < 0.02$). In the sensory rule domain, we found information transfers between the DAN and VIS in addition to the FPN, CON, and DMN (Supplementary Figure A.1d; averaged information transfer estimate=0.006; averaged t(31)=4.01; FWE-corrected $p < 0.05$). Lastly, in the motor rule domain, information transfers were between the DAN, CON, and the SMN (Supplementary Figure A.1e; averaged information transfer estimates=0.011; averaged t(31)=5.37; FWE-corrected $p < 0.01$).

Further, to ensure that information transfers between pairs of networks was dependent on the precise network-to-network FC topology, we performed permutation testing, permuting FC patterns between pairs of networks (see Supplementary Methods). Indeed, after statistical testing, we found that information transfers were identical to our results with parametric statistical testing, suggesting that the observed information transfers were dependent on the specific resting-state network FC topology (Supplementary Figure A.4). These empirical network-to-network information transfers, along with their dependence on specific resting-state FC patterns, establish a role for resting-state network topology in transferring cognitive task information.

We next focused on region-to-region mappings that, unlike the network-to-network transfers, are based on fine-grained vertex-wise patterns. As a prerequisite to testing for information transfer between pairs of regions, we first needed to establish whether regions contained decodable task-rule representations. Thus, we first quantified the information content of each rule domain in the C-PRO paradigm (logic, sensory and motor rule domains) for each of the 360 regions using activation patterns (at the vertex level) with a cross-validated representational similarity analysis (see Methods). We found that logic rules were relatively distributed, with highest-quality representations in frontal and parietal cortices (averaged information estimate across significant effects=0.02; averaged t(31)=5.24; FWE-corrected $p < 0.05$; Figure 2.5a). Sensory rule information was also relatively distributed (averaged information estimate across significant effects=0.02; averaged t(31)=4.97; FWE-corrected $p < 0.05$; Figure 2.5b), though the highest-quality representations were predominantly in visual areas. Lastly, we found that motor rule representations were significantly more localized, with the highest-quality representations in the somatomotor network (averaged information estimate across significant effects=0.06; averaged t(31)=6.80; FWE-corrected

$p < 0.05$; Figure 2.5c). The existence of distributed task-rule information in multiple cortical regions allowed us to next assess how task-rule-specific information in one region might be transferred to other regions.

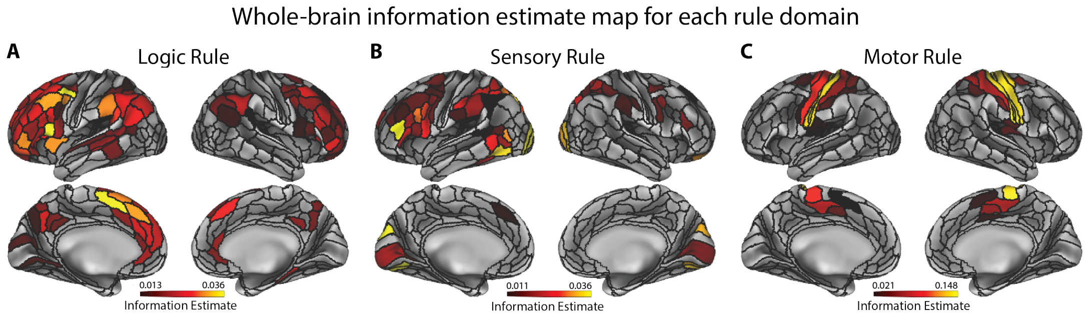


Figure 2.5: Information estimates of each region for each task-rule domain, prior to information transfer mapping. All reported results were statistically significant at FWE-corrected $p < 0.05$. A) Thresholded whole-brain logic rule information estimate map. A cross-validated representational similarity analysis (quantifying degree of information representation; see Methods) for the logic rule domain was computed using vertices within every region. For each region, an average information estimate was computed for each subject, and a one-sided t-test was computed against zero across subjects. B) Thresholded whole-brain sensory rule information estimate map. As in the logic rule analysis, rule representations were highly distributed across the entire cortex, though representations were especially prominent in visual areas. C) Thresholded whole-brain motor rule information estimate map. Unlike the logic and sensory rule representations, motor rule representations were more localized to the motor/tactile network.

We next performed region-to-region information transfer mapping (Figure 2.6). This approach utilized within-region vertex-level activation patterns along with vertex-to-vertex resting-state FC between regions to predict information content in each region (Figure 2.6a; also see Methods). We performed this procedure for every pair of 360 regions, and visualized our results as a 360x360 matrix for each rule domain (Figure 2.6b,d,f). However, given the difficulty in visually interpreting information transfers between every pair of regions (due to sparseness), we collapsed the region-to-region information transfer matrix by network to better visualize statistically significant region-to-region information transfers at the network level (Figure 2.6c,d,g; see Supplementary Figure A.2a-c for all 14 networks). In addition, to see the relative anatomical position of regions that transferred information (i.e., source regions), we computed the percent of statistically significant transfers from each cortical region for each rule domain, and plotted these percentages on the cortical surface (Figure 2.7a-c).

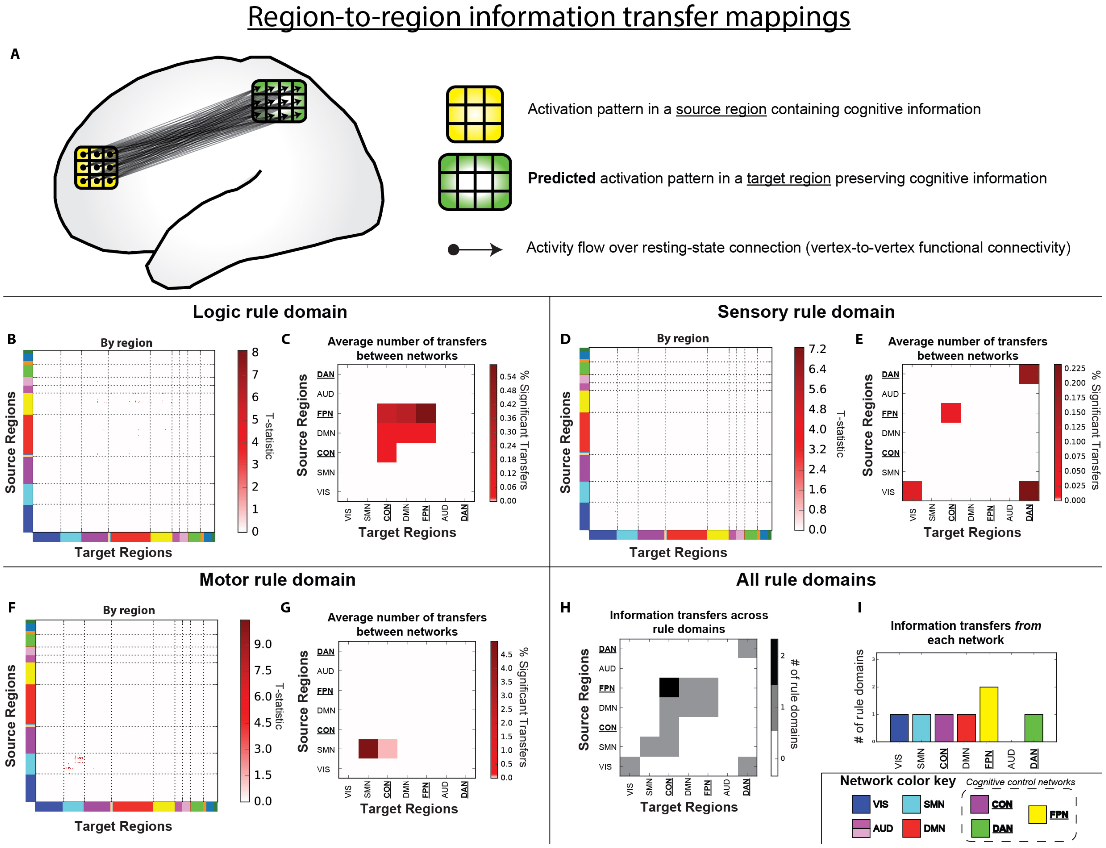


Figure 2.6: Information transfer mappings between all pairs of regions. All reported results were statistically significant at $p < 0.05$ (FWE-corrected). (See Supplementary Figure A.6 for results with an FDR-corrected $p < 0.05$ threshold.) A) Region-to-region information transfer mapping used the vertex-level activation pattern within one brain region and the fine-grained region-to-region resting-state FC topology to predict the vertex-level activation pattern in another brain region. B) Logic rule region-to-region information transfer mapping. C) Average number of statistically significant region-to-region transfers by network affiliations. To better visualize and assess how region-to-region transfer mappings may have been influenced by underlying network organization, we computed the percent of statistically significant rule transfers for every network-to-network configuration (i.e., the percentage of region-to-region transfers from a network A to a network B). (Note that visualizations for the full 14 network partition can be found in Supplementary Figure 2.2.) Cognitive control networks are underlined. Information transfer of logic rule information was distributed across frontal and parietal cortices. D,E) Statistically significant sensory rule region-to-region information transfers. Region-to-region information transfers were substantially sparser for sensory rule mappings, but involved DAN and VIS regions. F,G) Statistically significant motor rule region-to-region information transfers. Motor rule mappings were noticeably more localized within the motor network. H) Statistically significant information transfers between regions grouped by network affiliation across rule domains. Across the three rule domains (panels C, E, and G) we counted the number of rule domains information was transferred between networks. I) We performed a similar analysis as in panel H, but counted the number of rule domains a network contained a region that transferred information (as a source region) across the three rule domains.

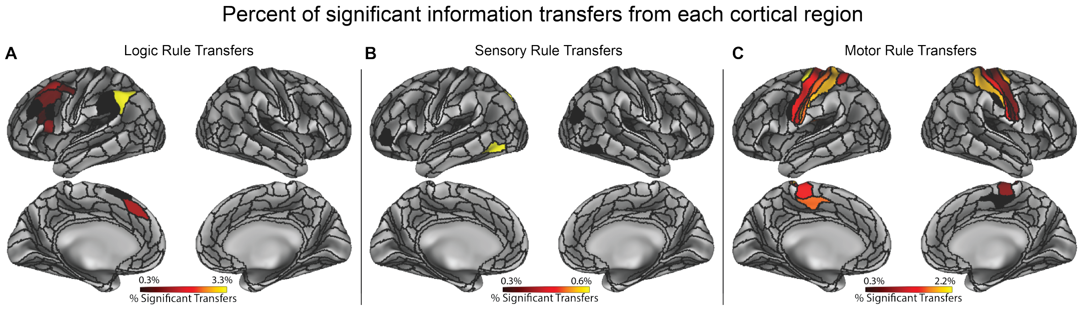


Figure 2.7: Percent of statistically significant information transfers from each cortical region. All reported information transfers were statistically significant at $p < 0.05$ (FWE-corrected). (See Supplementary Figure A.7 for results with an FDR-corrected $p < 0.05$ threshold.) A) Percent of statistically significant information transfers from each region for the logic rule domain. Percentages were computed by taking the number of significant transfers from each region, and dividing it by the total number of possible transfers from that region (359 other regions). Information transfers were relatively distributed, yet were predominantly from frontal parietal cortices. B) Percent of statistically significant information transfers from each region for the sensory rule domain. Information transfers were much sparser than in the logic rule domain. Most transfers were from higher-order visual areas and the DAN. C) Percent of statistically significant information transfers from each region for the motor rule domain. Transfers were predominantly from the motor network.

Overall, region-to-region information transfers were detected (FWE-corrected $p < 0.05$) for all three task rule domains, as described in detail below. However, given the conservative nature of FWE correction, we also provide region-to-region information transfer results for false discovery rate [Genovese and Wasserman, 2002] (FDR) corrected $p < 0.05$ thresholds, which potentially reduced false negatives but increased false positives (Supplementary Figures A.6 & A.7). We found that with FDR correction, information transfers between regions were significantly more distributed (particularly in the logic rule domain). In both cases, these findings support the hypothesis that resting-state FC topology describes the channels of information transfer across multiple functional networks and across multiple task-content domains.

For logic rule mappings, while information transfers were highly distributed, most statistically significant region-to-region information transfers predominantly involved the FPN and other frontoparietal regions (averaged information transfer estimate across significant effects=0.02; averaged t(31)=6.26; FWE-corrected $p < 0.05$). In particular, regions within the FPN transferred information to other regions in the FPN, as well as regions in other domain-general networks (CON and DMN) (Figure 2.6c). Further, source regions involved in the transfer of logic rule information were left-lateralized for FWE-corrected $p < 0.05$ (Figure 2.7a), although FDR-corrected $p < 0.05$ thresholds showed more distributed source regions across bilateral frontal and parietal cortices (Supplementary Figure A.7a). In both cases, these findings suggest that the FPN uses intrinsic FC topology to distribute abstract (e.g., logic) rule information broadly for task set implementation and maintenance.

For sensory rule mappings, we found high specificity and sparseness of region-to-region task information transfers (averaged information transfer estimate across significant effects=0.01; averaged t(31)=6.15; FWE-corrected $p < 0.05$; Figure 2.7b). Most notably, we found that sensory rule representations are predominantly transferred within and between the DAN and VIS networks, as well as the FPN and CON (Figure 2.6d,e). Previous studies have implicated a prominent role of the DAN and VIS in attentional processing of sensory information, consistent with the observed information transfers [Corbetta and Shulman, 2002]. These findings suggest sensory rule information may be transferred between cognitive control networks, with transfers between regions in the DAN and VIS implementing these top-down information transfers.

Lastly, we found the most information transfer specificity for motor rule information (averaged information transfer estimate across significant effects=0.09; averaged t(31)=7.38; FWE-corrected $p < 0.05$), consistent with the relatively localized representations of motor rule information (Figure 2.5c). In particular, transfer of motor rule information largely involved transfers from regions in the SMN (Figure 2.7c), while between-network information transfer with the SMN primarily involved the CON (Figure 2.6g).

We next characterized the rule-domain generality of information transfers between specific networks. We found that regions within the FPN transferred rule information to the CON across two out of the three rule domains (Figure 2.6h; see Supplementary Figure A.2d for all 14 networks). In addition, using an FDR-corrected threshold of $p < 0.05$, we found statistically significant information transfers from FPN to CON for all three rule domains (Supplementary Figure A.6d). This is consistent with theories suggesting the FPN coordinates with CON to maintain and implement task sets [Power and Petersen, 2013].

We next tested for networks that consistently transferred information across all rule domains, regardless of the target region's network affiliation. We found that regions in the FPN were consistently involved in transferring information to other regions in two rule domains (Figure 2.6h). When using FDR to correct for multiple comparisons, we found that the FPN, DAN, and DMN transferred task information in all three rule domains (Supplementary Figure A.6e). We next assessed whether a single region transferred information across multiple rule domains. We found that no individual region consistently transferred task-rule information across the rule domains with either FWE or FDR correction, which suggests that unique sets of regions within each network were involved in transferring distinct types of cognitive information. This suggests that the regions within the FPN (and the DAN and DMN for FDR-corrected $p < 0.05$ significance testing; Supplementary Figure A.6e) collectively act as flexible hub networks to communicate task-rules in different cognitive domains. Thus, the FPN likely plays an important role in task-rule transfers, regardless of cognitive domain.

These results uncover two key findings: (1) resting-state network topology describes the mappings likely underlying information transfer across distributed regions and functional networks, and (2) cognitive control networks likely play especially important roles in transferring a wide-range of task-rule information during complex cognitive tasks.

2.4.4 Behavioral relevance of cognitive information transfer

We next tested whether estimated information transfers are predictive of task performance, demonstrating a likely role of information-pattern transfers in supporting task performance. Given that successful task performance required cognitive encoding of all three rule types (i.e., logic, sensory, and motor rules), we hypothesized that information transfer of all three rules were important to performing a task correctly. We therefore constructed a decoder using multiple logistic regression that was trained on the miniblock-to-miniblock information transfer estimates for all three rule types, and predicted the overall accuracy for held-out miniblocks (i.e., predicted a 1 if greater than 50% of trials were performed correctly within a miniblock, and 0 otherwise). Successful decoding of task performance using information transfer between pairs of regions would suggest that task performance depends in part on the successful transfer of task-rule information between those regions.

We first sought to ensure that task-rule information coded in the activity patterns used for information transfer mapping could predict behavioral performance, as a prerequisite to performing the information transfer mapping procedure. Given our findings that transfers between the FPN and CON were involved in two out of the three rule domains and that the FPN and CON are known to be involved in task-set maintenance [Dosenbach et al., 2006] we constrained our search to regions within those two networks. We found that rule information estimates (see Supplementary Methods) in a single FPN region in the lateral prefrontal cortex (LPFC) could significantly decode task performance (decoding accuracy=52.6%; t(31)=3.97; FWE-corrected $p = 0.02$).

We then used this region as a source region and decoded task performance using information transfer estimates (across all rule domains) for transfers to every other region in the FPN and CON. We found that information transfer estimates from the LPFC region to an FPN region in the orbitofrontal cortex (OFC) could decode miniblock task performance significantly above chance (decoding accuracy=53.2%; t(31)=4.76; FWE-corrected $p = 0.003$; Supplementary Figure A.5). This result demonstrates that the transfer of cognitive task-rule information between the LPFC and OFC was significantly correlated with task performance. However, while we account for the imbalance of correct and error trials in our decoding model, given that the behavioral data contains significantly fewer incorrect versus correct miniblocks, we interpret these results cautiously. (On average, 85% of miniblocks were performed correctly.) It will be important for future work to investigate the behavioral relevance of information transfers using a dataset that contains more error trials, allowing for a more robust model fit to behavior. Nonetheless, the combination of linking resting-state FC topology to information transfers across multiple brain systems and multiple cognitive task domains as well as trial-by-trial task performance strongly supports a role for resting-state FC topology in cognitive information transfer and task information processing.

2.5 Discussion

Studies from neurophysiology, fMRI, and computational modeling emphasize the distributed nature of information processing in the brain [Eliasmith et al., 2012, Huth et al., 2012, Siegel et al., 2015]. However, fMRI studies often decode cognitive information from brain regions [Haxby et al., 2006] without considering how other brain regions might utilize that information [De-Wit et al., 2016]. In other words, current neuroscientific findings emphasize an experimenter-as-receiver framework (i.e., the experimenter decoding information in a brain region) rather than a cortex-as-receiver framework (i.e., brain regions decoding information transferred from other brain regions) [De-Wit et al., 2016]. The current emphasis on the experimenter-as-receiver framework clashes with the traditional understanding of information communication described by Shannon's Information Theory [Shannon, 1948], which provides a general theory of communication through the representation and transmission of information-bearing signals. Thus, understanding how cortical regions receive information from other regions bridges a crucial gap in understanding the nature of information processing in the brain. In light of recent findings relating resting-state fMRI to task-evoked cognitive activations [Cole et al., 2016a, Tavor et al., 2016], we hypothesized that resting-state FC describes the channels over which information can be communicated between cortical regions. Results strongly supported this hypothesis, suggesting that resting-state network topology describes the large-scale architecture of information communication in the human brain and demonstrating the relevance of resting-state network connectivity to cognitive information processing.

We developed a novel procedure to quantify information transfer between brain regions. The procedure requires an information-preserving mapping between a source region and a target region. In the neural network modeling literature, analogous mappings are typically estimated through machine learning techniques to approximate synaptic weight transformations between layers of a neural network (e.g., an artificial neural network model using backpropagation) [Yamins et al., 2014]. However, given that artificial neural networks are universal function estimators [Hartman et al., 1990] and would therefore fit any arbitrary mappings, we opted to take a more biologically principled approach that relied on FC estimation. Specifically, we used evidence that patterns of spontaneous activity can be used to successfully estimate the flow of task-related activity in both local and large-scale brain networks [Smith et al., 2006, Cole et al., 2016a, Timme et al., 2016] to obtain biophysically plausible, data-driven mappings between brain regions using resting-state fMRI. Thus, information transfer mapping unifies both biophysical and computational mechanisms into a single information-theoretic framework.

We used a computational model to validate the plausibility of this account of large-scale information transfer, finding that despite the slow dynamics of the blood-oxygen level dependent signal, resting-state FC with simulated fMRI accurately reflects the large-scale channels of information transfer. We then used empirical fMRI data to show that resting-state FC describes information-preserving mappings in cortex at two levels of organization: brain regions and functional networks. In other words, the connectivity-based mappings estimated via resting-state FC between a source and a target region preserved task information content (in the same representational geometry [Haxby et al., 2014, Diedrichsen and Kriegeskorte, 2017]). Note that the organization of activity patterns was necessarily distinct between brain regions (given their distinct sizes and shapes), such that accurately predicting activation patterns in a target region based on activity in a source region reflected accurate spatial transformation of information-carrying activity patterns between those brain regions. These findings suggest that resting-state FC estimates likely reflect the actual large-scale mappings that are implemented in the brain during task information transfer.

We used multiple regression (Figure 2.3c) rather than standard Pearson correlations (Figure 2.3b) to estimate resting-state FC for information transfer mapping. This decision was based on recent evidence that activations are better predicted when using multiple-regression FC as compared to Pearson-correlation FC [Cole et al., 2016a]. Importantly, multiple-regression FC strongly reduces the chance that estimated information transfers are indirect, since this method fits all regions/vertices simultaneously to identify unique shared variance between each pair of regions/vertices. Given that brain systems contain redundant neural signals [Tononi et al., 1999], however, multiple-regression FC estimates may be overly conservative. It will therefore be important for future research to validate appropriate regularization approaches to reduce the false negatives induced by multiple-regression FC. We expect that such a validated regularization approach would likely reveal that cognitive information transfers are even more widespread throughout the brain than reported here.

The evidence that fine-grained resting-state FC describes the information-preserving mappings between regions is important for advancing neuroscientific theory in a number of ways. First, the present results provide an empirically-validated theoretical account for how cognitive representations in different regions are likely mechanistically related to one another. Second, these results confirm the base assumption that decodable representations in a brain region are utilized by other regions through a biologically-plausible construct – information transfer via fine-grained patterns of activity flow. Third, these results expand the functional relevance of decades of resting-state FC findings [Biswal et al., 1995, Raichle, 2010], given that we demonstrated the ability to use resting-state FC to describe cognitively-meaningful fine-grained relationships between brain regions. Importantly, our modeling and empirical results showed that the topological organization of the intrinsic connectivity architecture described inter-region information-preserving mappings. Further supporting this conclusion, we verified via permutation testing that fine-grained FC topology (rather than, e.g., overall mean FC) was essential for the observed information transfer results.

Previous studies have focused on the role of task-evoked FC in shifting distributed task representations [Cole et al., 2013, Gratton et al., 2016]. We recently built on such findings to develop a flexible hub account of distributed task set reconfiguration via cognitive control networks [Cole et al., 2013, Cole et al., 2014b]. The present results advance these findings by describing a network mechanism involving resting-state FC topology (and cognitive control network hubs) in transferring task representations throughout cortex. Importantly, recent findings have demonstrated that task-evoked FC changes tend to be small relative to resting-state FC topology [Cole et al., 2014a, Krienen et al., 2014]. This suggests that the resting-state FC topology investigated here likely carries the bulk of the task-relevant information transfers, with task-evoked FC alterations to this topology contributing only small (but likely important) changes to this process.

The information transfer mapping approach involves estimating linear information transfer. Critically, however, neural information processing is thought to often depend on nonlinear transformations [Eliasmith, 2007], such as face-selective neurons in the ventral visual stream responding to whole faces but not facial components (e.g., eyes and ears) [Kanwisher et al., 1997, Tsao et al., 2006]. The present findings represent an important step toward understanding the network mechanisms underlying information transformations between brain regions, setting the stage for future research to identify the role of resting-state FC in nonlinear information transformations. This would go beyond the information transfer processes investigated here to better identify the role of resting-state FC in neural computation (not just communication).

In summary, we combined information decoding of brain activity patterns with resting-state FC to demonstrate how fine-grained intrinsic connectivity patterns relate to cognitive information transfer. Further, by estimating information transfer throughout cortex we found evidence that cognitive control networks play important roles in global transfer of cognitive task information. We expect that these findings will spur new investigations into the nature of distributed information processing throughout the brain, providing a deeper understanding of these fine-grained information channels estimated at rest and their contribution to task-relevant information transfers.

Chapter 3

Constructing neural network models from brain data reveals representational transformations underlying adaptive behavior

This chapter has been submitted for journal publication. The work is collaborative work with Guangyu Robert Yang, Patryk Laurent, Douglas H. Schultz, and Michael W. Cole

3.1 Abstract

The human ability to adaptively implement a wide variety of tasks is thought to emerge from the dynamic transformation of cognitive information. We hypothesized that these transformations are implemented via conjunctive representations in conjunction hubs – brain regions that selectively integrate sensory, cognitive, and motor representations. We used recent advances in mapping the representations of artificial neural networks to empirical brain data to construct a task-performing neural network model from empirical fMRI data during a cognitive control task. We verified the importance of conjunction hubs in cognitive computations by simulating neural activity flow over empirically-estimated neural network models. These simulations produced above-chance task performance (motor responses) by integrating sensory and task rule information in conjunction hubs. These findings reveal the role of conjunction hubs in supporting flexible cognitive computations, while demonstrating the feasibility of using empirically-estimated neural network models to gain insight into cognitive computations in the human

brain.

3.2 Introduction

The human brain exhibits remarkable cognitive flexibility. This cognitive flexibility enables humans to perform a wide variety of cognitive tasks, ranging from simple visual discrimination and motor control tasks, to highly complex context-dependent tasks. Key to this cognitive flexibility is the ability to use cognitive control, which involves goal-directed implementation of task rules to specify cognitive and motor responses to stimuli [Cole et al., 2017, Miller and Cohen, 2001]. Previous studies have investigated how task-relevant sensory, motor, and rule features are represented in the brain, finding that sensory stimulus features are represented in sensory cortices [Kanwisher, 2010, Nishimoto et al., 2011], motor action features are represented in motor cortices [Yokoi and Diedrichsen, 2018], while task rules are represented in prefrontal and other association cortices [Cole et al., 2015, Ito et al., 2017, Miller and Cohen, 2001, Reverberi et al., 2012, Rigotti et al., 2013]. However, exactly how and where in the brain different task representations mix to convert incoming stimuli to motor responses remains unclear [Kikumoto and Mayr, 2020]. In contrast, artificial neural network models (ANNs) can provide computationally rigorous accounts of how different task representations mix to implement cognitive computations [Mante et al., 2013, Yang et al., 2019]. Inspired by the formalization of ANNs, we constructed an empirically-estimated neural network (ENN) model from task fMRI data to provide insight into the representational transformations in the brain during a cognitive control task.

The flexible hub theory provides a network account of how large-scale cognitive control networks implement flexible cognition by updating task rule representations [Cocuzza et al., 2020, Cole et al., 2013]. This theory was built upon the guided activation theory – a seminal theory of the neural mechanisms underlying cognitive control – which posits that successful performance of a cognitive control task requires the selective mixing of task context with sensory stimulus information [Miller and Cohen, 2001]. The selective mixing of task context and sensory stimulus encoding activations would produce conjunctive (conditional association) representations that implement task rules on sensory stimuli. These conjunctive representations are thought to form through inter-area guided activations in "hidden units" located somewhere in association cortex, which we term conjunction hubs (Figure 3.1a). The outputs of conjunction hubs then activate motor representations to produce task-appropriate behavior. However, while the theory of interacting rule-guided neural activations between different task representations provides a framework to characterize representational transformations, it is unknown how the brain actually implements these computations.

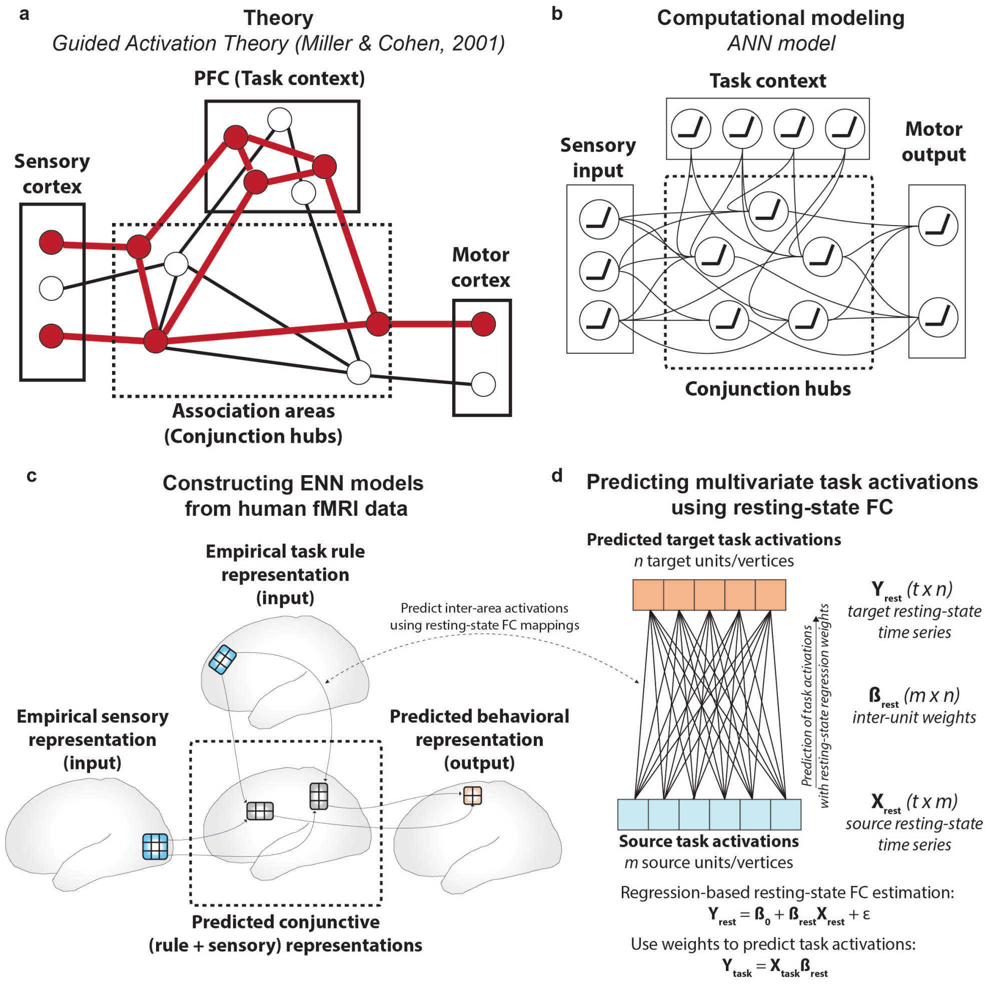

a
Theory
Guided Activation Theory (Miller & Cohen, 2001)
PFC (Task context)
Sensory cortex
Motor cortex
Association areas (Conjunction hubs)
b
Computational modeling
ANN model
Task context
Sensory input
Motor output
Conjunction hubs
c
Constructing ENN models from human fMRI data
Empirical task rule representation (input)
Predict inter-area activations using resting-state FC mappings
Empirical sensory representation (input)
Predicted behavioral representation (output)
Predicted conjunctive (rule + sensory) representations
d
Predicting multivariate task activations using resting-state FC
Predicted target task activations
n target units/vertices
Y_{rest} (t x n)
target resting-state time series
$ß_{rest}$ (m x n)
inter-unit weights
X_{rest} (t x m)
source resting-state time series
Prediction of task activations with resting-state regression weights
Source task activations
m source units/vertices
Regression-based resting-state FC estimation:
$Y_{rest} = ß_0 + ß_{rest} X_{rest} + \varepsilon$
Use weights to predict task activations:
$Y_{task} = X_{task} ß_{rest}$

Figure 3.1: Leveraging the guided activation theory to inspire ENN models of cognitive computation during task-based fMRI. a) A modified version of the guided activation theory, highlighting a potential key role for conjunction hubs. The guided activation theory posited that sensory cortices (left), which contain sensory stimulus-related representations, and prefrontal areas (top), which contain task context representations, integrate in association cortex to produce conjunctive representations through patterns of guided activations. Conjunctive representations are then guided to motor areas to generate motor response signals for task behavior. b) The guided activation theory can be instantiated more formally as a simple feedforward artificial neural network (ANN). This involves the task context and sensory stimuli representing the input layer, the association units representing a hidden layer, and the behavioral (motor) responses as the output layer. c) Testing the guided activation theory using task fMRI data collected in humans during context-dependent tasks. Using quantitative methods, we empirically test how different task representations (e.g., sensory stimuli and task context) form conjunctive representations to produce motor response representations using activity flow mapping [Cole et al., 2016a]. d) The guided activation theory can be empirically tested by projecting multivariate task activations between brain areas by estimating inter-area FC weight mappings obtained from resting-state fMRI data. Based on the activity flow principle, we estimated inter-vertex mappings using regression (see Methods) on resting-state fMRI data. This approach identifies a projection that maps across distinct spatial units in empirical data, similar to how inter-layer weights propagate activity across layers in a feedforward ANN.

We recently developed a method – activity flow mapping – that provides a framework for testing the guided activation theory with empirical brain data [Cole et al., 2016a]. Activity flow mapping involves several steps. First, a network model is derived from empirically-estimated connectivity weights. Second, empirical task activations (e.g., activity patterns from sensory regions) are used as inputs to simulate the activity flow (i.e., propagating activity) within the network. Finally, the predictions generated by simulated activity flow are tested against independent empirical brain data for model validation. Here we used activity flow mapping to test whether putative conjunction hubs could implement the context-dependent information transformations necessary to produce accurate behavioral (motor) representations in a 64-task cognitive paradigm.

We sought a principled approach to identify brain areas that form the conjunctive representations to produce flexible behavior. Recent studies have successfully used ANNs to probe the emergence of representational transformations in cognitive tasks [Mante et al., 2013, Song et al., 2016, Yang et al., 2019]. Importantly, the representational geometry of ANNs have often converged with the geometry of neural representations [Bashivan et al., 2019, Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014], reflecting the utility of ANNs in investigating task representations in the brain. Inspired by these previous studies, we constructed a simple feedforward ANN to investigate the representational transformations during the same task paradigm. This enabled us to characterize the representational transformations at each ANN layer that support successful task performance. These ANN representations provided a blueprint to search for analogous representations in empirical data: task context, sensory stimulus, conjunctive, and behavioral (motor) representations. The identification of these representations made it possible to empirically test the guided activation theory with activity flow mapping: Behavioral representations (in motor cortices) can be predicted through the formation of conjunctive representations through activity flow guided by task rule representations.

To summarize, we empirically tested the guided activation theory by constructing a task-performing ENN during a 64-task cognitive paradigm. This ENN was constructed directly from fMRI data, and illustrated the importance of conjunction hubs in facilitating representational transformations. This contrasts with many possible alternative hypotheses, such as the possibility that representations are transformed directly from task input areas (e.g., sensory systems) to motor cortices, bypassing association areas. We first trained a simple feedforward ANN to perform the 64-task paradigm to characterize the representational transformations required for successful task performance (Figure 3.1b). We next used the representational geometry in each ANN layer to map representationally similar brain areas in empirical data from subjects performing the same task. This allowed us to identify representational "layers" in the ENN, analogous to layers in a feedforward ANN. In contrast to ANNs, which use supervised learning to estimate connectivity weights, we show that representations in ENNs can be transformed via activity flow over functional connectivity (FC) weights estimated from resting-state fMRI (Figure 3.1d). This resulted in a task-performing, ENN model that transforms stimulus to response representations during a flexible cognitive control task. Critically, the transformations implemented by the ENN were carried out without classic optimization approaches such as gradient learning, demonstrating that the intrinsic architecture of the resting brain is suitable for implementing representational transformations. Together, these findings illustrate the computational relevance of functional network organization in supporting flexible cognitive computations in the human brain.

3.3 Methods

3.3.1 Participants

Data were collected from 106 human participants across two different sessions (a behavioral and an imaging session). Participants were recruited from the Rutgers University-Newark community and neighboring communities. Technical error during MRI acquisition resulted in removing six participants from the study. Four additional participants were removed from the study because they did not complete the behavior-only session. fMRI analysis was performed on the remaining 96 participants (54 females). All participants gave informed consent according to the protocol approved by the Rutgers University Institutional Review Board. The average age of the participants that were included for analysis was 22.06, with a standard deviation of 3.84. Additional details regarding this participant cohort have been previously reported [Schultz et al., 2019].

3.3.2 C-PRO task paradigm

We used the Concrete Permuted Operations (C-PRO) paradigm (Figure 3.2a) during fMRI acquisition, and used a computationally analogous task when training our ANN model. The details of this task are described below, and are adapted from a previous study [Ito et al., 2017].

The C-PRO paradigm is a modified version of the original PRO paradigm introduced in Cole et al., (2010) [Cole et al., 2010a]. Briefly, the C-PRO cognitive paradigm permutes specific task rules from three different rule domains (logical decision, sensory semantic, and motor response) to generate dozens of novel and unique task contexts. This creates a context-rich dataset in the task configuration domain akin in some ways to movies and other condition-rich datasets used to investigate visual and auditory domains [Nishimoto et al., 2011]. The primary modification of the C-PRO paradigm from the PRO paradigm was to use concrete, sensory (simultaneously presented visual and auditory) stimuli, as opposed to the abstract, linguistic stimuli in the original paradigm. Visual stimuli included either horizontally or vertically oriented bars with either blue or red coloring. Simultaneously presented auditory stimuli included continuous (constant) or non-continuous (non-constant, i.e., "beeping") tones presented at high (3000Hz) or low (300Hz) frequencies. Figure 3.2a demonstrates two example task-rule sets for "Task 1" and "Task 64". The paradigm was presented using E-Prime software version 2.0.10.353 [Schneider et al., 2002].

Each rule domain (logic, sensory, and motor) consisted of four specific rules, while each task context was a combination of one rule from each rule domain. A total of 64 unique task contexts (4 logic rules x 4 sensory rules x 4 motor rules) were possible, and each unique task set was presented twice for a total of 128 task miniblocks. Identical task sets were not presented in consecutive blocks. Each task miniblock included three trials, each consisting of two sequentially presented instances of simultaneous audiovisual stimuli. A task block began with a 3925 ms instruction screen (5 TRs), followed by a jittered delay ranging from 1570 ms to 6280 ms (2 - 8 TRs; randomly selected). Following the jittered delay, three trials were presented for 2355 ms (3 TRs), each with an inter-trial interval of 1570 ms (2 TRs). A second jittered delay followed the third trial, lasting 7850 ms to 12560 ms (10-16 TRs; randomly selected). A task block lasted a total of 28260 ms (36 TRs). Subjects were trained on four of the 64 task contexts for 30 minutes prior to the fMRI session. The four practiced rule sets were selected such that all 12 rules were equally practiced. There were 16 such groups of four task sets possible, and the task sets chosen to be practiced were counterbalanced across subjects. Subjects' mean performance across all trials performed in the scanner was 84% (median=86%) with a standard deviation of 9% (min=51%; max=96%). All subjects performed statistically above chance (25%).

3.3.3 ANN construction

We trained a simple feedforward ANN with a single hidden layer on a computationally analogous form of the C-PRO task. This enabled us to directly compare the representations of the ANN with the representations extracted from our empirical data.

To model the task context input layer, we designated an input unit for each task rule across all rule domains. Thus, we had 12 units in the task context layer. A specific task context (or rule set) would selectively activate three of the 12 units; one logic rule, one sensory rule, and one motor rule. Input activations were either 0 or 1, indicating an active or inactive state.

To model the stimulus input layer, we designated an input unit for a stimulus pair for each sensory dimension. To isolate visual color stimulus pairings, we designated input units for a red-red pairing, red-blue pairing, blue-red pairing, and blue-blue pairing. (Note that each unit represented a stimulus pair because the ANN had no temporal dynamics to present consecutive stimuli.) To isolate visual orientation stimulus pairings, we designated inputs for a vertical-vertical, vertical-horizontal, horizontal-vertical, and horizontal-horizontal stimulus pairing. To isolate auditory pitch stimulus pairings, we designated input units for high-high, high-low, low-high, and low-low frequency combinations. Finally, to isolate auditory continuity stimulus pairings (i.e., whether an auditory tone was constant or beeping), we designated input units for constant-constant, constant-beeping, beeping-constant, and beeping-beeping. Altogether, across the four sensory domains, we obtained 16 different sensory stimulus pair input units. For a given trial, four units would be activated to simulate a sensory stimulus combination (one unit per sensory domain). For example, a single trial might observe red-red (color), vertical-horizontal (orientation), low-high (pitch), constant-beeping (continuity) stimulus combination. Thus, to simulate an entire trial including both context and sensory stimuli, 7/28 possible input units would be activated.

We constructed our ANN with 1280 hidden units. This choice was due to recent counterintuitive evidence suggesting that the learning dynamics of extremely high-dimensional ANNs (i.e., those with many network parameters to tune) naturally protect against overfitting, supporting generalized solutions [Advani and Saxe, 2017]. Moreover, we found that across many initializations, the representational geometry identified in the ANN's hidden layer was highly replicable. Finally, our output layer contained four units, one for each motor response (corresponding to left middle, left index, right middle, right index finger presses).

The ANN transformed a 28-element input vector (representing a specific trial instance) into a 4-element response vector, and obeyed the equation

$$Y = f_s(X_{hidden} W_{out} + b) \tag{3.1}$$

where Y corresponds to the 4-element response vector, f_s is a sigmoid function, W_{out} corresponds to the connectivity weight matrix between the hidden and output layer, b is a bias term, and X_{hidden} is the activity vector of the hidden layer. X_{hidden} was obtained by the equation

$$X_{hidden} = f_r((X_{input} + I)W_{hidden} + b) \tag{3.2}$$

Where f_r is a rectified linear function (ReLU), W_{hidden} is the connectivity weight from the input to hidden layer, X_{input} corresponds to the input layer activations that contain trial information, and I is a 28-element noise vector sampled from a normal distribution with 0-mean and $\frac{1}{n}$-variance, where n refers to the number of hidden units.

3.3.4 ANN training

The ANN was trained by minimizing the mean squared error between the network's outputs and the correct target output. The mean squared error was computed using a mini-batch approach, where each mini-batch comprised of 192 distinct trials. (Each of the 64 unique task contexts were presented three times (with randomly sampled stimuli) in each mini-batch. Training was optimized using Adam, a variant of stochastic gradient descent [Kingma and Ba, 2017]. We used the default parameters in PyTorch (version 1.0.1), with a learning rate of 0.0001. Training was stopped when the last 1000 mini-batches achieved over 99.5% average accuracy on the task. This performance was achieved after roughly 10,000 mini-batches (or 1,920,000 trials). Weights and biases were initialized with a uniform distribution $U(-\sqrt{k}, \sqrt{k})$, where $k = \frac{1}{targets}$, where 'targets' represents the number of units in the next layer. Note that no cross-validation was performed (nor was it necessary), since we were only interested in representational geometry of the hidden layer. We also note that the representational geometry we observed in the hidden layer was robust to different initializations and hyperparameter choices.

3.3.5 ANN representational analysis

We extracted the representational geometry of the ANN's hidden layer using representational similarity analysis (RSA) [Kriegeskorte et al., 2008]. This was done to understand how task rule and stimulus information was transformed in the hidden layer. To extract the representational geometry of the hidden layer, we systematically activated a single unit in the input layer (which corresponded to either a task rule or sensory stimulus pair), and estimated the corresponding hidden layer activations (using trained connectivity weights). This resulted in a total of 28 (12 task rules and 16 sensory stimuli combinations) activation patterns. The representational similarity matrix (RSM) was obtained by computing the Pearson's correlation between the hidden layer activation patterns for all 28 conditions.

3.3.6 fMRI acquisition and preprocessing

The following fMRI acquisition details is taken from a previous study that used the identical protocol (and a subset of the data) [Ito et al., 2017].

Data were collected at the Rutgers University Brain Imaging Center (RUBIC). Whole-brain multiband echo-planar imaging (EPI) acquisitions were collected with a 32-channel head coil on a 3T Siemens Trio MRI scanner with TR=785 ms, TE=34.8 ms, flip angle=55°, Bandwidth 1924/Hz/Px, in-plane FoV read=208 mm, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. Whole-brain high-resolution T1-weighted and T2-weighted anatomical scans were also collected with 0.8 mm isotropic voxels. Spin echo field maps were collected in both the anterior to posterior direction and the posterior to anterior direction in accordance with the Human Connectome Project preprocessing pipeline [Glasser et al., 2016b]. A resting-state scan was collected for 14 minutes (1070 TRs), prior to the task scans. Eight task scans were subsequently collected, each spanning 7 minutes and 36 seconds (581 TRs). Each of the eight task runs (in addition to all other MRI data) were collected consecutively with short breaks in between (subjects did not leave the scanner).

3.3.7 fMRI acquisition and preprocessing

The following details are adapted from a previous study that used the same preprocessing scheme on a different data set [Ito et al., 2020a].

Resting-state and task-state fMRI data were minimally preprocessed using the publicly available Human Connectome Project minimal preprocessing pipeline version 3.5.0. This pipeline included anatomical reconstruction and segmentation, EPI reconstruction, segmentation, spatial normalization to standard template, intensity normalization, and motion correction. After minimal preprocessing, additional custom preprocessing was conducted on CIFTI 64k grayordinate standard space for vertex-wise analyses using a surface based atlas [Glasser et al., 2016b]. This included removal of the first five frames of each run, de-meaning and detrending the time series, and performing nuisance regression on the minimally preprocessed data [Ciric et al., 2017]. We removed motion parameters and physiological noise during nuisance regression. This included six motion parameters, their derivatives, and the quadratics of those parameters (24 motion regressors in total). We applied aCompCor on the physiological time series extracted from the white matter and ventricle voxels (5 components each extracted volumetrically) [Behzadi et al., 2007]. We additionally included the derivatives of each component time series, and the quadratics of the original and derivative time series (40 physiological noise regressors in total). This combination of motion and physiological noise regressors totaled 64 nuisance parameters, and is a variant of previously benchmarked nuisance regression models [Ciric et al., 2017].

3.3.8 fMRI task activation estimation

We performed a standard task GLM analysis on fMRI task data to estimate task-evoked activations from different conditions. Task GLMs were fit for each subject separately, but using the fully concatenated task data set (concatenated across 8 runs). We obtained regressors for each task rule (during the encoding period), each stimulus pair combination (during stimulus presentation), and each motor response (during button presses). For task rules, we obtained 12 regressors that were fit during the encoding period, which lasted 3925ms (5 TRs). For logic rules, we obtained regressors for “both”, “not both”, “either”, and “neither” rules. For sensory rules, we obtained regressors for “red”, “vertical”, “high”, and “constant” rules. For motor rules, we obtained regressors for “left middle”, “left index”, “right middle”, and “right index” rules. Note that a given encoding period contained overlapping regressors from each of the logic, sensory, and motor rule domains. However, the regressors were not collinear since specific rule instances were counterbalanced across all encoding blocks.

To obtain activations for sensory stimuli, we fit regressors for each stimulus pair. For example, for the color dimensions of a stimulus, we fit separate regressors for the presentation of red-red, red-blue, blue-red, and blue-blue stimulus pairs. This was done (rather than fitting regressors for just red or blue) due to the inability to temporally separate individual stimuli with fMRI’s low sampling rate. Thus, there were 16 stimulus regressors (four conditions for each stimulus dimension: color, orientation, pitch, continuity). Stimulus pairs were presented after a delay period, and lasted 2355ms (3 TRs). Note that a given stimulus presentation period contained overlapping regressors from four different conditions, one from each stimulus dimension. However, the stimulus regressors were not collinear since stimulus pairings were counterbalanced across all stimulus presentation periods (e.g., red-red stimuli were not exclusively presented with vertical-vertical stimuli).

Finally, to obtain activations for motor responses (or finger button presses), we fit regressors for each motor response. There were four regressors for motor responses, one for each finger (i.e., left middle, left index, right middle, right index fingers). Responses overlapped with the stimulus period, so we fit regressors for each button press during the 2355ms (3 TR) window during stimulus presentations. Note, however, that while response regressors overlapped with stimulus regressors, response regressors were not collinear with stimulus presentations. This is because a response is statistically independent from a stimulus pair, enabling the extraction of meaningful response activation patterns. A strong validation was that the finger representations could be reliably extracted according to the appropriate topographic organization in somatomotor cortex (Figure 3.4c).

(For a schematic of how task GLMs were performed, see Supplementary Figure B.3. For the task design matrix of an example subject, see Supplementary Figure B.4.)

3.3.9 fMRI decoding: Identifying sensory stimulus representations

Decoding analyses were performed to identify the brain areas that contained relevant task context and sensory stimulus representations. To identify the brain areas that contained relevant sensory stimulus representation, we performed four, four-way decoding analyses on each stimulus dimension: color (vision), orientation (vision), pitch (audition), constant (audition). For color stimulus information, we decoded activation patterns where the stimulus pairs were red-red, red-blue, blue-red, and blue-blue. For orientation stimulus information, we decoded activation patterns where the stimulus pairs were vertical-vertical, vertical-horizontal, horizontal-vertical, horizontal-horizontal. For pitch stimulus information, we decoded activation patterns where the stimulus pairs were high-high, high-low, low-high, and low-low. Finally, for constant (beeping) stimulus information, we decoded activation patterns where the stimulus pairs were constant-constant, constant-beeping, beeping-constant, beeping-beeping.

Decoding analyses were performed using the vertices within each parcel as decoding features. We limited decoding to visual network parcels for decoding visual stimulus features, and auditory network parcels for decoding auditory stimulus features. Visual parcels were defined as the VIS1 and VIS2 networks in Ji et al. (2019) [Ji et al., 2019], and auditory networks as the AUD network.

We performed a group-level decoding analysis, with a leave-8-subjects out cross-validation scheme. The choice of leaving 8 (out of 96) subjects out was due to recent studies suggesting that test sets should contain roughly 10% of the entire data set to yield stable predictive estimates of the test-set [Varoquaux, 2018]. Moreover, of the 88 subjects that remained in the train set pool (for each cross-validation fold), the training set was randomly sampled (with replacement, number of bootstrapped samples per fold = 88). We used a minimum-distance classifier (based on Pearson's correlation score), where a test set sample would be classified as the condition whose centroid is closest to in the multivariate activation pattern space [Mur et al., 2009]. P-values were calculated using a binomial test. Statistical significance was assessed using a False Discovery Rate (FDR) corrected threshold of p<0.05 across all 360 regions.

3.3.10 fMRI decoding: Identifying task rule representations

To identify the brain areas that contained task rule information, we performed a 12-way decoding analysis on the activation patterns for each of the 12 task rules. We used the same decoding and cross-validation scheme as above (for identifying sensory stimulus representations). However, we ran the decoding analyses on all 360 parcels, given previous evidence that task rule information is widely distributed across cortex [Ito et al., 2017]. P-values were calculated using a binomial test. Statistical significance was assessed using an FDR-corrected threshold of p<0.05 across all 360 regions.

3.3.11 fMRI activation analysis: Identifying motor response activations

To identify the brain areas/vertices that contained motor response information, we performed univariate analyses to identify the finger press activations in motor cortex. We performed two univariate activation contrasts, identifying index and middle finger activations on each hand. For each hand, we performed a two-sided group paired (by subject) t-test contrasting index versus middle finger representations. We constrained our analyses to include only vertices in the somatomotor network. Statistical significance was assessed using an FDR-corrected $p<0.05$ threshold, resulting in a set of vertices that were selective to button press representations in motor cortex (see Figure 3.4c).

We subsequently performed a decoding analysis on these sets of vertices (see Figure 3.7h). We decoded finger presses on each hand separately. Note that this decoding analysis is circular, since we had already determined that the selected vertices contained relevant information with regards to motor responses (via a univariate t-test). However, this provided an important benchmark to evaluate how well we could predict motor button responses using only context and stimulus activations (described below) relative to cross-validation of motor button response activations (i.e., a noise ceiling). Similar to the previous decoding analyses, we performed a leave-8-out cross validation scheme using a minimum-distance classifier, bootstrapping training samples for each fold. Moreover, because the decoding analysis was limited to a single ROI (as opposed to across many parcels/ROIs), we were able to compute confidence intervals (by bootstrapping cross-validation folds) and run nonparametric permutation tests since it was computationally tractable. We ran each cross-validation scheme 1000 times to generate confidence intervals. Null distributions were computed by randomly permuting labels 1000 times. P-values were computed by comparing the null distribution against the mean of the bootstrapped accuracy values.

3.3.12 fMRI representational similarity analysis: Identifying conjunction hubs

We compared the representational geometry of the ANN's hidden layer to the representational geometry of each brain parcel. This was possible because we extracted the exact same set of activation patterns (e.g., activations for task rules and sensory stimuli) in empirical data as our ANN model, enabling a direct comparison of representations. The representational geometry was estimated as the representational similarity matrix (RSM) of all task rules and sensory stimuli conditions.

We first estimated the empirical RSMs for every brain parcel separately in the Glasser et al. (2016) atlas. This was done by comparing the activation patterns of each of the 28 task conditions using the vertices within each parcel (12 task rule activations, 16 sensory stimulus activations). We then applied a Fisher's z-transform on both the empirical and ANN's RSMs, and then estimated the Spearman's rank correlation between the Fisher's z-transformed ANN and empirical RSMs (using the upper triangle values only). This procedure was performed on the RSM of every brain parcel, providing a similarity score between each brain parcel's and the ANN's representational geometry. For our main analysis, we selected the top 10 parcels with highest similarity to the ANN's hidden layer. However, we also performed additional analyses using the top 20, 30, and 40 parcels.

3.3.13 Inter-layer FC weight estimation

We estimated the inter-layer resting-state FC to identify weights between regions and layers in our empirical model. This was similar to a previously published approach which identified FC weights between pairs of brain regions [Ito et al., 2017]. This involved identifying FC weight mappings between the task rule input layer to the hidden layer, sensory stimulus input layer to the hidden layer, and the hidden layer to the motor output layer. For each inter-layer FC mapping, we estimated the vertex-to-vertex FC weights using principal components linear regression. We used principal components regression because most layers had more vertices (i.e., predictors) than samples in our resting-state data (resting-state fMRI data contained 1065 TRs). For all inter-layer FC estimations, we used principal components regression with 500 components. Specifically, inter-layer weights were estimated by fitting principal components to the regression equation

$$Y = \beta_0 + \sum_{i}^{500} X_i \beta_i + \epsilon \tag{3.3}$$

where Y corresponds to the t x n matrix with t time points and n vertices (i.e., the target vertices to be predicted), β_0 corresponds to a constant term, β_i corresponds to the 1 x n matrix reflecting the mapping from the component time series onto the n target vertices, X_i corresponds to the t x 1 component time series for component i, and ϵ corresponds to the error in the regression model. Note that X corresponds to the t x 500 component matrix obtained from a PCA on the resting-state data from the source layer. Also note that these loadings onto these 500 components are saved for later, when task activation patterns from a source layer are projected onto a target layer. The loadings project the original vertex-wise task activation patterns in the source layer onto a lower-dimensional space enabling faster computations. A similar approach was used in a previous study [Anzellotti et al., 2016]. FC weights were computed for each individual separately, but then averaged across subjects to obtain a group inter-layer weight FC matrix.

Note that in some cases, it was possible for overlap between the source and target vertices. (For example, some hidden area vertices may have coincided with the same vertices in the context layer.) In these cases, these overlapping vertices were excluded in the set of predictors (i.e., removed from the source layer) in the regression model.

3.3.14 Simulating sensorimotor transformations with multi-step activity flow mapping

We generated predictions of motor responses (in motor cortex) by assessing the correct motor response given a specific task context and sensory stimulus activation pattern (for additional details see Supplementary Figure B.1). For each subject, we simulated 960 trials. This consisted of the 64 unique task contexts paired with 15 randomly sampled stimulus combinations. For a trial, the task context input activation pattern was obtained by extracting the activation vector for the logic, sensory, and motor rule, and computing the mean rule vector (i.e., additive compositionality). The sensory stimulus input activation pattern was obtained by extracting the relevant sensory stimulus activation pattern. (Note that for a given trial, we only extracted the activation pattern for the sensory feature of interest. For example, if the rule was “Red”, only color activation patterns would be extracted, and all other stimulus activations would be set to 0.) Thus, the context and sensory stimulus activation patterns could be defined as

$$X_{context} = (R_{logic} + R_{sensory} + R_{motor})/3 \tag{3.4}$$

$$X_{stimulus} = X_{sensory} \tag{3.5}$$

where $X_{context}$ corresponds to the input activation pattern for task context, R_{logic} corresponds to extracted logic rule activation pattern (e.g., “Both”, “Not Both”, "Either", "Neither") obtained from the task GLM, $R_{sensory}$ corresponds to the extracted sensory rule activation pattern from the task GLM, R_{motor} corresponds to the extracted motor rule activation pattern from the task GLM, and $X_{stimulus}$ corresponds to the extracted sensory stimulus activation pattern that is indicated by the task context.

$X_{context}$ and $X_{stimulus}$ reflect the input activation patterns that were used to predict motor response conditions. Importantly, these input activation patterns were both spatially and representationally distinct from the motor response activations (in motor cortex). They were representationally distinct because these input activation patterns contained no information about the motor response required for a correct response. (In addition, we also used cross-validation to predict the motor response of a held-out subject, described below).

We used the inter-layer FC weight maps to project $X_{context}$ and $X_{stimulus}$ onto the hidden layer vertices. The projections (or predicted activation patterns on the hidden layer) were then thresholded to remove any negative BOLD predictions. This thresholding is equivalent to a rectified linear unit (ReLU), a commonly used nonlinear function in artificial neural networks [Yang and Wang, 2020]. Thus, the hidden layer was defined by

$$X_{hidden} = f_r(X_{context}W_{context2hidden} + X_{stimulus}W_{stimulus2hidden}) \tag{3.6}$$

where X_{hidden} corresponds to the predicted hidden layer activation pattern, f_r is a ReLU function (i.e., $f_r(x) = max(x, 0)$), $W_{context2hidden}$ corresponds to the inter-layer resting-state FC weights between the context and hidden layer, and $W_{stimulus2hidden}$ corresponds to the inter-layer resting-state FC weights between the stimulus and hidden layer. Note that all inter-layer FC weights (W_x) were computed using a principal component regression with 500 components. This requires that the vertex-wise activation space (e.g., $X_{context}$) be projected onto component space such that we define

$$W_x = U\hat{W}_{pc} \tag{3.7}$$

where U corresponds to an m x 500 matrix which maps the source layer's m vertices into component space, and $\hat{W}_{pc}$ is a 500 x n matrix that maps the components onto the target layer's n vertices. (Note that $\hat{W}_{pc}$ corresponds to the regression coefficients from equation 3.3, and that both U and $\hat{W}_{pc}$ are estimated from resting-state data.) Thus, W_x is an m x n transformation from a source layer's spatial pattern to a target layer's spatial pattern that is achieved through principal component regression on resting-state fMRI.

Finally, we generated a predicted motor output response by computing

$$X_{output} = X_{hidden}W_{hidden2output} \tag{3.8}$$

where X_{output} corresponds to the predicted motor response (in motor cortex), and $W_{hidden2output}$ corresponds to the inter-layer resting-state FC weights between the hidden and output layer. The full model computation can thus be formalized as

$$X_{output} = f_r(X_{context}W_{context2hidden} + X_{stimulus}W_{stimulus2hidden})W_{hidden2output} \tag{3.9}$$

X_{output} only yields a predicted activation pattern for the motor cortex for a given context and stimulus input activation pattern. To evaluate whether X_{output} could successfully predict the correct motor response for a given trial, we constructed an ideal 'task solver' that would indicate the correct motor response on a given trial (Supplementary Figure B.1). This solver would then be used to extract the correct motor response activation pattern, and compare the predicted motor cortex activation with the actual motor cortex activation pattern.

We simulated 960 trials per subject, randomly sampling context and stimulus input activation patterns. Because we sampled across the 64 task contexts equally (15 trials per context), the correct motor responses were equally balanced across 960 trials. Thus, of the 960 simulated trials for each subject, 240 trials yielded a left middle, left index, right middle, and right index response each. Each of these 240 predicted motor response patterns were subsequently averaged across trials such that we only obtained 4 predicted motor response patterns for each subject. Averaging was performed to remove any potential biases that a trial may have (e.g., a task context with the 'left middle' motor rule might be more biased towards a 'left middle' motor response).

3.3.15 Statistical and permutation testing of predicted motor response activations

The simulated empirical model generated predicted activations of motor activations in motor cortex. However, the predictions would only be interesting if they resembled actual motor response activations directly estimated during the response period via task GLM. In other words, without a ground truth reference to the actual motor response activation pattern, the predicted activation patterns would hold little meaning. The simulated empirical model generated four predicted activation patterns corresponding to predicted motor responses for each subject. We also had four actual activation patterns corresponding to motor responses that were extracted from the motor response period using a standard task GLM for each subject. To test whether the predicted activation patterns actually conformed to the actual motor response activation patterns, we trained a decoder on the predicted motor response activations and tested on the actual motor response activations of held-out subjects. We used the same cross-validation decoding scheme as before, with the exception that training was exclusively performed on predicted activation patterns of 88 subjects, while testing was exclusively performed on the actual activation patterns of 8 held-out subjects. Training a decoder on the predicted activations and decoding the actual activations made this analysis consistent with a prediction perspective – we could test if, in the absence of any motor task activation information, the ENN could predict actual motor response activation patterns that correspond to behavior. All other details (e.g., minimum-distance classifier, leave-8-subjects out cross-validation) remained the same.

Statistical significance was assessed using permutation tests. We permuted the labels of the predicted motor responses while testing on the actual motor responses. Null distributions are visualized in gray (Figure 3.7h). Statistical significance was assessed by comparing the mean of the bootstrapped predicted-to-actual accuracy scores, and comparing them against a non-parametric p-value that was estimated from the null distribution. Statistical significance was defined by a $p<0.05$ threshold.

3.4 Results

3.4.1 Training an ANN to perform the C-PRO cognitive task paradigm

The guided activation theory [Miller and Cohen, 2001] hypothesizes that sensory stimulus and task context information integrate in association cortex to form conjunctive (conditional association) representations (Figure 3.1a,c). We began by developing an ANN that formalizes the guided activation theory (Figure 3.1b), characterizing the computational properties of context-dependent sensorimotor transformations. Due to its comprehensive assessment of rule-guided behavior across 64 task contexts, we used the Concrete Permuted Rule Operations (C-PRO) task paradigm (Figure 3.2a). Briefly, the C-PRO paradigm is a highly context-dependent cognitive control task, with 12 distinct rules that span three rule domains (four rules per domain; logical gating, sensory gating, motor selection). These rules were permuted within rule domains to generate 64 unique task contexts, and up to 16384 unique trials possibilities (with various stimulus pairings; see Methods). The ANN included an input, hidden, and output layer (Figure 3.2b). The model was trained on all 64 C-PRO task contexts until the model achieved 99.5% accuracy (see Methods).

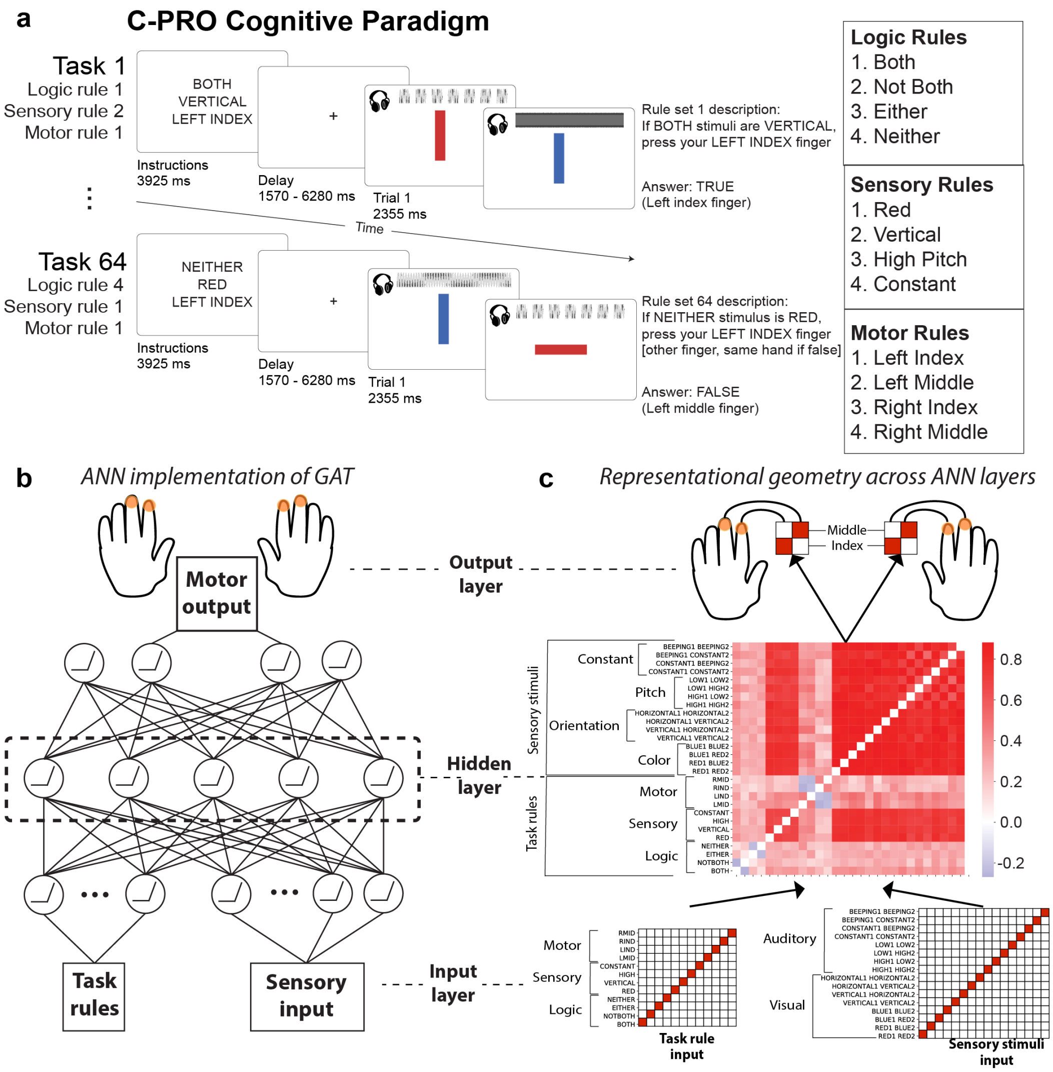


Figure 3.2: Training an ANN on a context-dependent task to identify representational transformations during cognitive computations. a) The Concrete Permuted Rule Operations (C-PRO) task paradigm [Ito et al., 2017]. For a given trial, subjects were presented with a task rule set (encoding), in which they were presented with three rules sampled from three different rule domains (i.e., logical gating, sensory gating, and motor selection domains). After a delay period, subjects applied the task rule set to two consecutively presented sensory stimuli (simultaneous audio-visual stimuli) and responded accordingly with button presses (index and middle fingers on either hand). We employed a miniblock design, in which for a given task rule set, three trials were presented separated by an inter-trial interval (1570ms). See Methods for additional details. b) We trained an ANN on the C-PRO paradigm. c) Training an ANN on the C-PRO task paradigm enabled us to identify the representational transformations that occur during a given trial, from stimulus to response. Transformations in representational content can be observed by the changes to the representational geometry (indicated by the representational similarity matrix) at each layer of the ANN [Kriegeskorte and Kievit, 2013]. Note that the diagonal matrices in the input and output layers serve to indicate that each feature (i.e., task rule, stimulus feature, or motor responses) was decodable. The RSA on the hidden layer yielded a 28 x 28 matrix consisting of 12 task rules (spanning all three rule domains) and the 16 stimulus pairings (spanning all sensory dimensions) (see Methods).

We were specifically interested in identifying the conjunctive representations that integrated stimulus and context information in the hidden layer, since these are the critical representations for selecting the correct motor responses [Kikumoto and Mayr, 2020]. To identify these conjunctive representations, we performed a representational similarity analysis (RSA) on the hidden layer of the ANN (Figure 3.2c) [Kriegeskorte et al., 2008]. The representational similarity matrix (RSM) of the hidden layer consisted of 28 task features: 12 task rules (which spanned the 3 rule domains), and 16 stimulus pairings (which spanned each sensory dimension). The hidden layer's RSM revealed the representational geometry that the ANN learned to perform the task correctly. (Note that we illustrated the RSMs at the input and output layers, though these RSMs merely indicate that input task rules/stimuli and output responses were decodable.) The analysis of the ANN at each layer provided a representational blueprint to identify similar representations in human fMRI data during the same task.

3.4.2 Identifying brain areas containing task-relevant information

Using RSA to map the representational geometry of the ANN to the fMRI data, we first identified the set of cortical areas that contained decodable sensory stimulus representations (Figure 3.3a). Because our stimuli were multimodal (audiovisual), this involved the identification of surface vertices that contained the relevant visual (color and orientation) and auditory (pitch and continuity) dimensions. We performed a four-way multivariate pattern analysis [Norman et al., 2006] (using a minimum-distance classifier [Mur et al., 2009]) to decode stimulus pairs for each of the four stimulus dimensions (e.g., red-red vs. red-blue vs. blue-red vs. blue-blue). Decoding analyses were performed within each brain parcel using the

Glasser et al. atlas [Glasser et al., 2016a], using vertices within each parcel as decoding features. For all decoding analyses, statistical thresholding was performed using a one-sided binomial test (greater than chance=25%), and corrected for multiple comparisons using an FDR-corrected p<0.05 threshold. We collectively defined the units in the ENN (i.e., vertices) that contained sensory stimulus information to be the set of all vertices within the parcels that contained decodable stimulus information (Figure 3.3f; Supplementary Tables 1-4).

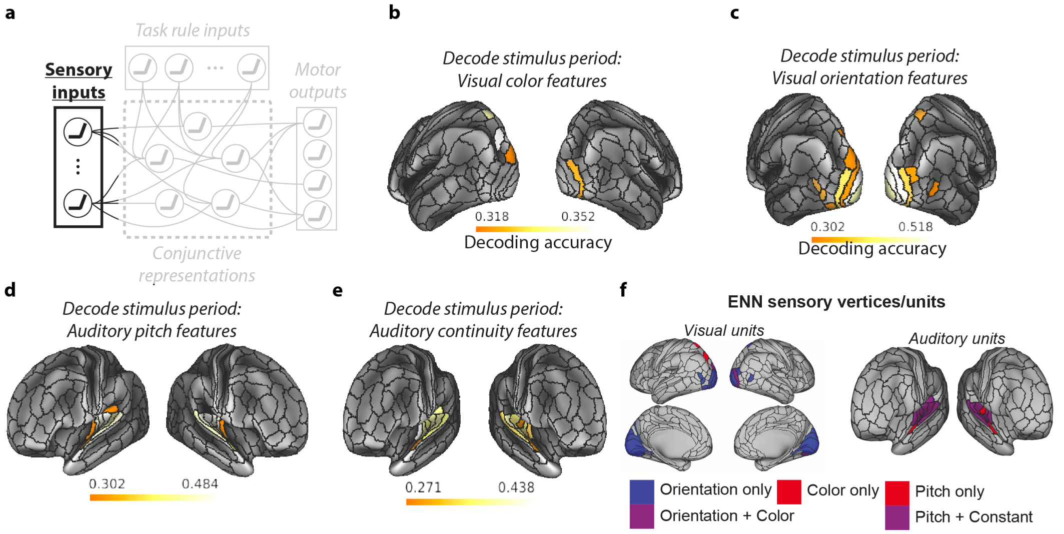

a
Task rule inputs
Sensory inputs
Motor outputs
Conjunctive representations
b
Decode stimulus period: Visual color features
0.318
0.352
Decoding accuracy
c
Decode stimulus period: Visual orientation features
0.302
0.518
Decoding accuracy
d
Decode stimulus period: Auditory pitch features
0.302
0.484
e
Decode stimulus period: Auditory continuity features
0.271
0.438
f
ENN sensory vertices/units
Visual units
Auditory units
🟦 Orientation only
🟥 Color only
🟪 Orientation + Color
🟥 Pitch only
🟪 Pitch + Constant

Figure 3.3: Identifying sensory stimulus input units (vertices) of the ENN using multivariate pattern classification analysis. a) We identified the sensory stimulus representations in empirical data using multivariate pattern decoding of stimulus activations. This corresponded to the sensory input component of the guided activation theory. To decode visual features (i.e., color and orientation stimulus features) we decoded the vertices within each parcel in the visual network using a recent functional network atlas [Ji et al., 2019]. To decode auditory features (i.e., pitch and continuity) we decoded the vertices within each parcel in the auditory network (see Methods). b) Decoding of color features using task activation estimates (from a task GLM) during the stimulus presentation period of the C-PRO task. Chance is 25%; cortical maps were thresholded using an FDR-corrected threshold of p<0.05. c) 4-way decoding of orientation features. d) 4-way decoding of auditory pitch features. e) 4-way decoding of auditory continuity features. f) The ENN sensory units, which were derived from a mask of the vertices that could successfully decode stimulus features.

Next, we performed a 12-way decoding analysis – isolated to the fMRI activation during the task instruction period – across all 12 task rules to identify the set of vertices that contained task rule information. Our previous study illustrated that rule representations are widely distributed across cortex [Ito et al., 2017], such that we tested for rule representations in every parcel in the Glasser et al. atlas (360 total parcels [Glasser et al., 2016a]). We again found that task rule representations were widely distributed across cortex (Figure 3.4b; FDR-corrected p<0.05 threshold; Supplementary Table 6). The set of vertices that survived statistical thresholding were included as "task rule" input units in the ENN (Figure 3.1c).

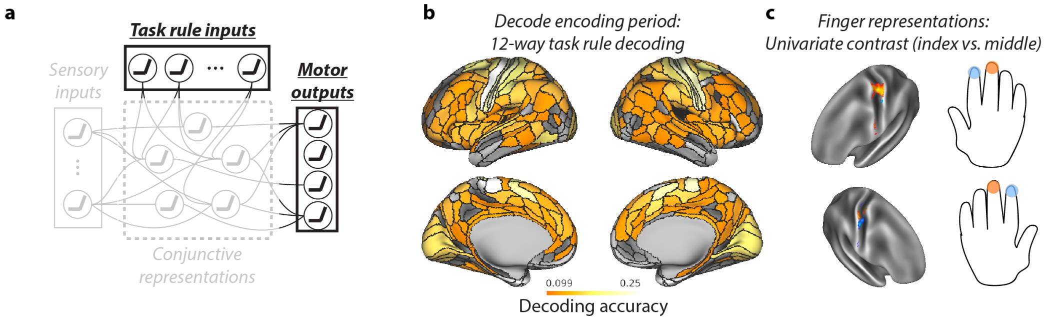


Figure 3.4: Identifying ENN units (i.e., fMRI vertices) containing relevant task rule (encoding) and motor response (behavior) representations. a) We identified the task rule inputs and motor output representations in empirical data using MVPA and univariate task activation contrasts. b) A 12-way decoding of each of the task rules (across the 3 rule domains) using task activations (estimated from a task GLM) during the encoding period of the C-PRO task. We applied this 12-way decoding to every parcel, given that task rule representations have been previously shown to be widely distributed across cortex [Ito et al., 2017]. Chance decoding was 8.33%; statistical maps were thresholded using an FDR-corrected p<0.05 threshold. c) To identify the motor/output representations, we performed a univariate contrast, contrasting the middle versus index finger response activations for each hand separately. Finger response activations were estimated during the response period, and univariate contrasts were performed on a vertex-wise basis using all vertices within the somatomotor network [Ji et al., 2019]. Contrast maps were statistically thresholded using an FDR-corrected p<0.05 threshold. The resulting finger representations matched the placement of finger representations in the well-established somatomotor homunculus in the human brain.

The C-PRO task paradigm required button presses (using index and middle fingers on either hand) to indicate task responses. Thus, to isolate finger representations in empirical neural data, we performed a univariate contrast of the vertex-wise response-evoked activation estimates during index and middle finger response windows (see Methods). For each hand, we performed a two-sided paired t-test (paired across subjects) for middle versus index finger responses in the somatomotor network [Ji et al., 2019]. Contrast maps were corrected for multiple comparisons (comparisons across vertices) using an FDR-corrected threshold of

$p<0.05$ (Figure 3.4c). Vertices that survived statistical thresholding were then selected for use as output units in the ENN (Figure 3.1c).

3.4.3 Identifying conjunction hubs

We next sought to identify conjunctive representations that could plausibly implement the transformation of inputs to outputs across the 64 task contexts (Figure 3.5a). Using the ANN trained on our experimental paradigm (C-PRO) (Figure 3.2c), we mapped the ANN's hidden layer (Figure 3.5b) to empirical fMRI data using RSA. The ANN's hidden layer necessarily contains the conjunction of task rule and sensory stimulus information, providing a blueprint to identify conjunction hubs – brain regions with analogous conjunctive representations in empirical fMRI data.

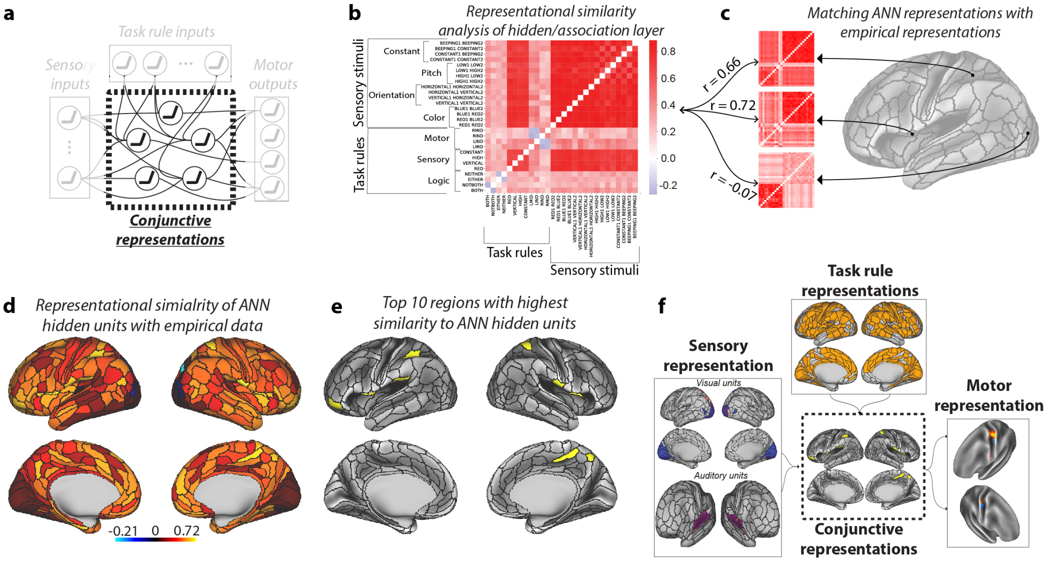


Figure 3.5: Identifying conjunction hubs: brain areas (vertices) that contain task-relevant conjunctions of sensory stimulus and task rule information. a) The guided activation theory hypothesized that there were a specific set of association (or hidden) areas that integrated sensory stimulus and task context information to select appropriate motor response representations. Computationally, this corresponded to the hidden layer in our ANN implementation (Figure 3.2b). b) We therefore used the representational similarity matrix (RSM) of the ANN's hidden layer as a blueprint to identify analogous conjunctive representations in empirical data. c) We constructed RSMs for each brain parcel (using the vertices within each parcel as features). We evaluated the correspondence between the representational geometry of the ANN's hidden layer and each brain parcel's representational geometry. Correspondence was assessed by taking the correlation of the upper triangle of the ANN and empirical RSMs. d) The representational similarity of ANN hidden units and each brain parcel. e) We showed the top 10 regions with highest similarity to the ANN hidden units. f) The full ENN architecture for the C-PRO task. We identified the vertices that contained task-relevant rule, sensory stimulus, conjunctive, and motor output representations.

To evaluate the similarity of the ANN's hidden layer representational geometry with each brain parcel, we computed the similarity (using Spearman's correlation) of the ANN's RSM with the brain parcel's RSM (Figure 3.5c). This resulted in a cortical map, which showed the representational similarity between each brain region and the ANN's hidden layer (Figure 3.5d). For our primary analysis, we selected the top 10 parcels with highest similarity to the ANN's hidden layer to represent the set of spatial units that contain conjunctive representations in the ENN (Figure 3.5e). The conjunction hubs were most strongly represented by the cingulo-opercular network, a network previously reported to be involved in task set maintenance (Supplementary Figure B.2; Supplementary Table 5)

[Power and Petersen, 2013]. However, we also performed ENN simulations using the top 20, 30, and 40 regions with highest similarity to the ANN hidden layer below.

3.4.4 Task-performing neural network simulations via empirical connectivity

The previous sections provided the groundwork for constructing an ENN model from empirical data. After estimating the connectivity weights between the surface vertices between ENN layers using resting-state fMRI (see Methods), we next sought to evaluate whether we could use this ENN to produce representational transformations sufficient for performing the C-PRO paradigm. This would demonstrate that the empirical input representations and the estimated connectivity patterns between ENN layers are sufficient to approximate the cognitive computations involved in task performance.

The primary goal was to predict the motor response pattern (i.e., behavior) yielding correct task performance. The only inputs to the model were a combination of activation patterns for a specific task context (rule combination) and sensory stimulus pair sampled from empirical data (Figure 3.6a). The outputs of the model were the predicted motor response activation pattern in motor cortex that should correspond to the correct button press (Figure 3.6c). High correspondence between the predicted and actual motor activation patterns would constitute an empirical demonstration of representational transformation in the brain, where task rule and sensory stimulus information is transformed into task-appropriate behavioral activation patterns.

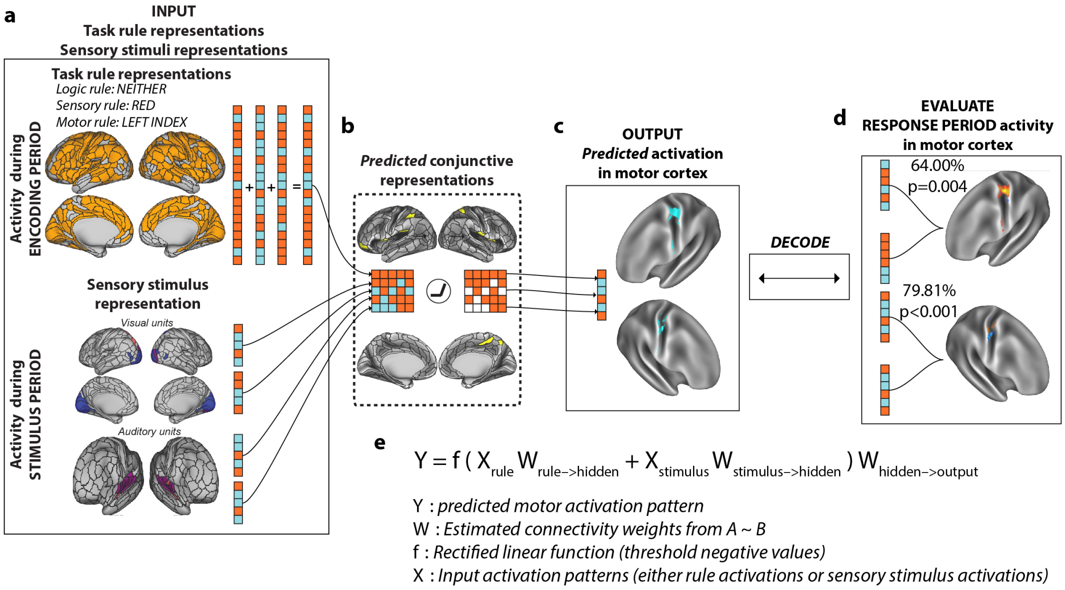

a
INPUT
Task rule representations
Sensory stimuli representations
Task rule representations
Logic rule: NEITHER
Sensory rule: RED
Motor rule: LEFT INDEX
Activity during ENCODING PERIOD
Sensory stimulus representation
Visual units
Auditory units
Activity during STIMULUS PERIOD
b
Predicted conjunctive representations
c
OUTPUT
Predicted activation in motor cortex
DECODE
d
EVALUATE
RESPONSE PERIOD activity in motor cortex
64.00%
p=0.004
79.81%
p<0.001
e
Y = f (X_rule W_rule->hidden + X_stimulus W_stimulus->hidden) W_hidden->output
Y : predicted motor activation pattern
W : Estimated connectivity weights from A ~ B
f : Rectified linear function (threshold negative values)
X : Input activation patterns (either rule activations or sensory stimulus activations)

Figure 3.6: Simulating context-dependent sensorimotor transformations with empirically-estimated task activations and inter-unit functional connectivity estimates. We constructed the ENN by identifying the vertices that contained task rule, sensory stimulus, and motor response representations (via decoding) and by estimating the resting-state FC weights between them. a) The input layer, consisting of vertices with decodable task rule information and sensory stimulus representations. b) Through activity flow mapping, input representations were mapped onto surface vertices in conjunction hubs. The activity flow-mapped vertices were passed through a nonlinearity, which removed any negative values. This threshold was chosen given the difficulty in interpreting predicted negative BOLD values. c) The predicted conjunctive representations were then activity flow-mapped onto the motor output vertices, generating a predicted motor activation pattern. d) These predicted motor activations were then tested against the actual motor response activations of other subjects using a leave-8-subject out cross validation scheme. A decoder was trained on the predicted motor response activations and tested on the actual motor response activations of the held-out cohort (see Methods and Supplementary Figure B.1 1). e) An equation summarizing the ENN model's computations.

Simulating activity flow in the ENN involved first extracting the task rule activation patterns (inputs) for a randomly generated task context (see Methods and Supplementary Figure B.1). Independently, we sampled sensory stimulus activation patterns for each stimulus dimension (color, orientation, pitch, continuity) (Figure 3.3). Then, using activity flow mapping with resting-state FC weights, we projected the activation patterns from the input vertices onto the conjunction hub vertices (Figure 3.6b). The predicted conjunction hub activation pattern was then passed through a simple rectified linear function, which removed any negative values (i.e., any values lower than resting-state baseline;

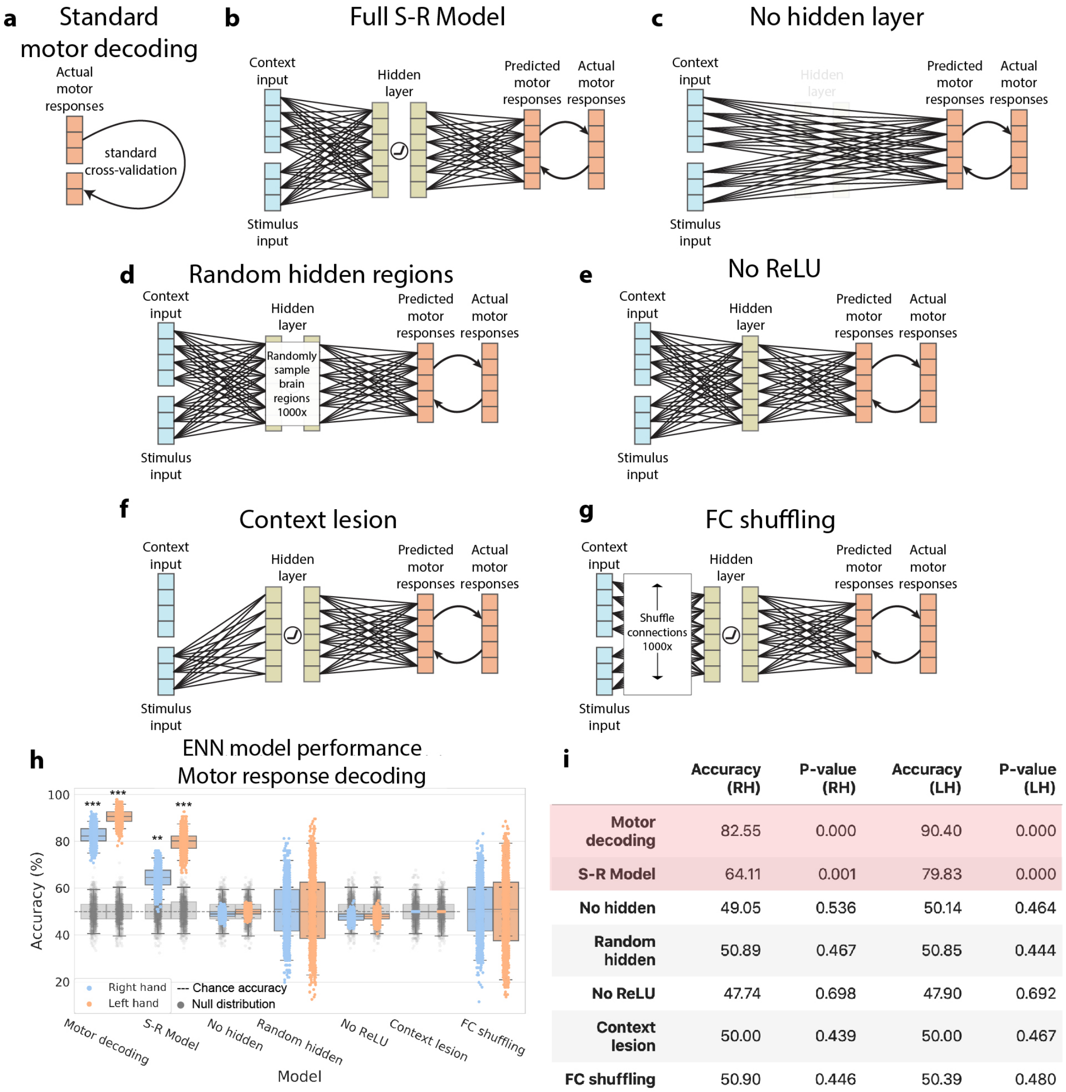
see Methods). Thresholded values were then projected onto the output layer vertices in motor cortex (Figure 3.6c), yielding a predicted response activation pattern. The sequence of computations performed to generate a predicted motor activation pattern (Figure 3.6a-c) is encapsulated by the equation in Figure 3.6e. Thus, predicted motor activation patterns can be generated by randomly sampling different task context and sensory stimuli combinations for each subject.

While the above procedure yielded a predicted activation pattern in the motor output layer, these predictions may not actually yield meaningful activation patterns. Thus, we evaluated whether the predicted motor activation patterns would accurately predict the actual motor response activation pattern extracted (via GLM) during the response period. Activity flow simulations generated predicted motor responses for each subject (Supplementary Figure B.1). This yielded four predicted motor response activations per subject, one for each behavioral response. Importantly, the predicted motor response activations were generated using only input task activations from the task encoding period and stimulus presentation period (Figure 3.6a). Independently, each subject also had four corresponding real motor response activations, which were estimated from the task GLM during the response period. Using a leave-8-subjects out cross-validation scheme, we trained a decoder on the four predicted motor responses and decoded the four actual motor responses (Figure 3.6c,d). Training a decoder on the predicted activations and decoding the actual activations (rather than vice versa) made this analysis more in line with a prediction perspective – we could test if, in the absence of any motor task activation information, the ENN could predict actual motor response activation patterns that correspond to behavior.

We note that this decoding analysis is highly non-trivial, given that the predicted motor responses are independent from the test set (actual motor responses) in three ways: 1) The predicted motor responses were generated from task rule and stimulus activation patterns, which (due to temporal separation in the task paradigm and counterbalancing) were statistically independent from the motor responses; 2) The motor response predictions were generated via activity flow mapping, and thus from a spatially independent set of vertices (see Methods); 3) The actual motor responses in the test set were sampled from independent subjects. By simulating neural network computations from stimulus and task context activations to predict motor response, we accurately decoded the correct finger response on each hand separately: decoding accuracy of right hand responses = 64.00%, non-parametric p=0.004; decoding accuracy of left hand responses = 79.81%, non-parametric p<0.001. These results demonstrate that task rule and sensory stimulus representations can be transformed into motor output representations by simulating multi-step neural network computations using activity flow mapping on empirical fMRI data.

3.4.5 The importance of the conjunctive representations

We next evaluated whether specific components of the ENN model were necessary to produce accurate stimulus-response transformations. We first sought to evaluate the role of the conjunction hubs (hidden layer) in model performance. This involved re-running the ENN with the conjunction hubs removed (Figure 3.7c), which required resting-state FC weights to be re-estimated between the input and motor output layer directly. We found that the removal of conjunction hubs severely impaired task performance to chance accuracy (RH accuracy=49.05%, p=0.54; LH accuracy=50.14%, p=0.46; Figure 3.7h,i). This illustrated the importance of conjunction hub computations in producing the conjunctive representations required to perform context-dependent stimulus-response mappings.



	Accuracy (RH)	P-value (RH)	Accuracy (LH)	P-value (LH)
Motor decoding	82.55	0.000	90.40	0.000
S-R Model	64.11	0.001	79.83	0.000
No hidden	49.05	0.536	50.14	0.464
Random hidden	50.89	0.467	50.85	0.444
No ReLU	47.74	0.698	47.90	0.692
Context lesion	50.00	0.439	50.00	0.467
FC shuffling	50.90	0.446	50.39	0.480

Figure 3.7: Systematic alteration of ENN model architecture verifies validity of "full S-R model" results. a) We first benchmarked the motor response decoding accuracy for each hand separately using a standard cross-validation scheme on motor activation patterns for each hand (tested across subjects). This standard motor decoding was done independently of modeling sensorimotor transformations. b) The full stimulus-response model, taking stimulus and context input activations to predicting motor response patterns in motor cortex. c) The ENN model after entirely removing the hidden layer. d) The ENN model, where we randomly sampled regions in the hidden layer (conjunction hubs) 1000 times and estimated task performance. e) The ENN model after removing the nonlinearity (ReLU) function in the hidden layer. f) The ENN model after lesioning connections from the task context input activations. g) The ENN model, where we shuffled the connectivity patterns from the stimulus and context layers 1000 times. h) Benchmarking the performances of all model architectures. Accuracy distributions were obtained by bootstrapping samples (leave-8-out cross-validation scheme and randomly sample within the training set). Boxplot whiskers reflect the 95% confidence interval. Grey distributions indicate the null distribution generated from permutation tests (permuting labels 1000 times). (*** = p<0.001; ** = p<0.01; * = p<0.01) i) Summary statistics of model performances. Reported accuracies are the mean of the bootstrapped samples.

We next replaced conjunction hubs with randomly sampled parcels in empirical data. This assessed the importance of using the ANN's hidden layer RSM to identify conjunction hubs (Figure 3.7d). We sampled random parcels 1000 times, recomputing the inter-layer vertex-wise FC each time. The distribution of randomly selected conjunction hubs did not yield task performance accuracies that were statistically different than chance for both hands (RH mean accuracy=50.87%, p=0.47; LH mean accuracy 50.85%, p=0.44; Figure 3.7h,i). However, the overall distribution had high variance, indicating that there may be other sets of conjunction hubs that would yield above-chance (if not better) task performance. However, compared to the conjunction hubs we identified by matching empirical brain representations with ANN representations, we found that the ANN-matched conjunction hubs performed better than 85.2% of all randomly selected conjunction hubs for RH responses, and greater than 97.7% of all randomly selected conjunction hubs for LH responses.

In addition, we evaluated whether the precise number of hidden regions was critical to task performance. We ran the full S-R model, but instead of using only the top 10 regions with highest similarity to the ANN's hidden layer's representations, we constructed ENN variants containing the top 20, 30, and 40 hidden regions. We found that we were able to reproduce correct task performance using 20 hidden regions (RH accuracy=63.90%, $p<0.001$; LH accuracy=76.95%, $p<0.001$). Using 30 hidden regions yielded reduced yet above-chance accuracies for RH responses, but not for LH responses (RH accuracy=59.83%, p=0.024; LH accuracy=43.54%, p=0.917). Inclusion of an additional 10 hidden regions (totaling 40 hidden regions) did not yield above-chance predictions of motor responses for either hand. These results suggest that conjunction hubs were better identified the greater the similarity of a region's representational geometry was to that of the ANN's hidden layer.

3.4.6 The importance of nonlinearities when combining rule and stimulus information

We next removed the thresholding of negative BOLD values (i.e., those lower than resting baseline) in the hidden layer. This is equivalent to removing the nonlinearity (ReLU) in an ANN (Figure 3.7e). We found that the removal of the ReLU function significantly impaired model performance (RH accuracy=47.74%, p=0.70; LH=47.90%, p=0.692; Figure 3.7h,i). This is likely due to the fact that context-dependent sensorimotor transformations require a nonlinear mapping between stimulus-response pairs, as predicted by prior computational studies [Cohen et al., 1990, Cohen et al., 2004].

3.4.7 Removing task context impairs task performance

We next sought to evaluate the importance of including task rule information in model performance. To remove context information, we lesioned all connections from the rule input layer to the hidden layer. This was achieved by setting all resting-state FC connections from the context input layer to 0 (Figure 3.7f). We ran the model on the exact same set of tasks, and found that as hypothesized, model performance was at chance without task context information (RH accuracy=50.00%, p=0.44; LH=50.00%, p=0.47; Figure 3.7h,i). This illustrated that the model implemented a representational transformation from task context and sensory stimulus representations to the correct motor responses.

3.4.8 The influence of specific functional network topography

We next evaluated whether the empirically-estimated connectivity topography was critical to successful task performance. This involved shuffling the connectivity weights within the context and stimulus input layers 1000 times (Figure 3.7g). While we hypothesized that the specific resting-state FC topography would be critical to task performance, we found that shuffling connectivity patterns yielded a very high variance distribution of task performance (Figure 3.7h). While the mean across all connectivity shuffles were approximately at chance for both hands (RH mean accuracy=50.90%, p=0.45; LH mean accuracy=50.39%, p=0.48), we found that there were some connectivity configurations that would significantly improve task performance, and other connectivity configurations that would yield significant below chance task performance. Notably, the FC topography that was estimated from resting-state fMRI (the full S-R model, without shuffling; Figure 3.7b) performed greater than 85.3% of all connectivity reconfigurations in RH responses, and greater than 97.7% of all connectivity reconfigurations for LH responses. This indicates that while there may exist better connectivity patterns for task performance, the weights derived from resting-state fMRI were sufficient to model correct task performance. We note that while the distribution of performance accuracies when shuffling FC weights and randomly sampling hidden layers are quite similar, these two permutation analyses control for fundamentally distinct properties of the ENN: specificity of FC topography versus specificity of conjunction hubs.

3.5 Discussion

Characterizing how different cognitive representations are transformed throughout the brain would fill a critical gap in understanding how the brain implements cognitive computations [Brette, 2019, De-Wit et al., 2016, Ito et al., 2020b]. To address this gap, we built a task-performing ENN from empirical data to characterize representational transformations during a cognitive control task. First, we performed representational similarity analysis on an ANN trained to perform an analogous task. Second, we used the representations identified in ANNs to find analogous representations in empirical data during the same task. Importantly, this enabled us to characterize how rule encoding and stimulus representations were selectively integrated to produce conjunctive representations. Finally, using activity flow mapping, we found that incoming sensory and task rule representations were transformed via conjunction hubs to produce above-chance behavioral predictions of outgoing motor responses. These findings suggest that flexible cognitive control is implemented by guided activations, as originally suggested by the guided activation theory.

The present results build on prior findings emphasizing the role of cognitive control networks (CCNs) in highly flexible cognition [Cole et al., 2017, Dosenbach et al., 2007, Power and Petersen, 2013, Waskom and Kiani, 2018]. The present results are largely consistent with previous accounts, showing that the task rule layer and conjunction hubs are most strongly affiliated with CCNs (e.g., cingulo-opercular and frontoparietal networks) (Supplementary Figure B.2) [Dosenbach et al., 2007, Power and Petersen, 2013]. (However, we note that other functional networks also represented task rules, though to a lesser extent.) Several studies of rapid instructed task learning found that CCNs represent rules compositionally in activity [Cole et al., 2015, Reverberi et al., 2012, Waskom et al., 2014] and FC [Cocuzza et al., 2020, Cole et al., 2013] patterns, which are considered essential for flexible reuse of task components [Cole et al., 2013, Reverberi et al., 2012, Yang et al., 2019]. The present results also demonstrate that the CCN and other networks use compositional rule representations, since the ENN rule inputs contained three rules whose fMRI activity patterns were simply added compositionally to create the full task context. Critically, we found that these compositional codes were not enough to enable flexible task performance – rather, conjunctive representations were required to interact non-linearly with these compositional representations. Moreover, our results showed that without conjunctive representations producing conditional interactions (e.g., through conjunction hub lesioning), the task performance of the ENN was substantially impaired. It will be important for future research to determine the exact relationship between compositional and conjunctive representations in implementing flexible cognitive programs.

The ENN characterized the representational transformations required to transform task input activations to output activations (in motor cortex) directly from data. Model parameters, such as unit identification and inter-unit connectivity estimation, were estimated *without optimizing for task performance*. This contrasts with mainstream machine learning techniques that iteratively train ANNs that directly optimize for behavior [Song et al., 2016, Yamins et al., 2014, Yang and Wang, 2020]. Our approach enabled the construction of functioning ENNs with above-chance task performance without optimizing for behavior; instead, we were able to derive parameters from empirical neural data alone. These results suggest that the human brain's intrinsic network architecture, as estimated with human fMRI data, is informative regarding the design of task-performing functioning models of cognitive computation.

We showed that the specific FC topography could predict inter-area transformations. In contrast, shuffling these specific inter-area FC topographies yielded ENNs with highly variable task performances, suggesting the computational utility of the empirically-estimated FC patterns. Previous work has illustrated that the functional network architecture of the brain emerges from a structural backbone [Deco et al., 2013a, Demirtaş et al., 2019, Wang et al., 2019, Tschopp et al., 2018, Hagmann et al., 2008]. Building on this work, we recently proposed that the functional network architecture of the brain can be used to build network coding models – models of brain function that describe information encoding and decoding processes constrained by empirically-estimated connectivity [Ito et al., 2020b]. Related proposals have also been suggested in the electron microscopy connectomics literature, suggesting that structural wiring diagrams of the brain (e.g., in drosophila) can inform functional models of biological systems (e.g., the drosophila's visual system) [Litwin-Kumar and Turaga, 2019, Tschopp et al., 2018]. Consistent with these proposals, our findings establish that the intrinsic functional network architecture in humans provides a meaningful foundation from which to implement cognitive computations.

Despite strong evidence that the estimated functional network model can perform tasks, there are several theoretical and methodological limitations. First, though we perform numerous control analyses by either lesioning or altering the ENN architecture (Figure 3.7), the space of alternative possible models that can potentially achieve similar (if not better) task performances is large. For example, here we assumed only a single hidden layer (one layer of 'conjunction hubs'). However, it is possible – if not probable – that such transformations actually involve a large sequence of transformations, similar to how the ventral visual stream transforms visual input into object codes, from V1 to inferior temporal cortex [Khaligh-Razavi and Kriegeskorte, 2014, Yamins et al., 2014]. It is therefore likely that the identification of conjunction hubs is likely dependent on both specific task demands and the targeted level of analysis (e.g., neuronal circuits versus large-scale functional networks). Here we opted for the simplest possible network model that involved conjunction hubs at the level of large-scale functional networks. Starting from this simple model allowed us to reduce potential extraneous assumptions and model complexity (such as modeling the extraction of stimulus features from early visual areas) which likely would have been necessary in more complex and detailed models. However, the current findings provide a strong foundation for future studies to unpack the mechanisms of finer-grained computations important for adaptive behavior.

Another assumption in the ENN was that activations were guided by additive connectivity weights. Additive connectivity weights assume inter-area predicted activations are the sum of source activations weighted by connections. One potential alternative (among others) would have been multiplicative guided activations; weighted activations that are multiplied (rather than summed) from incoming areas, which has been previously proposed as a potential alternative to designing ANNs [Wu et al., 2016]. However, several recent studies have suggested that inter-area activations are predicted via additive connectivity weights in both human fMRI [Cole et al., 2016a, Ito et al., 2017], the primate visual system [Bashivan et al., 2019], and the drosophila's visual system [Tschopp et al., 2018]. Nevertheless, it will be important for future work to systematically test alternative network architectures and dynamics in producing functional ENN models.

Finally, another limitation is that we constructed an ENN model that did not model realistic dynamics. Typical experimental paradigms include separate intervals for encoding, delay, stimulus, and response periods, since cognitive processing occurs over time. Here, we did not explicitly model temporal dynamics based on the empirical data when simulating the ENN. (However, we note that activation estimates for task encoding and trials were obtained from temporally distinct intervals.) Nevertheless, though it is likely that temporal dynamics (with recurrent feedback) likely play a role in shaping cognitive computations, we illustrate here that simple dynamics (i.e., rules + sensory inputs → conjunction hubs → motor outputs) involving the interplay of static activations are sufficient to model representational transformations. It will be important for future studies to construct task-performing brain models that can simulate temporal and recurrent dynamics constrained by empirical data, as this can provide a more detailed computational account of the representational transformations that contribute to behavioral variability.

In conclusion we constructed an ENN model capable of performing adaptive cognitive control tasks. This model provides strong evidence for the well-known guided activation theory by providing a computational implementation of the theory that is directly estimated from empirical data. We first identified the relevant brain representations associated with different task features. We then used an ANN to identify conjunction hubs that were critical to the selective integration of task input information for motor response selection. Finally, by estimating FC patterns from resting-state fMRI data, we parameterized a network model to generate predictive stimulus-to-response transformations using activity flow mapping. We expect that these findings will drive new investigations into characterizing the neural implementation of cognitive computations, providing dual insight into how the brain implements cognitive processes and how such knowledge can inform the design of ANN architectures.

Chapter 4
Modeling context-dependent sensorimotor transformations in multi-unit spiking activity

This chapter is in progress, and is a collaborative project with Scott L. Brincat, Markus Siegel, Earl K. Miller, and Michael W. Cole

4.1 Abstract

During flexible sensorimotor tasks, the brain transforms environmental and sensory information into motor actions. While the brain areas responsible for representing sensory and motor information are well characterized, how sensory information is converted to behavioral signals during context-dependent tasks is not well known. In this chapter, we provide neurophysiological evidence for how sensory and context information integrate to produce behavioral signals using multisite, multi-unit spiking activity obtained from non-human primates (NHP). By measuring the functional connectivity (FC) between multi-units using task-free spiking activity, we build a network model that describes how spiking activity in different cortical areas map onto each other to predict trial-to-trial motor activity. This provides a network explanation at the level of multi-units of how task representations are used, manipulated, and transformed in the brain to produce behavior during task execution.

4.2 Introduction

The brain extracts and transforms complex information into meaningful signals important for behavior. Recent advances have begun to characterize the computational and neural bases of these behaviorally relevant transformations. For example, in visual object recognition, inputs from the visual field are relayed through a sequence of computations from the retina, lateral geniculate nucleus, and down through the ventral visual stream to incrementally transform visual features into meaningful object representations in inferior temporal (IT) cortex [Yamins et al., 2014, Castelo-Branco et al., 1998]. In the motor cortex, noise robust, “untangled” population motor signals facilitate easy read out of motor intentions in downstream peripheral areas [Russo et al., 2018]. However, exactly how high-level sensory information is converted to motor cortex signals through intermediate computations during flexible, context-dependent tasks remains unclear.

Previous work has shown that during flexible behaviors, there is a transient bottom-up sweep of information followed by top-down flow of sustained task information [Siegel et al., 2015]. However, that study focused on characterizing the temporal dynamics of task information within each brain area, rather than addressing exactly how information in one area was related to information in another area. More specifically, understanding the temporal dynamics within each region provides only a partial understanding regarding how sensory representations integrate with context representations to guide motor decisions. In this chapter, we provide an empirically-derived network model that models sensorimotor transformations to predict neural spiking activity directly related to task behavior.

In Chapter 3, we demonstrated that an empirically-estimated neural network (ENN) could be extracted from whole-brain human fMRI data. Importantly, this

ENN modeled the integration of sensory and context representations using input activations from sensory and context representing brain areas. This transformation was modeled and parameterized using voxel-to-voxel FC weights obtained from resting-state data, demonstrating the computational relevance of intrinsic FC in supporting representational transformations. However, that study relied on extracted fMRI blood oxygenated level-dependent (BOLD) signals. While the BOLD signal has been shown to be correlated with local field potentials and spiking activity [Ma et al., 2016], it is still an indirect measure of neural activity.

In previous work, we illustrated that we could accurately predict the spread of task-evoked activity using activity flow modeling over empirically-estimated functional connections [Cole et al., 2016a, Ito et al., 2017]. Importantly, activity flow modeling tests the relationship between intrinsic functional network organization and local task-evoked activations, assessing whether local activations can be predicted from distributed network activity. Here we extend this notion of activity flow from whole-brain fMRI modeling to multi-unit spiking activity, the neural activity thought to be most mechanistically relevant to behavior. Using an analogous approach to Chapter 3, we build an ENN model to predict neural transformations using multi-site and multi-unit spiking activity. This addresses whether the activity flow framework previously developed for fMRI can be used to estimate spiking activity flow. We demonstrate that neural spiking activity in a target unit can be predicted using spiking activity flow over inter-unit FC estimates. Furthermore, we show that spiking activity flow predictions can accurately model sensorimotor transformations during flexible sensorimotor tasks.

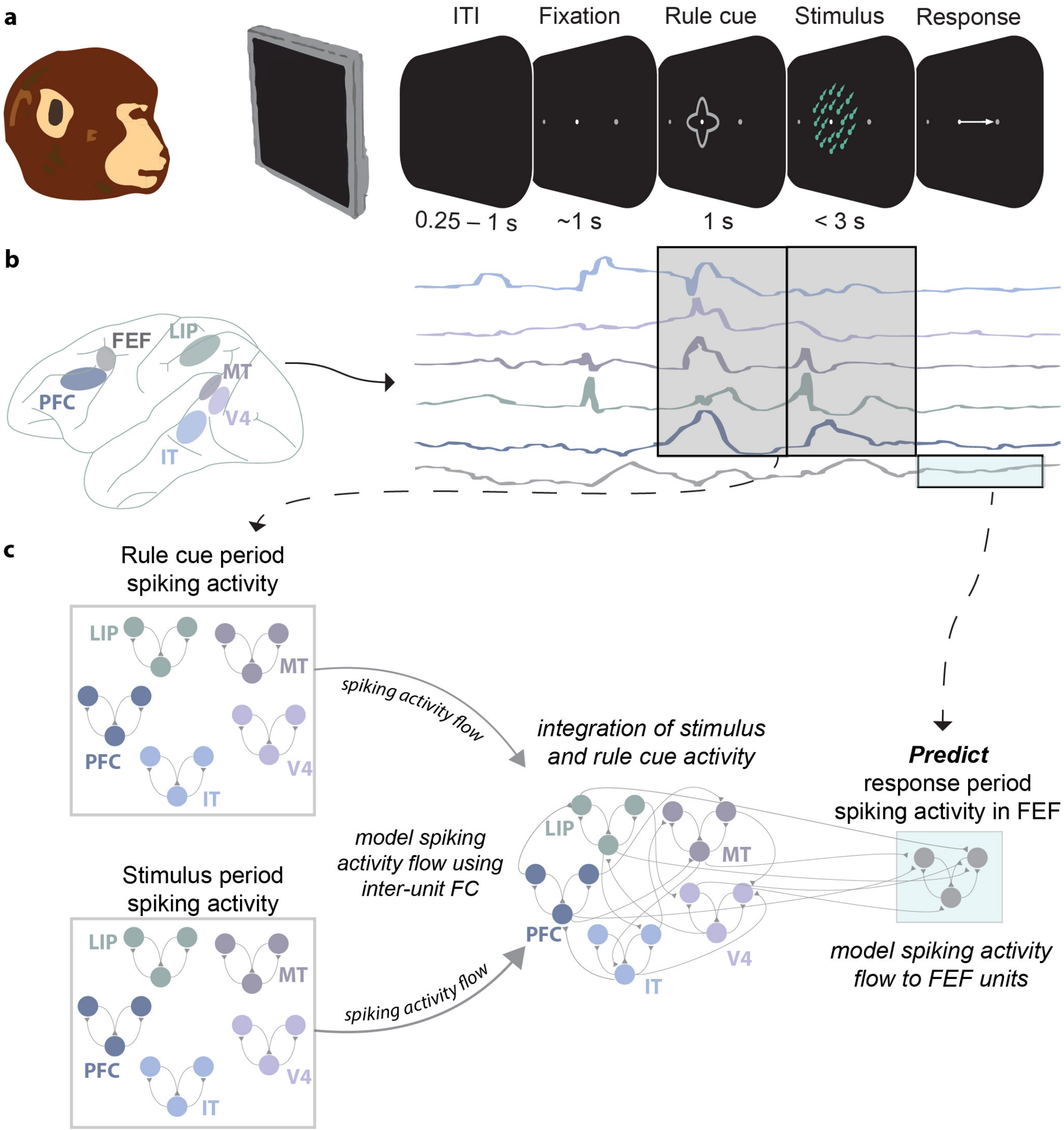

a
ITI
Fixation
Rule cue
Stimulus
Response
0.25 – 1 s
~1 s
1 s
< 3 s
b
FEF
LIP
MT
PFC
V4
IT
c
Rule cue period spiking activity
LIP
MT
PFC
V4
IT
spiking activity flow
integration of stimulus and rule cue activity
model spiking activity flow using inter-unit FC
Stimulus period spiking activity
spiking activity flow
LIP
MT
PFC
V4
IT
Predict response period spiking activity in FEF
model spiking activity flow to FEF units

Figure 4.1: Data, task paradigm, and methodological approach. a) Task paradigm. Two NHPs were trained to perform a motion-color categorization task [Siegel et al., 2015]. Depending on the task context (i.e.,, attend to color or motion), subjects were asked to categorize green/red colors or up/down motion of dot stimuli. NHPs responded with a saccade to either the left/right direction. b) Multi-unit spiking activity was extracted from six different cortical areas: V4, IT, MT, LIP, PFC, and FEF. c) We predicted FEF spiking activity during the response period using activity from other cortical areas during the task rule and stimulus periods. Note that the response period was a non-overlapping interval that occurs *after* the rule and stimulus intervals. Spiking activity flow was implemented by first estimating the FC between pairs of multi-units, and then by calculating the matrix multiplication of the firing rate vector with the inter-unit FC matrix (see Methods).

We report that during flexible behavior, behavioral signals in motor areas can be predicted as a sequence of network computations that takes task rule and stimulus-related activity as inputs to an ENN. We used a previously published data set that obtained multi-unit activity from six cortical areas during a motion-color categorization task (visual area 4 (V4); medial temporal area (MT); inferior temporal cortex (IT); lateral intraparietal cortex (LIP); prefrontal cortex (PFC); frontal eye fields (FEF)) [Siegel et al., 2015]. By estimating inter-unit FC, we construct a network model to estimate the spiking activity flow between task context and stimulus-representing areas to motor output areas (FEF). These results provide an explanation for how flexible sensorimotor transformations are implemented through network-based computations, while also illustrating the viability of translating network modeling approaches developed in the fMRI literature to electrophysiology data.

4.3 Methods

4.3.1 Spiking data: Data collection

Details regarding the data set have been previously reported in [Siegel et al., 2015, Brincat et al., 2018]. Additional details were reported in [Ito et al., 2020a] (Chapter 5), and were reported as follows. Data was collected in vivo from two (one female) behaving adult rhesus macaques (Macaca mulatta) across 55 sessions. Data from six distinct cortical regions were recorded simultaneously from acutely inserted electrodes. Cortical regions included: MT, V4, PIT, LIP, FEF, and LPFC (Figure 5.1a). Multi-unit spikes from each region were sorted offline. For each trial, spikes were sorted for a 5s period, beginning 2.5s prior to stimulus onset, and until 3.5s after stimulus onset. Further details regarding electrophysiological data collection can be found here: `http://www.sciencemag.org/content/348/6241/1352/suppl/DC1` and here: `http://www.pnas.org/content/pnas/suppl/2018/07/09/1717075115.DCSupplemental/pnas.1717075115.sapp.pdf`

All statistical analyses in the article were performed on two monkeys.

4.3.2 Baseline decoding

We first ran a baseline decoding to assess how much decodable information was contained in each of the cortical areas (Figure 4.2). This involved decoding task rule, sensory stimuli (color and motion features separately), and response activity decoding. Decoding analyses were performed using the mean firing rates (averaged across time, within trial) during a task condition from each unit. Decoding was performed for each cortical area separately (using each unit as a decoding feature), and was performed within session across trials. Some regions contained more sessions than others, since not all six regions were recorded from for every session.

For task rule decoding, we obtained the spike rate for every unit within a cortical area. The mean spike rate was calculated as the average of spikes from cue onset to offset, and a spike rate was obtained for every trial. (Note that the cue interval occurs prior to the stimulus and response intervals.) Each trial was labeled as either a cue associated with the "color" or "direction" rule. Note that there were four cues in total, two corresponding to "color" and two corresponding to "direction". We used a stratified 10-fold cross-validation scheme, ensuring that the number of labels were balanced across conditions within each training fold. We used a linear decoder (logistic regression with an L_2 regularization parameter $C = 1.0$). For each session, we obtained an average decoding accuracy (the average across all trial-wise predictions). Statistical significance was assessed using a t-test (chance = 50%) using the averaged values from each session ($p < 0.05$). P-values were corrected for multiple comparisons (across cortical areas) using False Discovery Rate [Genovese and Wasserman, 2002].

There were two sensory features of interest: color and direction. Thus, we performed two, two-way classifications, decoding color (red versus green) and direction (up versus down). The mean spike rate was calculated as the average of spikes during the stimulus period interval for each trial. While sensory stimuli were sampled from a 2-dimensional space comprising of 7 distinct color/color coherence values and 7 distinct direction/direction coherence values, 6 of 7 coherence values could be grouped as either red/green or up/down. (In each color/direction domain, there existed an ambiguous stimulus; trials which contained ambiguous stimuli were excluded from this classification.) Color and direction decoding were both implemented using the same decoding scheme as above (stratified 10-fold using a logistic decoder).

Motor response classifications decoded either left or right saccades. We computed the response period spiking activity as the mean number of spikes from the saccade onset to offset. Response period activity was temporally distinct from both stimulus and rule period activity, since the response cue emerged after stimulus offset. Trials were labeled as left or right movements based on behavior only, and independent of task correctness. Response decoding was implemented using the same decoding scheme as above.

As a control, we also assessed whether response information could be decoded during the cue and stimulus interval (Figure 4.4). This involved assigning the motor response label associated with a given trial to either cue or stimulus period activity. We note that while cue period activity should contain no information about the response period, it was conceivable that some response information was formed during the stimulus period. Decoding analyses were implemented using the same decoding schemes as above.

4.3.3 Inter-unit FC estimation

We estimated inter-unit FC to identify weights between units. This was similar to Chapter 3, but using individual units rather than individual voxels. However, unlike in the previous chapter, our NHP data set did not have a true resting state. Thus, we used the mean spike rate during the inter-trial interval (ITI), which preceded the fixation that indicated trial onset. (The ITI was stimulus-free.) FC was computed for each session separately, since recorded units were not necessarily the same across sessions.

Inter-unit FC was estimated using cross-validated ridge regression (using Python's scikit-learn RidgeCV function). We included all ITIs within a session. We estimated two sets of weights: 1) inter-unit FC for PFC, LIP, MT, V4, IT; 2) inter-unit FC between units in PFC, LIP, MT, V4, IT to FEF (see Figure 4.3b). Specifically, we estimated the weights for unit x_i (where unit x_i belongs in either PFC, LIP, MT, V4, or IT), by fitting the linear model (via ridge regression)

$$x_i = \beta_0 + \sum_{j \in PFC} \beta_j x_j + \sum_{j \in LIP} \beta_j x_j + \sum_{j \in V4} \beta_j x_j + \sum_{j \in MT} \beta_j x_j + \sum_{j \in IT} \beta_j x_j + \epsilon \quad (4.1)$$

where $i \neq j$, but $i \in \{PFC, LIP, V4, MT, IT\}$.

To identify FC weight mappings to the output area (FEF), we estimated the weights for unit $x_{i \in FEF}$ by fitting the linear equation

$$x_{i \in FEF} = \beta_0 + \sum_{j \in PFC} \beta_j x_j + \sum_{j \in LIP} \beta_j x_j + \sum_{j \in V4} \beta_j x_j + \sum_{j \in MT} \beta_j x_j + \sum_{j \in IT} \beta_j x_j + \epsilon \quad (4.2)$$

where x_i corresponds to the mean ITI spiking activity across trials and β_x corresponds to the estimated coefficients. This procedure was repeated for each session, and R^2 values were calculated for each session, too.

4.3.4 Spiking activity flow estimation

Our aim was to predict the FEF activity during the response interval (saccade) using spiking activity from other cortical areas during the stimulus and task rule intervals. Thus, we constructed an empirically-derived neural network (ENN) that is conceptually similar to the model in Chapter 3. Broadly, this parameterizes a functional neural network model that predicts the spike rate in output units (in

FEF) using spike rate activity from input units (Figure 4.1c).

We used the trial-to-trial spike rates during the stimulus and task rule intervals from input areas as inputs to our ENN. Input units were defined as all sorted units within each region excluding FEF (i.e., PFC, LIP, MT, IT, V4). We excluded FEF units to avoid the circularity of using a region's units to predict itself.

We activity flow mapped spiking task rule and stimulus activations onto all units in the input areas (Figure 4.3c). Conceptually, this can be thought of as a recurrent computation, whereby both stimulus and rule activations are mapped onto each other. More formally, however, this can be conceived as an unrolled feedforward neural network, whereby stimulus and task rule activations simultaneously project onto a shared hidden layer, namely all units within PFC, LIP, MT, IT, and V4 (Figure 4.3c). Mathematically, we define this mapping as

$$\mathbf{X_{hidden1}} = \mathbf{f}(\mathbf{X_{stimulus}}\beta_{\mathbf{ITI}} + \mathbf{X_{rule}}\beta_{\mathbf{ITI}}) \tag{4.3}$$

where $\mathbf{X_{hidden1}}$ refers to the predicted integration of stimulus and task rule spiking activity (i.e., the hidden layer in an unrolled recurrent computation), $\mathbf{X_{stimulus}}$ refers to the spike rate during the stimulus interval for all input areas, $\mathbf{X_{rule}}$ refers to the spike rate during the task rule interval for all input areas, and $\beta_{\mathbf{ITI}}$ corresponds to the inter-unit FC matrix estimated from ITI activity. f corresponds to a rectified linear function, which thresholds any negative predictions as 0 (i.e., $f(x) = max(x, 0)$).

We subsequently performed the activity flow computation once more. However, we only use predicted activations (i.e., $\mathbf{X_{hidden1}}$ as input). Conceptually, this is similar to performing an additional recurrent computation. Formally, this was calculated as

$$\mathbf{X_{hidden2}} = \mathbf{f}(\mathbf{X_{hidden1}}\beta_{\mathbf{ITI}}) \tag{4.4}$$

$\mathbf{X_{hidden2}}$, which contain predicted spike rates from input units, was then projected to FEF via another activity flow step. Thus, our final predicted FEF spiking activity was obtained by calculating

$$\mathbf{X_{FEF}} = \mathbf{f}(\mathbf{X_{hidden2}}\beta_{\mathbf{ITI,FEF}}) \tag{4.5}$$

where $\mathbf{X_{FEF}}$ is the predicted spiking activity of FEF, and $\beta_{\mathbf{ITI,FEF}}$ is the FC estimates between input areas and FEF. Importantly, a unique $\mathbf{X_{FEF}}$ was generated for every trial using the spike rate obtained for the task rule and stimulus intervals associated with that trial.

We next assessed how accurately $\mathbf{X_{FEF}}$ predicted spiking activity associated with behavioral responses (i.e., saccades during the response period). We compared our predicted FEF spiking activity with the actual FEF spiking activity using a cross-validated decoding scheme. This scheme is conceptually similar with the one employed in [Ito et al., 2017] (Figure 4.3c). Cross-validation was applied across trials using a 10-fold scheme. We trained a linear decoder on the predicted FEF spiking activations, and classified the held-out trials (using the actual response period activity). Training and testing on different trials ensured that no spontaneous or trial-to-trial variance could explain the successful prediction of FEF spiking activity. Thus, this cross-validation scheme ensured three forms of statistical independence: 1) task condition independence (since task cue and stimulus activity were used to predict response activity); 2) spatial independence, since only PFC, LIP, MT, IT, and V4 activity was used to predict FEF; 3) temporal independence, since distinct sets of trials were used to predict the spiking activity of a held-out trial.

As above, we used a linear decoder that was trained using logistic regression. Parameters are the same as in the above sections.

4.4 Results

4.4.1 Baseline decoding of different task conditions

We first established a baseline of how much each cortical area contained information for different task conditions. This involved decoding task rule conditions, sensory stimulus features, and behavioral responses using spiking activity. Given that our hypothesis was that the integration of task rule and sensory stimulus activity could be transformed to predict response activity, it was critical that at least one cortical area contained representations for each task condition.

We performed task rule, sensory stimuli, and behavioral response decoding. Note that each task condition decoded from temporally adjacent yet non-overlapping intervals during a trial (Figure 4.1b). For task rule decoding, we classified spiking activity as either associated with the color rule or motion direction rule. We classified the the task rule interval using the multi-unit spike rates within a cortical region as features. We found that all six regions contained significantly decodable activity related to task rule conditions (Figure 4.2c; PFC accuracy=59.46%, p=8.30e-12; FEF accuracy=59.54%, p=9.08e-11; IT accuracy=74.62%, p=3.19e-9; LIP accuracy=63.63%, p=3.00e-12; MT accuracy=63.90%, p=8.30e-7; V4 accuracy=85.93%, p=4.58e-13). This is consistent with previous work on this same data set, which showed that all cortical areas contained some percentage of units that were responsive to task rule information (using a task variance analysis rather than decoding analysis) [Siegel et al., 2015].

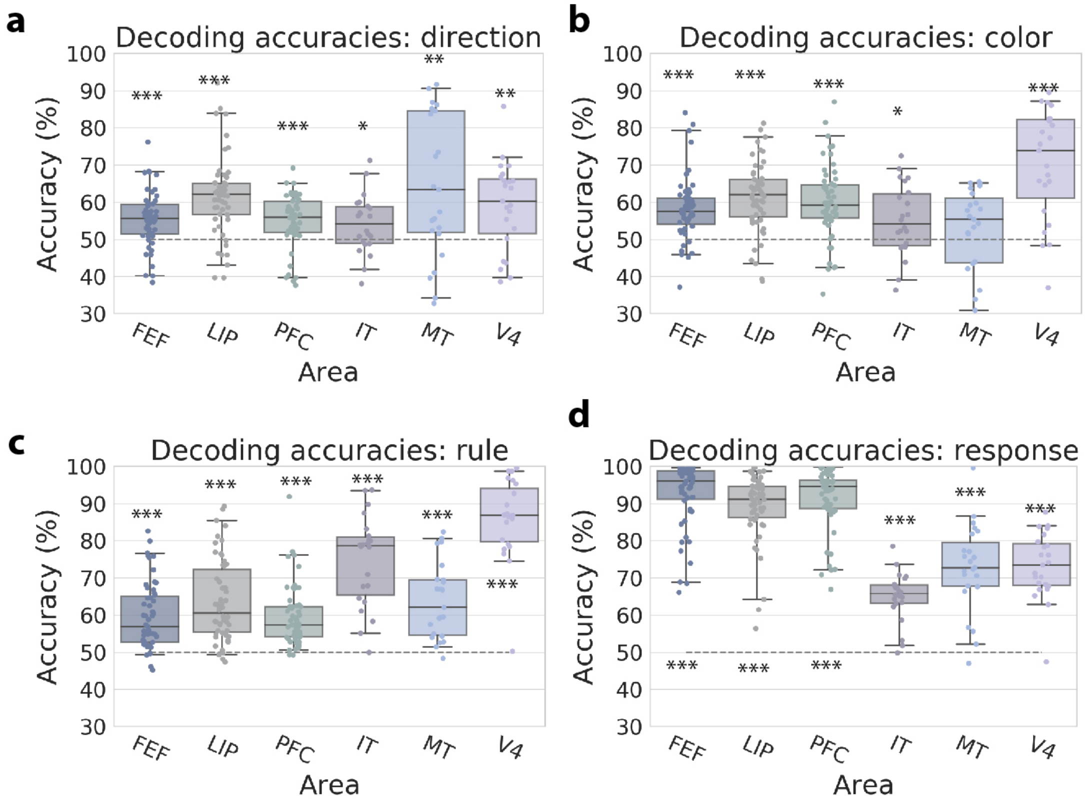


Figure 4.2: Baseline decoding of stimulus, task rules, and response period activity in each cortical area. We performed a decoding analysis using a linear decoder on trial-to-trial spiking activity, using the spike rate of multi-units as decoding features. For each recording session, we performed a decoding across trials on the spike rate of either the task rule, stimulus, or response period. We decoded a) direction (up/down) of moving dots during the stimulus period, b) color of the dots during the stimulus period, c) the rule (indicated by the cues), and d) the behavioral response (saccade). Note that rules were indicated by four possible cue stimuli. Two cues indicated the color categorization rule, and two cues indicated the direction categorization rule. Boxes represent the interquartile range of the distribution, whiskers represent 95% of the distribution, and the grey line represents the median. Each dot in the strip plot represents the average accuracy for a recording session. Gray line indicates chance decoding. (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.0001$.)

Next, we sought to identify which cortical areas contained sensory stimulus information. Since the task required the subjects to categorize either color or direction features, this involved decoding two sensory stimuli features: color (red versus green) and direction (upward or downward motion). We first decoded trial-to-trial motion information using the spiking activity during the stimulus interval. We found that all six areas could successfully decode direction information (Figure 4.2a; PFC accuracy=54.92%, p=6.20e-6; FEF accuracy=55.55%, p=1.63e-6; IT accuracy=54.03%, p=0.01; LIP accuracy=61.59%, p=4.45e-10; MT accuracy=64.56%, p=0.0009; V4 accuracy=59.06%, p=0.0009). We also found that 5/6 areas could successfully decode color information (Figure 4.2b; PFC accuracy=59.57%, p=1.26e-9; FEF accuracy=58.23%, p=1.95e-6; IT accuracy=55.13%, p=0.01; LIP accuracy=61.38%, p=1.07e-11; MT accuracy=52.26%, p=0.18; V4 accuracy=70.11%, p=1.09e-6). These findings are also consistent with a previous analysis on the same data set, which showed that stimulus information is widely distributed across these six cortical areas [Siegel et al., 2015].

Finally we were interested in identifying which areas contained decodable response information. While we primarily hypothesized that response information would likely exist within FEF (given that the task required saccade responses), we nevertheless performed decoding analyses on all six cortical areas. Again, we found that all six cortical areas could reliably decode trial-to-trial response information using spiking activity during the response interval (PFC accuracy=91.46%, p=1.34e-39; FEF accuracy=92.60%, p=2.23e-39; IT accuracy=64.90%, p=7.35e-10; LIP accuracy=89.07%, p=3.79e-36; MT accuracy=72.62%, p=3.49e-9; V4 accuracy=73.15%, p=1.41e-11; Figure 4.2d). Although all six cortical regions contained motor response information, as expected, we found that FEF had the highest response decoding accuracy. The identification of distributed representations laid the foundation to test how different task representations interact to transform stimulus and rule representations into FEF motor representations.

4.4.2 Transforming task rule and stimulus activity into FEF response period activity

We were able to decode task rule, stimulus, and response information using the spiking activity of most cortical areas. However, how do these distinct pieces of information interact during task performance to form the correct motor response signals (i.e., eye movements)? Here we focus on addressing how motor response information is computed across cortical areas through network computations.

Our previous work illustrated that different pieces of information in different brain areas, like task rules and stimulus representations, can be integrated to form response representations in motor cortex during tasks [Chapter 3]. Moreover, these transformations can be implemented by measuring intrinsic FC mappings between brain areas. Thus, to obtain weight mappings between sets of units, we first estimated the inter-unit FC maps using spiking activity from inter-trial intervals (ITI). We used ITI activity since there was no true 'resting-state' in the current data set. Importantly, the ITI was stimulus-free in that it did not contain any task-relevant information (Figure 4.3a). Thus, we believed that estimating FC from the ITI would be useful in approximating the true intrinsic FC structure between sets of units.

We computed the FC weights to a multi-unit using cross-validated ridge regression (see Methods). We identified two sets of FC weights. We first identified FC weights between all multi-units within input areas. Given that we found that most areas contained decodable information about the task, we defined input areas as all cortical areas (multi-units) excluding FEF (Figure 4.3b). Specifically, for a given unit in an input area, we estimated the cross-validated regression weights of all other multi-units to that target unit using ITI activity. This provided a weight mapping that enabled us to predict the activity of a unit using the spiking activity of all other units (excluding those in the FEF). Importantly, this

FC mapping enabled us to integrate the spiking activity of task rule and sensory stimulus activity by modeling activity flow between all input units (Figure 4.3c). Activity flow modeling of stimulus and task rule activity is modeled as the sum of weighted activity flow of spiking activity from both conditions onto each unit in the input areas. We applied a rectified linear function f, that removes any negative spike rate predictions, since those are biologically implausible.

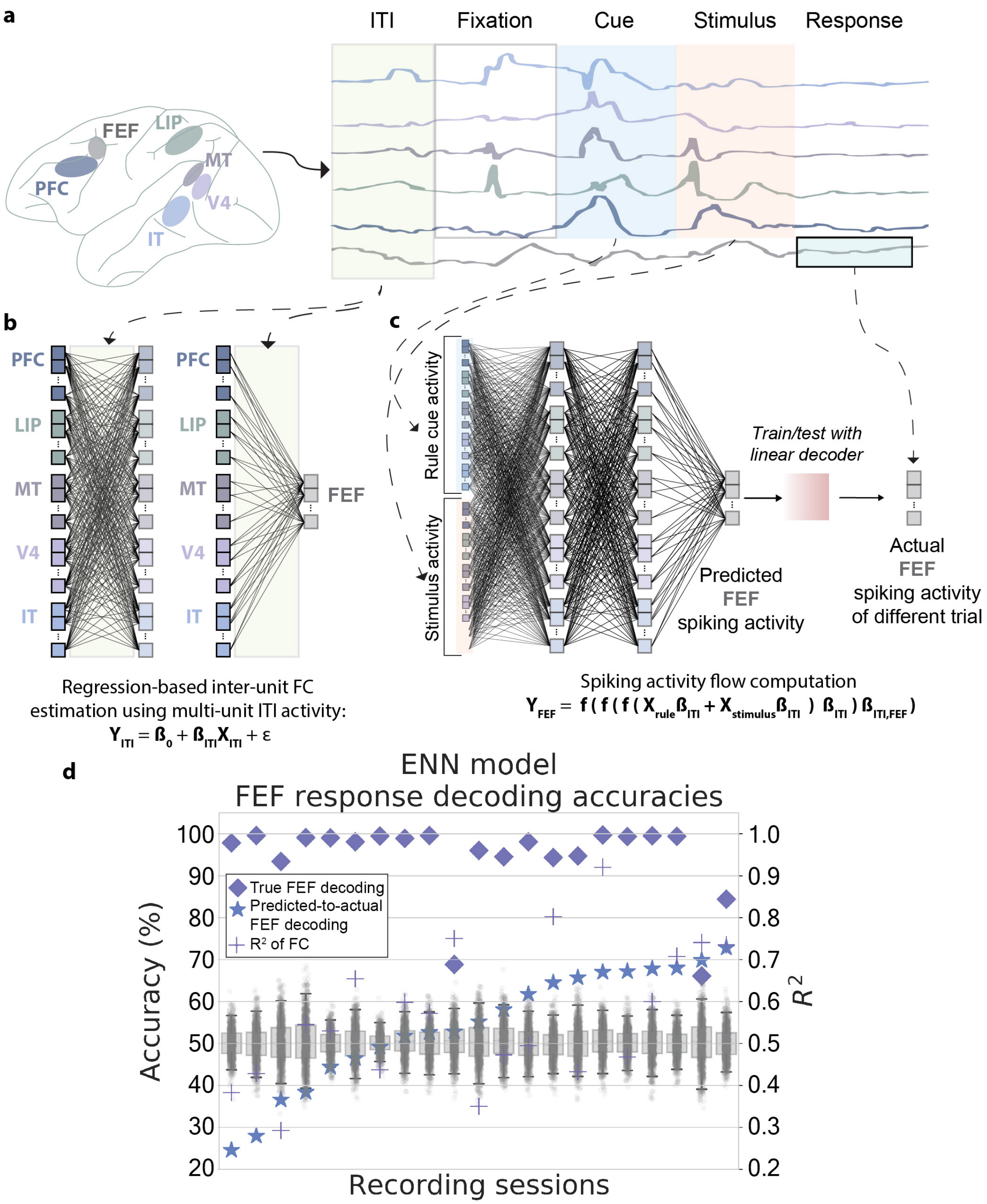

a
ITI
Fixation
Cue
Stimulus
Response
FEF
LIP
MT
PFC
V4
IT
b
PFC
LIP
MT
V4
IT
PFC
LIP
MT
V4
IT
FEF
c
Rule cue activity
Stimulus activity
Train/test with linear decoder
Predicted FEF spiking activity
Actual FEF spiking activity of different trial
Regression-based inter-unit FC estimation using multi-unit ITI activity:
$Y_{ITI} = ß_0 + ß_{ITI}X_{ITI} + ε$
Spiking activity flow computation
$Y_{FEF} = f(f(f(X_{rule}ß_{ITI} + X_{stimulus}ß_{ITI})\ ß_{ITI})\ ß_{ITI,FEF})$
d
ENN model
FEF response decoding accuracies
True FEF decoding
Predicted-to-actual FEF decoding
R^2 of FC
Accuracy (%)
100
90
80
70
60
50
40
30
20
R^2
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Recording sessions

Figure 4.3: Constructing ENNs from spiking data to predict inter-area sensorimotor transformations. a) We extracted the spike rate from different intervals during task execution from six cortical areas. b) We performed a cross-validated ridge regression between all units in PFC, LIP, MT, IT, and V4 to estimate inter-unit FC. Separately, to identify the weights to the FEF, we estimated the regression weights between PFC, LIP, MT, IT, and V4 to FEF. c) To predict FEF response period activity associated with saccades, we used stimulus and task rule activity from all recording areas *excluding* FEF. We modeled activity flow between PFC, LIP, MT, IT, and V4, prior to modeling activity flow to FEF. The activity flow computation involves a linear weighted sum of the spike rates of the source units weighted by their connectivity coefficients, passed through a nonlinearity f. d) Summary of the FEF predictions from the ENN, sorted by predicted-to-actual FEF decoding. We could successfully predict the actual FEF response spike rates for 10/21 recording sessions, after correcting for multiple comparisons using an FDR-corrected $p<0.05$ threshold. Interestingly, the predicted-to-actual decoding accuracies were significantly correlated ($r=0.56$; $p<0.01$) with the R^2 value of the FC regression fit. Gray box plots represent a null distribution obtained from shuffling labels.

We next estimated inter-unit FC weights between input areas and the FEF (Figure 4.3b). This provided a simple linear model to transform neural spiking activity from other areas into the spatial geometry of the FEF. Moreover, this would enable us to predict whether or not the integration of task rule and stimulus activity could accurately predict motor response patterns in the FEF.

Our objective was to accurately predict the response period activity of FEF using stimulus and task rule activity from units outside the FEF. Thus, we modeled the spiking activity flow of stimulus activity and task-rule activity as recurrent interactions among input units. We modeled two recurrent interactions, which can be formalized as a feedforward neural network with two hidden layers (see Figure 4.3c). Importantly, the weights between the two hidden layers are fixed, and estimated empirically (via regression on ITI activity). To generate a predicted FEF response, we model spiking activity flow from the input areas onto FEF using FC estimates.

We generated predicted spike rates for all FEF units on a trial-by-trial basis. To evaluate the validity of the activity flow model in predicting sensorimotor transformation, we compared the predicted FEF spiking patterns to the actual spiking patterns of FEF units. In line with a prediction perspective (and previous work [Chapter 3]), we trained a decoder on predicted FEF spiking patterns. The decoder was trained using labels determined by the actual response of that trial (Figure 4.3c). (Moreover, only correct trials were included in this decoding scheme.) We then applied the decoder to classify the actual FEF response patterns of held-out trials. In other words, we applied the decoder on trials in which the decoder was not trained on (see Methods). This ensured that our decoding scheme was not circular in three ways: 1) FEF predictions were spatially independent, since predictions were generated from all units outside the FEF; 2) FEF predictions were condition independent, since while we predicted response patterns, only task rule and stimulus interval activity was used for prediction; 3) the decoder was tested on different trials than it was trained on. We found that of all 21 recording sessions included, we could accurately predict FEF response activations for 10/21 sessions (average accuracy of significant sessions=66.29%, FDR-corrected $p < 0.05$; Figure 4.3d). These results suggest that though some recording session were unable to predict FEF response periods, we were able to accurately model neural transformations to FEF to predict behavioral responses in roughly half the recording sessions.

4.4.3 Successful inter-area predictions depend on robust FC estimates

Roughly half recording session failed to predict FEF response spiking patterns. Why could some recording sessions accurately predict neural transformations of FEF responses, while others could not? Unlike whole-brain fMRI imaging, which has access to the entire brain (albeit at limited spatial resolution), our electrophysiological recording had limited spatial coverage. Moreover, this spatial coverage varied from recording session to session. Thus, we hypothesized that variable prediction of FEF response patterns was due to the variability in sampling units in input areas that were "functionally connected" to FEF units.

For each recording session, we calculated the cross-validated regression-based FC values from input units to FEF units using ITI activity (Figure 4.3b). We then calculated the average R^2 of the model fit to all FEF units for each session. This provided a metric that described how well FEF units were "functionally connected" to the input units for each session. We then correlated across sessions this mean R^2 value with the predicted-to-actual decoding accuracy (Figure 4.3d). Consistent with our hypothesis, we found that the R^2 of the FC was positively correlated with the predicted-to-actual decoding accuracies in FEF (r=0.56; p<0.01). This demonstrated that the ability to accurately decode FEF responses was dependent on the ability to robustly estimate the FC between input and FEF units during the ITI periods.

Thus, the current results demonstrated that our ability to accurately capture network computations that describe flexible sensorimotor transformations was limited by the ability to estimate useful FC between input units and FEF units.

4.4.4 Response information begins to form during the stimulus interval

Our primary hypothesis was that the integration of task rule and stimulus information would accurately predict response information that emerges after the stimulus interval. In our previous study [Chapter 3], we were able to statistically disentangle (i.e., orthogonalize) stimulus from motor response information through a task GLM. However, in the present study, we focused on trial-to-trial response estimates, given that the response interval was temporally independent (occurred after) stimulus offset. However, it is conceivable that response information begins to form during the stimulus interval, given that once the subject receives both task rule and stimulus information, subjects can form a response decision.

Thus, we assessed whether response information might have emerged during the stimulus interval. However, as a control, we tested whether response information might emerge during the task rule interval. (No response can be made with only task rule information.) As expected, we found that none of the six cortical areas contained decodable response information during the rule interval (Figure 4.4a).

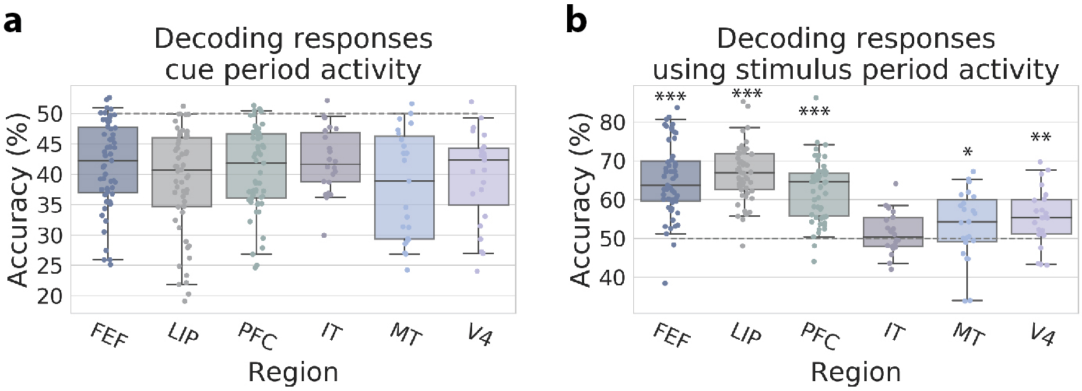


Figure 4.4: Evaluating whether response information existed during the task rule and stimulus interval. Our objective was to predict response signals in FEF using only context and stimulus information. Thus, as a baseline, we sought to evaluate whether it was possible to decode response information during either the task rule or stimulus interval. a) We first trained a decoder to classify response information in the task rule interval. Theoretically, response information for a given trial should not exist during the task rule interval, since prior to the stimulus presentation no correct decision can be made. We found that no cortical areas could decode the response information during a trial using task rule activity. b) We next decoded response information during stimulus interval. Theoretically, it may be possible to decode response information during the stimulus interval since the subject has enough information (rule and stimulus) to make an appropriate motor response decision. Indeed, we found that almost all regions (except IT) could decode motor responses during stimulus interval. Boxes represent the interquartile range of the distribution, whiskers represent 95% of the distribution, and the grey line represents the median. Each dot in the strip plot represents the average accuracy for a recording session. Gray line indicates chance decoding. (* = p < 0.05, ** = p < 0.01, *** = p < 0.0001.)

Next, we evaluated whether we could decode response information during the stimulus interval. Indeed, response information appeared to form during stimulus period in all recording areas except for IT (Figure 4.4b; FEF accuracy=65.06%; PFC accuracy=62.47%; LIP accuracy=66.81%; MT accuracy=53.46%; V4 accuracy=55.39%; FDR-corrected $p < 0.05$ except for IT). However, though response information could be decoded during the stimulus period (while subjects were required to fixate), response decoding accuracies were notably lower than during the response period (Figure 4.2d). Nevertheless, this suggests that the response predictions in FEF reported in the previous section may have been due to small amounts of response information beginning to form during the stimulus period. Given our initial hypothesis that we could predict sensorimotor transformations from input areas to FEF, future follow-up analyses will need to rule out the possibility that response information during the stimulus interval played a role in generating FEF activity predictions that could decode behavioral responses.

4.5 Discussion

Cognitive information during tasks has been shown to be widely encoded in multi-unit activity across cortex. Previous work on this same data set has reported that following a bottom-up sweep of task information, task information flows from frontoparietal to visual areas in a sustained manner [Siegel et al., 2015]. However, that studied focused on the temporal dynamics of task information in each area, without asking how the brain communicates that information between areas. Using previously developed methods for modeling inter-area activity flow in fMRI data, we constructed ENN models that could account for the flow of spiking activity between pairs of units during flexible behavior.

Using ITI activity, we estimated inter-unit FC mappings that we hypothesized may reflect functional pathways for neural communication. Statistically, this approach is conceptually similar to the estimation of intrinsic FC that is common to the fMRI literature, though we replace voxel and/or parcel time series with neural spike rate data. Importantly, constructing a functional network model of inter-unit interaction enabled us to test whether a technique developed for mapping task activations in fMRI data – activity flow mapping – could predict multi-unit spiking activity during a flexible sensorimotor task. We had two main findings: 1) That for some recording sessions, we could accurately predict FEF response activity from other areas using empirically-estimated functional network connectivity; 2) Successful prediction was dependent on the ability to record from units that were "functionally connected" (as determined by the R^2 fit of the FC model) to downstream FEF units.

The present results extend the concept of FC, which are common in the fMRI literature, to neural spiking data. While the spiking literature has developed a rich literature studying neural correlations (i.e., noise correlations) [Aertsen et al., 1989, Cohen and Kohn, 2011, Pillow et al., 2008], few studies have investigated how units in one area are predictive of units in another area. This is likely due to the technical difficulties in performing experiments that obtain the widespread recordings in behaving animals. However, one recent study was able to identify a communication subspace by which activity in V1 units could predict activity in downstream V2 [Semedo et al., 2019]. However, this study was limited to recordings in V1 and V2 during the presentation of simple visual stimuli. Using a previously published data set [Siegel et al., 2015], we were able to extend these ideas to predict response activity in FEF by modeling the spiking activity flow from units in five other cortical areas during a flexible sensorimotor task.

Recent work in whole-brain fMRI imaging has shown a strong correspondence between the FC estimated during task-free and task-evoked states [Cole et al., 2014a, Krienen et al., 2014, Bolt et al., 2017]. Importantly, we recently leveraged this strong correspondence between task-free and task-evoked states to parameterize functional network models to estimate inter-area activity flow [Cole et al., 2016a, Ito et al., 2017]. (The use of task-free states to parameterize network models was important to rule out the potential circularity of using the same data for both network estimation and task activation prediction.) In the present study, we found that network estimates of stimulus-free ITI activity can be used to accurately model multi-unit spiking activity flow. This is consistent with findings in fMRI studies, and suggests that there is likely a strong correspondence between stimulus-free and stimulus-dependent functional network organization in electrophysiological recordings. Additionally, in the subsequent chapter, we find consistent changes across intrinsic and task-evoked states in both large-scale fMRI

FC and mean-field spike count correlations, suggesting that despite differences in the data acquisition techniques (e.g., BOLD versus multi-unit spikes), there likely exist similarities in the underlying processes that govern neural interactions.

We assessed how well FC-based techniques could predict spiking activity flow between sets of multi-units. However, one limitation is that neural spiking data has limited spatial coverage, even with multi-unit recordings from six different areas. Despite having more direct neural recordings and higher signal-to-noise ratio, this spatial undersampling contrasts with human fMRI data, which has whole-brain spatial coverage. Whole-brain sampling increases the likelihood that we can build better network connectivity estimates to a target region. This is because we can include all other brain areas as potential predictors to a target region and optimize for target prediction using regression-based techniques [Cole et al., 2016a]. Despite this limitation, we still performed a statistically analogous analysis by optimizing for prediction on FEF target units using all other input units as predictors and using cross-validated regularized regression. Interestingly, our results showed that so long as we can sample from units that are informative with respect to a target area (e.g., high R^2 via linear regression to predict FEF activity), we can accurately build functional network models that can predict spiking activity in target units.

Our primary hypothesis was that the integration of stimulus and task rule activity would successfully predict motor response spiking activity through activity flow modeling. While the stimulus and task rule intervals preceded the response interval and did not overlap, we found in follow-up analyses that response behavior began to form during the stimulus interval. This finding is also consistent with previous accounts with the same data, which found that choice information began to form as early as 100ms before response onset [Siegel et al., 2015]. This indicates that the projection of stimulus information from input areas onto FEF areas likely contains some response information. This suggests the possibility that our mapping of stimulus activity onto FEF units may not truly reflect the transformation of cognitive information, but instead reflect only information transfer [Ito et al., 2017]. However, previous work showed that response information only formed 100ms prior to response onset, a small fraction of the stimulus interval (lasting up to 3 seconds) [Siegel et al., 2015]. Nevertheless, future work will focus on dissociating this possibility and removing any response related confounds in stimulus period spike rates.

In conclusion, we modeled sensorimotor transformations during flexible behavior by applying the activity flow framework to multi-unit spiking data. Our results demonstrate the feasibility of bridging theories and concepts used in human neuroscience onto animal neurophysiology studies, bridging two subfields in neuroscience. We showed that network estimation techniques in fMRI, such as multiple regression-based techniques, accurately predicted the spiking activity of target multi-units during flexible sensorimotor processing. Thus, the present work illustrates that we can begin to probe the network mechanisms of cognitive processes in neural spiking data by combining network estimation techniques during stimulus-free periods and task-based manipulations.

Chapter 5

Task-evoked activity quenches neural correlations and variability across cortical areas

This chapter has been published in PLOS Computational Biology [Ito et al., 2020a]. The contents have been reformatted for this thesis.

5.1 Abstract

Many large-scale functional connectivity studies have emphasized the importance of communication through increased inter-region correlations during task states. In contrast, local circuit studies have demonstrated that task states primarily reduce correlations among pairs of neurons, likely enhancing their information coding by suppressing shared spontaneous activity. Here we sought to adjudicate between these conflicting perspectives, assessing whether co-active brain regions during task states tend to increase or decrease their correlations. We found that variability and correlations primarily decrease across a variety of cortical regions in two highly distinct data sets: non-human primate spiking data and human functional magnetic resonance imaging data. Moreover, this observed variability and correlation reduction was accompanied by an overall increase in dimensionality (reflecting less information redundancy) during task states, suggesting that decreased correlations increased information coding capacity. We further found in both spiking and neural mass computational models that task-evoked activity increased the stability around a stable attractor, globally quenching neural variability and correlations. Together, our results provide an integrative mechanistic

account that encompasses measures of large-scale neural activity, variability, and correlations during resting and task states.

5.2 Introduction

Measures of neural correlations and variability are widely used in neuroscience to characterize neural processes. During task states, neural variability has consistently been shown to be reduced during tasks across human functional magnetic resonance imaging (fMRI) [He, 2011, He, 2013, Fegen, 2012], local neural populations [Churchland et al., 2010, Hennequin et al., 2018, Jacobs et al., 2018], and both spiking [Litwin-Kumar and Doiron, 2012, Hennequin et al., 2018] and mean-field rate models [Deco and Hugues, 2012, Ponce-alvarez et al., 2015]. Despite this convergence in the neural variability literature, there are disparities in the use and interpretation of neural correlations. In the human fMRI literature, neural correlations are often estimated by measuring the correlation of blood oxygenated level-dependent (BOLD) signals and is commonly referred to as functional connectivity (FC) [Biswal et al., 1995]. In the non-human primate (NHP) spiking literature, neural correlations have been measured by computing the correlation between the spike rate of two or more neurons and is commonly referred to as the spike count correlation (or noise correlation) [Cohen and Kohn, 2011]. Yet despite the use of different terms, the target statistical inference behind these two techniques is consistent: to characterize the interaction among neural units.

In the human fMRI literature, studies have identified large-scale functional brain networks through clustering sets of correlated brain regions using resting-state activity [Power et al., 2011, Yeo et al., 2011, Ji et al., 2019]. During task states, the FC structure has been demonstrated to dynamically reconfigure [Cole et al., 2014a, Krienen et al., 2014, Gonzalez-Castillo and Bandettini, 2017]. Though it has been suggested that correlation increases and decreases respectively facilitate and inhibit inter-region communication [Tomasi et al., 2014, Gonzalez-Castillo and Bandettini, 2017], the mechanistic bases of these FC changes remain unclear.

Studies in the local circuit literature using electrophysiological recordings in animals have characterized the correlation structure between neuron spikes across a range of task demands. These studies have found that the spike count correlation (or noise correlation) between neuron spikes generally decreases during task states, particularly for neurons that are responsive to the task [Cohen and Maunsell, 2009, Ecker et al., 2010, Ruff and Cohen, 2014, Pinto et al., 2019]. Moreover, these empirical studies have been accompanied by theoretical work, which has suggested that the reduction in noise correlations may enhance information coding by suppressing shared spontaneous activity and reducing neural noise [Aertsen et al., 1989, Averbeck et al., 2006, Cohen and Kohn, 2011, da Silveira and Berry, 2014, Doiron et al., 2016]. Thus, the theoretical framework behind noise correlations may also provide a solid foundation from which to advance understanding of fMRI FC [Aertsen et al., 1989, Averbeck et al., 2006, Cohen and Kohn, 2011, Doiron et al., 2016].

Here we sought to quantify the relationship between neural correlations (i.e., FC) in large-scale human imaging and (local circuit) animal neurophysiology (spike count correlations). In particular, it is unclear whether observations at the local circuit level would be consistent with observations made across large cortical areas. To further complicate this issue, we recently found that task activations can inflate task functional connectivity estimates in human fMRI data [Cole et al., 2019], suggesting some previous neuroimaging studies may have erroneously reported correlation increases due to inaccurate removal of the mean-evoked response. Importantly, the removal of the mean task-evoked response is a standard procedure in the spiking literature, a critical step designed to dissociate signal correlations (task-to-neural associations) from noise correlations (neural-to-neural associations) [Aertsen et al., 1989, Averbeck et al., 2006, Cohen and Kohn, 2011]. (In the fMRI literature, signal correlations and noise correlations are both statistically and conceptually analogous to task co-activations and functional connectivity, respectively.) Thus, to accurately bridge the FC literature with the spike count correlation literature, it was necessary to analyze the data in a statistically consistent way. This enabled us to adjudicate the differing perspectives in the neural correlation literature while simultaneously confirming and extending previous findings on task-state neural variability reduction.

We report multiple sources of empirical and theoretical evidence demonstrating that task-evoked activity quenches neural correlations and variability across cortical areas. First, we characterize task-evoked neural variability and correlations in empirical data using two highly distinct data sets: multi-site and mean-field NHP spike rates and whole-brain human fMRI (Figure 5.1). This allowed us to test whether there were consistent large-scale variability and correlation changes during task states independent of data acquisition technique. Moreover, this allowed us to take advantage of the more direct neural recording with NHP electrophysiology along with the more comprehensive coverage of human fMRI (in addition to testing for translation of findings to humans). Next, to provide a mechanistic account capable of explaining our empirical findings, we used both spiking and neural mass models to parsimoniously explain variability and correlation suppression across mean-field cortical areas. This led to theoretical insight using dynamical systems analysis, demonstrating that task-evoked activity strengthens the system's attractor dynamics around a stable fixed point in neural mass models, quenching neural correlations and variability. The combination of simultaneously recorded mean-field spike rate recordings from six cortical sites, whole-brain fMRI obtained from seven different cognitive tasks, and dynamical systems modeling and analysis provide a comprehensive account of task-related correlation and variability changes spanning species and data acquisition techniques.

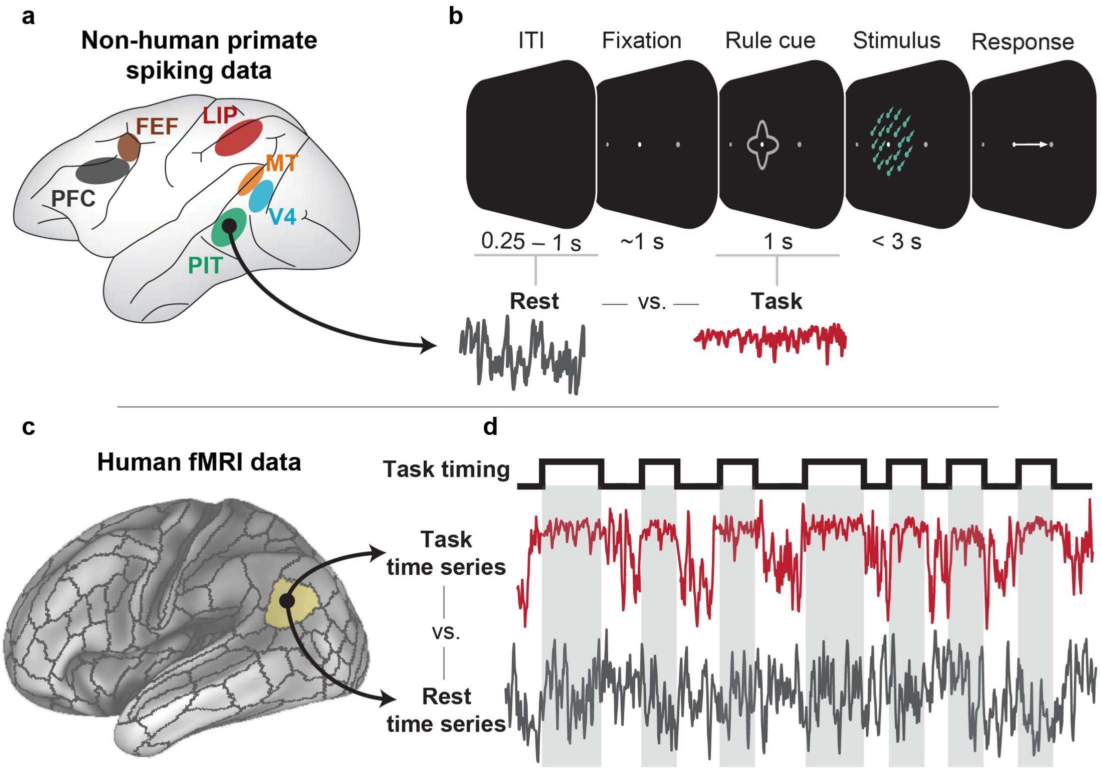


Figure 5.1: Testing the hypothesis that task-evoked neural variability and correlations are quenched across cortical areas in NHP spiking and human fMRI data sets. We used two highly distinct data sets to test the hypothesis that task-evoked activity globally quenches neural variability and correlations to suppress background spontaneous activity/noise. This contrasts with the alternate hypothesis, namely that task-evoked activity increases variability and correlation to facilitate inter-region communication. Importantly, the two data sets were analyzed in a statistically consistent manner, including the removal of the mean task-evoked response to isolate neural-to-neural interactions. a,b) Using mean-field spike rate data collected simultaneously from six different cortical areas [Siegel et al., 2015], we compared the spiking variability and spike count correlations between task-state (i.e., following task cue onset) and rest-state spiking activity. We defined rest state as the inter-trial interval (ITI) directly preceding the trial. This was performed by estimating the mean-field spike rate by averaging across multi-units in each cortical area, allowing us to target the activity of large neural populations. c,d) Using human fMRI data obtained from the Human Connectome Project [Barch et al., 2013], we compared the neural variability and correlations (i.e., FC) of the BOLD signal during task block intervals to equivalent resting-state intervals. We used seven highly distinct cognitive tasks. Time series and task timings are illustrative, and do not reflect actual data.

5.3 Methods

5.3.1 Spiking data: Data collection

The behavioral paradigm for each monkey was a motion-color categorization task (Figure 5.1b). Experimental methods for electrophysiology data collected for NHP was previously reported in [Siegel et al., 2015] and [Brincat et al., 2018]. Data was collected in vivo from two (one female) behaving adult rhesus macaques (Macaca mulatta) across 55 sessions. Data from six distinct cortical regions were recorded simultaneously from acutely inserted electrodes. Cortical regions included: MT, V4, PIT, LIP, FEF, and LPFC (Figure 5.1a). Spikes from each region were sorted offline into isolated neurons. However, given our interest in inter-region neural correlations across large scale neural systems, we pooled spikes from each functional area into a single spike rate time series. For each trial, spikes were sorted for a 5s period, beginning 2.5s prior to stimulus onset, and until 3.5s after stimulus onset. Further details regarding electrophysiological data collection can be found here: `http://www.sciencemag.org/content/348/6241/1352/suppl/DC1` and here: `http://www.pnas.org/content/pnas/suppl/2018/07/09/1717075115.DCSupplemental/pnas.1717075115.sapp.pdf`

All statistical analyses in the main article (detailed below) were performed on a single monkey. Independent replication was performed on the second monkey, and is reported in Supplementary Figure C.1.

5.3.2 Spiking data: Task versus rest variability analysis

Neural variability analysis was analyzed using an analogous approach to both the computational model and fMRI data. However, since we had no true 'resting-state' activity for the monkey data set, we used the inter-trial interval (ITI; 0.5s - 1s variable duration, see Figure 5.1b) as "resting-state activity". We used the

0.5s - 1s interval immediately preceding the trial's fixation period to avoid any reward/feedback signals from the previous trial. (Reward/feedback from the previous trial was provided more than 1.5s prior to the fixation period.) Spike counts were calculated by taking a 50ms sliding window with 10ms increments, consistent with previous studies [Churchland et al., 2010]. The mean-evoked response across all trials for a given task rule (e.g., motion rule versus color rule) was calculated and removed from each trial, as is common in the spike count literature [Cohen and Kohn, 2011] and the fMRI literature [Cole et al., 2019]. (Statistically this is equivalent to performing the task activation regression in the fMRI data, described below.) The mean task-evoked response of the ITI period associated with each task condition was also removed. This was to control for any artifacts that might be induced due to removal of the mean-evoked response. Trials with less than 500ms (or 50 time points) worth of spiking data for either the ITI and/or task cue presentation were excluded. This was done to reduce variability of the estimated spike count correlations, since correlations with few observations are highly variable.

We computed the variance across 25 consecutive trials using the spike rate from each cortical recording during either the ITI or the task cue period. This was repeated for all trials for each subject. We used across-trial variance to calculate variability rather than Fano factor [Churchland et al., 2010]. This choice was due to the insight from our model illustrating that the mean-evoked activity and the corresponding variance interact in a nonlinear manner, and that the Fano factor is computed as the variance over the mean. Cross-trial variance was computed as

$$Var = \sum_{trial=x}^{x+n} \frac{(r_{trial} - \bar{r})^2}{n-1} \tag{5.1}$$

Where $n = 25$ trials, r_{trial} reflected the spike rate of each trial, and $\bar{r}$ the cross-trial mean firing rate for the task condition (i.e., either the cross-trial mean firing rate during the color or motion task cue period).

The statistical difference in task versus rest neural variability was computed by using a two-way, paired t-test across all bins of 25 consecutive trials. The global neural variability change was computed by averaging the variance across all recording areas for each bin. Statistics for the regional neural variability change were corrected for multiple comparisons using an FDR-corrected threshold of $p < 0.05$.

In addition, we computed the variance during the ITI and task cue period within trial (across time points) (Supplementary Figure C.2). This analysis demonstrated that moment-to-moment variability (rather than trial-to-trial variability) was also quenched from rest to task periods, suggesting that variability quenching also occurs at faster timescales. The statistical difference in task versus rest neural variability was computed by using a two-way, paired t-test (paired by trial) across all trials for each monkey separately. The global neural variability change was computed by averaging the variance across all recording areas for each trial (Supplementary Figure C.2). Statistics for the regional neural variability change were corrected for multiple comparisons using an FDR-corrected threshold of $p < 0.05$.

5.3.3 Spiking data: Task versus rest state correlation analysis

Neural correlations for spiking data using the same preprocessing steps mentioned above for spike rate variability analysis. Specifically, the mean-evoked response across all trials for each task condition was removed from each trial. Spike count correlations were then computed across trials, using groups of 25 trials as described above (Figure 5.2e-f).

The difference in task versus rest neural correlations was calculated using a two-way, paired t-test (paired by each bin of 25 trials) for each subject separately using Fisher's z-transformed correlation values. The global neural correlation change was computed by averaging the Fisher's z-transformed correlation values between all pairs of cortical regions, and comparing the averaged task versus rest correlation for each bin (Figure 5.2d). Statistics for the pairwise neural correlation change (Figure 5.2e-g) were corrected for multiple comparisons using an FDR-corrected threshold of $p < 0.05$.

In addition, we computed the spike count correlation during the ITI and task cue period separately within trial (across time points) (Supplementary Figure C.2). This analysis demonstrated that moment-to-moment correlations (rather than trial-to-trial correlations) were also quenched from rest to task periods, suggesting that correlation quenching also occurs at faster timescales. The statistical difference in task versus rest neural correlations was computed by using a two-way, paired t-test (paired by trial) across all trials for each monkey separately. The global neural correlation change was computed by averaging the correlation across all pairs of recording areas for each trial (Supplementary Figure C.2).

5.3.4 fMRI: Data and paradigm

The present study was approved by the Rutgers University institutional review board. Data were collected as part of the Washington University-Minnesota Consortium of the Human Connectome Project (HCP) [Van Essen et al., 2013]. A subset of data (n = 352) from the HCP 1200 release was used for empirical analyses. Specific details and procedures of subject recruitment can be found in [Van Essen et al., 2013]. The subset of 352 participants was selected based on: quality control assessments (i.e., any participants with any quality control flags were excluded, including 1) focal anatomical anomaly found in T1w and/or T2w scans, 2) focal segmentation or surface errors, as output from the HCP structural pipeline, 3) data collected during periods of known problems with the head coil,

4) data in which some of the FIX-ICA components were manually reclassified; low-motion participants (i.e., exclusion of participants that had any fMRI run in which more than 50% of TRs had greater than 0.25mm framewise displacement); removal according to family relations (unrelated participants were selected only, and those with no genotype testing were excluded). A full list of the 352 participants used in this study will be included as part of the code release.

All participants were recruited from Washington University in St. Louis and the surrounding area. We split the 352 subjects into two cohorts of 176 subjects: an exploratory cohort (99 females) and a replication cohort (84 females). The exploratory cohort had a mean age of 29 years of age (range=22-36 years of age), and the replication cohort had a mean age of 28 years of age (range=22-36 years of age). All subjects gave signed, informed consent in accordance with the protocol approved by the Washington University institutional review board. Whole-brain multiband echo-planar imaging acquisitions were collected on a 32-channel head coil on a modified 3T Siemens Skyra with TR=720 ms, TE=33.1 ms, flip angle=52°, Bandwidth=2,290 Hz/Px, in-plane FOV=208x180 mm, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. Data for each subject were collected over the span of two days. On the first day, anatomical scans were collected (including T1-weighted and T2-weighted images acquired at 0.7 mm isotropic voxels) followed by two resting-state fMRI scans (each lasting 14.4 minutes), and ending with a task fMRI component. The second day consisted with first collecting a diffusion imaging scan, followed by a second set of two resting-state fMRI scans (each lasting 14.4 minutes), and again ending with a task fMRI session. Each of the seven tasks was collected over two consecutive fMRI runs. The seven tasks consisted of an emotion cognition task, a gambling reward task, a language task, a motor task, a relational reasoning task, a social cognition task, and a working memory task. Briefly, the emotion cognition task required making valence judgements on negative (fearful and angry) and neutral faces. The gambling reward task consisted of a card guessing game, where subjects were asked to guess the number on the card to win or lose money. The language processing task consisted of interleaving a language condition, which involved answering questions related to a story presented aurally, and a math condition, which involved basic arithmetic questions presented aurally. The motor task involved asking subjects to either tap their left/right fingers, squeeze their left/right toes, or move their tongue. The reasoning task involved asking subjects to determine whether two sets of objects differed from each other in the same dimension (e.g., shape or texture). The social cognition task was a theory of mind task, where objects (squares, circles, triangles) interacted with each other in a video clip, and subjects were subsequently asked whether the objects interacted in a social manner. Lastly, the working memory task was a variant of the N-back task.

Further details on the resting-state fMRI portion can be found in [Smith et al., 2013], and additional details on the task fMRI components can be found in [Barch et al., 2013]. All fMRI results reported in the main article reflect results found with the first cohort of subjects. Independent replication of these effects are reported in Supplementary Figure C.4 with the replication cohort.

5.3.5 fMRI: Preprocessing

Minimally preprocessed data for both resting-state and task fMRI were obtained from the publicly available HCP data. Minimally preprocessed surface data was then parcellated into 360 brain regions using the [Glasser et al., 2016a] atlas. We performed additional standard preprocessing steps on the parcellated data for resting-state fMRI and task state fMRI to conduct neural variability and FC analyses. This included removing the first five frames of each run, de-meaning and de-trending the time series, and performing nuisance regression on the minimally preprocessed data [Ciric et al., 2017]. Nuisance regression removed motion parameters and physiological noise. Specifically, six primary motion parameters were removed, along with their derivatives, and the quadratics of all regressors (24 motion regressors in total). Physiological noise was modeled using aCompCor on time series extracted from the white matter and ventricles [Behzadi et al., 2007]. For aCompCor, the first 5 principal components from the white matter and ventricles were extracted separately and included in the nuisance regression. In addition, we included the derivatives of each of those components, and the quadratics of all physiological noise regressors (40 physiological noise regressors in total). The nuisance regression model contained a total of 64 nuisance parameters. This was a variant of previously benchmarked nuisance regression models reported in [Ciric et al., 2017].

We excluded global signal regression (GSR), given that GSR artificially induces negative correlations [Murphy et al., 2009, Power et al., 2014], which would bias analyses of the difference of the magnitude of correlations between rest and task. We included aCompCor as a preprocessing step here given that aCompCor does not include the circularity of GSR (regressing out some global gray matter signal of interest) while including some of the benefits of GSR (some extracted components are highly similar to the global signal) [Power et al., 2018]. This logic is similar to a recently-developed temporal-ICA-based artifact removal procedure that seeks to remove global artifact without removing global neural signals, which contains behaviorally relevant information such as vigilance [Wong et al., 2013, Glasser et al., 2018]. We extended aCompCor to include the derivatives and quadratics of each of the component time series to further reduce artifacts. Code to perform this regression is publicly available online using python code (version 2.7.15) (`https://github.com/ito-takuya/fmriNuisanceRegression`).

Task data for task FC analyses were additionally preprocessed using a standard general linear model (GLM) for fMRI analysis. For each task paradigm, we removed the mean evoked task-related activity for each task condition by fitting the task timing (block design) for each condition using a finite impulse response (FIR) model [Cole et al., 2019]. (There were 24 task conditions across seven cognitive tasks.) We used an FIR model instead of a canonical hemodynamic response function given recent evidence suggesting that the FIR model reduces both false positives and false negatives in the identification of FC estimates [Cole et al., 2019]. This is due to the FIR model's ability to flexibly fit the mean task-evoked response across all blocks. Removing the mean-evoked response of a task condition (i.e., main effect of task) is critical to isolate the spontaneous neural activity (and similarly the background connectivity [Norman-Haignere et al., 2012]). Importantly, this procedure is standard when performing in spike count correlations [Cohen and Kohn, 2011, Cole et al., 2019]. Analogous statistical preprocessing steps were critical when comparing neural correlation measures across human fMRI data and NHP spiking data.

FIR modeled task blocks were modeled separately for task conditions within each of the seven tasks. Thus, the mean task-evoked activation was differentially accounted for according to each specific task condition. In particular, two conditions were fit for the emotion cognition task, where coefficients were fit to either the face condition or shape condition. For the gambling reward task, one condition was fit to trials with the punishment condition, and the other condition was fit to trials with the reward condition. For the language task, one condition was fit for the story condition, and the other condition was fit to the math condition. For the motor task, six conditions were fit: (1) cue; (2) right hand trials; (3) left hand trials; (4) right foot trials; (5) left foot trials; (6) tongue trials. For the relational reasoning task, one condition was fit to trials when the sets of objects were matched, and the other condition was fit to trials when the objects were not matched. For the social cognition task, one condition was fit if the objects were interacting socially (theory of mind), and the other condition was fit to trials where objects were moving randomly. Lastly, for the working memory task, 8 conditions were fit: (1) 2-back body trials; (2) 2-back face trials; (3) 2-back tool trials; (4) 2-back place trials; (5) 0-back body trials; (6) 0-back face trials; (7) 0-back tool trials; (8) 0-back place trials. Since all tasks were block designs, each time point for each block was modeled separately for each task condition (i.e., FIR model), with a lag extending up to 25 TRs after task block offset.

5.3.6 fMRI: Task state activation analysis

We performed a task GLM analysis on fMRI task data to evaluate the task-evoked activity. The task timing for each of the 24 task conditions was convolved with the SPM canonical hemodynamic response function to obtain task-evoked activity estimates for each task condition separately [Friston et al., 1994]. FIR modeling was not used when modeling task-evoked activity. Coefficients were obtained for each parcel in the Glasser et al. (2016) cortical atlas for each of the 24 task conditions.

5.3.7 fMRI: Task state versus resting-state variability analysis

To compare task state versus resting-state variability, we regressed out the exact same task design matrix used on task-state regression on resting-state data. This was possible given that the number of timepoints of the combined resting-state scans in the HCP data set exceeded the number of timepoints of the combined task-state scans (4800 resting-state TRs > 3880 task-state TRs). This step was to ensure that any spurious change induced through the removal of the mean task-evoked response would also induce spurious changes in the resting-state data. However, results were qualitatively identical without the regression of the task design matrix on resting-state data.

After task regression, we obtained the residual time series for both resting-state and task state fMRI data. We then z-normalized each task run with zero-mean and unit variance such that we could appropriately compare the neural variability of task blocks across different runs. We emphasize that task activation regression (removal of the mean task-evoked response) was removed prior to z-scoring the time series. Additionally, Supplementary Figure C.5 shows results without z-normalization, and the results are qualitatively identical.) This enabled us to evaluate whether the variability during task blocks significantly decreased relative to inter-block intervals by evaluating the variance of task blocks relative to 1. We then extracted the time series variance during task blocks, and then averaged the variance across all task conditions to obtain our statistic of task-evoked neural variability. To identify the resting-state neural variability, we applied the same exact procedure to resting-state time series using the task-state design matrix. A sanity check for our analysis was that the 'intrinsic-state' neural variability is close to 1 (given that the time series was normalized to have unit variance), while the task-state neural variability is significantly less than 1 (Figure 5.3a). This ensured that variability measures were not biased by the normalization step.

We compared the neural variability of the entire brain during task state periods versus resting-state periods. For each subject, we computed the variance during task and resting state separately, and then averaged across all brain regions. This resulted in two values per subject, representing task state and resting-state variability. We then performed a two-way group paired t-test across subjects to assess statistical significance (Figure 5.3a). We also computed the task state versus resting-state difference in neural variability for each brain region separately (Figure 5.3b). We corrected for multiple comparisons using an FDR-corrected threshold of $p < 0.05$ (Figure 5.3b,c). Cortical surface visualizations were constructed using Connectome Workbench (version 1.2.3) [Van Essen et al., 2013].

5.3.8 fMRI: Task state versus resting-state correlation analysis

We compared task-state versus resting-state FC (i.e., neural correlations), after performing the exact same preprocessing steps as mentioned above. Results without z-normalization (and using covariance rather than correlations) on the task and rest residual time series are reported in Supplementary Figure C.5.

We computed the correlation between all pairs of brain regions for each task condition during task block periods. We then averaged the Fisher's z-transformed correlation values across all task conditions to obtain a general task state FC matrix (Figure 5.3e). We repeated the same procedure (i.e., using the same task-timed blocks) on resting-state FC to obtain an equivalent resting-state FC matrix for each subject (Figure 5.3d). We directly compared task-state FC to resting-state FC by performing two-way group paired t-tests for every pair of brain regions using the Fisher's z-transformed correlation values. Statistical significance was assessed using an FDR-corrected threshold of $p < 0.05$ (Figure 5.3f). To compare the average global correlation during task state and resting state, we computed the average correlation between all pairs of brain regions during task and resting-state, performing a group paired t-test (Figure 5.3g). To compare the average global connectivity profile of every brain region [Cole et al., 2010b], we computed the average Fisher z-transformed correlation of a single region to all other brain regions during task and rest and performed a two-way group paired t-test between task and rest (Figure 5.3h,i). Statistical significance was assessed using an FDR-corrected threshold of $p < 0.05$.

5.3.9 fMRI: Task state versus resting-state variability/correlation analysis without task regression

To compare task-state versus resting-state variability/correlations without regressing out task effects using FIR [Cole et al., 2019], we calculated the variance/correlations for each time point across blocks. This approach is similar to previous studies that measured variability changes after task/stimulus onset [Churchland et al., 2010, He, 2013]. Importantly, because variance/correlations explicitly account for the mean across a sample, and variance/correlations are computed for each time point separately, this approach accurately accounts for task-locked effects.

Statistics (i.e., variance/correlations) were calculated across blocks at each time point, for each condition separately. To accurately compare task-state to resting-state statistics, we computed cross-block statistics for rest data using the same task block design (i.e., sham/control blocks). This controlled for the number of task blocks and temporal spacing between blocks. We included the first 15 time points following block onset for both the rest and task data. Thus, any task blocks that contained fewer than 15 time points were excluded. This was performed for all ROIs for every subject. Summary statistics were aggregated across ROIs, task conditions and time points (within rest or task states) and visualized in Figure 5.4.

(For this analysis, we used minimally preprocessed data (from the HCP). Additional nuisance regression was performed for both rest and task data as described above, excluding task regression.)

5.3.10 Information-theoretic analysis

We evaluated the information-theoretic relevance of rest and task states by characterizing the dimensionality of neural activity. To estimate the statistical dimensionality of neural data, we used the 'participation ratio', as previously described in [Litwin-Kumar et al., 2017]. We first obtain the covariance matrix W of activity for rest and task states separately. We then calculated

$$dim_W = \frac{(\sum_i^m \lambda_i)^2}{\sum_i^m \lambda_i^2} \tag{5.2}$$

Where dim_W corresponds to the statistical dimensionality of W, and λ_i corresponds to the eigenvalues of the covariance matrix W. Intuitively, this is related to finding the number of components needed to explain variance greater than some fixed threshold, with more needed components reflecting a higher dimensionality of the data.

For human fMRI data, we estimated the task-state and resting-state dimensionality by calculating the whole-brain covariance matrix for each state. For task state this was done by estimating the covariance matrix using task block periods. For resting state this was done by calculating the covariance matrix across the equivalently lengthed resting-state periods (using the same data in the FC analysis above). We applied equation 5.2 to the task-state and resting-state covariance matrix for each subject. Finally, we applied a two-way, group paired t-test comparing the dimensionality of task-state activity to resting-state activity (Figure 5.5a). We replicated this finding in the replication cohort. In addition, we performed this analysis for each fMRI task separately (Supplementary Figure C.10).

For NHP spiking data, we estimated the task (task cue period) and rest (ITI) dimensionality by calculating the covariance matrix between all pairs of population recordings. We then applied equation 5.2 to task and rest periods for each covariance matrix. (Each covariance matrix was calculated using bins of 25 consecutive trials.) Finally, we applied a two-way group paired t-test (across bins) comparing the dimensionality of task activity to rest activity (Figure 5.5b). We replicated this effect in the held-out second monkey.

5.3.11 Spiking model: Estimating the transfer function of a neural population with a balanced spiking model

Our goal was to evaluate the effects of evoked activity across large neural populations, rather than within populations. Thus, we first estimated the transfer function of a neural population using a previously established balanced neural spiking model, with 4000 excitatory and 1000 inhibitory units [Litwin-Kumar and Doiron, 2012]. All parameters are taken directly from [Litwin-Kumar and Doiron, 2012] with the description paraphrased below. Units within the network were modeled as leaky integrate-and-fire neurons whose membrane voltages obeyed the equation

$$\frac{dV}{dt} = \frac{1}{\tau}(\mu - V) + I_{syn} \tag{5.3}$$

where τ indicates the membrane time constant, μ is the bias term, and I_{syn} is the synaptic input. When neurons reached $V_{th} = 1$ a spike was emitted, and voltages were reset to $V_{re} = 0$ for an absolute refractory period of 5ms. τ was 15ms and 10ms for excitatory and inhibitory neurons, respectively. For excitatory neurons, μ was randomly sampled from a uniform distribution between 1.1 and 1.2. For inhibitory neurons, μ was randomly sampled from a uniform distribution between 1 and 1.05.

Synapses to a neuron were modeled as the sum of excitatory and inhibitory synaptic trains x_E and x_I, respectively, and was calculated as the normalized difference of exponentials describing the synaptic rise and decay times caused by each presynaptic event. This effectively captured the weighted effect of all presynaptic neurons to a target neuron, and specifically obeyed the equations

$$I_{y,syn} = x_E(t) + x_I(t) \tag{5.4}$$

$$x_Z(t) = \frac{x_{Z,decay} - x_{Z,rise}}{\tau_{Z,decay} - \tau_{Z,rise}}, \quad Z \in \{E, I\} \tag{5.5}$$

where the synaptic rise and decay of x_E and x_I was modeled as the first order differential equations

$$\frac{dx_{Z,decay}}{dt} = \sum_j J_{ij} s_j - \frac{x_{Z,decay}}{\tau_{Z,decay}} \tag{5.6}$$

$$\frac{dx_{Z,rise}}{dt} = \sum_j J_{ij} s_j - \frac{x_{Z,rise}}{\tau_{Z,rise}} \tag{5.7}$$

J_{ij} refers to the synaptic weight from neuron j to i, s_j indicates whether neuron j emitted a spike. Synaptic rise times were the same for excitatory and inhibitory neurons, with $\tau_{E,rise} = \tau_{I,rise} = 1$ms, while $\tau_{E,decay} = 3$ms and $\tau_{I,decay} = 2$ms. Connection probabilities p^{xy} from neurons in population y to x were $p^{EI} = p^{IE} = p^{II} = 0.5$, and on average, $p^{EE} = 0.2$. However, if two neurons were both excitatory and belonged to the same cluster, the connection strength was multiplied by 1.9. (We employed only the homogenous clustered networks, as described by [Litwin-Kumar and Doiron, 2012], with parameters $J^{EE} = 0.024$, $J^{EI} = -0.045$, $J^{IE} = 0.014$, and $J^{II} = -0.057$. Excitatory stimulation was performed by increasing μ to the first 400 excitatory neurons by 0.5 in 0.05 increments. Inhibitory stimulation was performed by decreasing μ by 0.5 in 0.05 increments to 400 inhibitory neurons.

To estimate the population transfer function, we simulated 30 trials lasting 2s each at each stimulation amplitude. Spike train statistics were estimated across trials in 50ms sliding windows with 10ms shifts. Only excitatory neurons were included when calculating the population spike train statistics (i.e., mean and variance at each stimulation amplitude).

Model code was originally adapted from [Litwin-Kumar and Doiron, 2012], and was simulated with Julia (version 1.1.1).

5.3.12 Model: One-dimensional minimal network model

We use the simplest model to mathematically characterize the relationship between evoked activity and neural variability: a one-dimensional mean-field model. We used Wilson-Cowan-type firing rate dynamics to simulate neural population activity [Wilson and Cowan, 1972]. Specifically, our population's activity obeyed the equation

$$\tau_i \frac{dx_i}{dt} = -x_i + f(w_{ii}x_i + b_i + s_i + I) \tag{5.8}$$

where x_i denotes the firing rate (or a measure of activity), $\tau_i = 0.1$ denotes the time constant, $w_{ii} = 1$ refers to the local coupling (auto-correlation), $b_i = -0.5$ refers to the input threshold for optimal activity (or a bias term), s_i refers to the evoked stimulation ($s_i = 0$ for intrinsic activity), I refers to the background spontaneous activity sampled from a Gaussian distribution with mean 0 and standard deviation 0.25, and f is a sigmoid input-output activation function, which is defined as

$$f(x) = \frac{1}{1 + e^{-k*x}} \tag{5.9}$$

where $k = 1$. Numerical simulations were computed using a Runge-Kutta second order method with a time step of dt=10ms [Burden and Faires, 2001]. We simulated neural population activity injecting a fixed input (boxcar input) with amplitudes ranging from $s_i \in [-5, 5]$ in 0.01 increments (Figure 5.7C). Neural variability for each input strength was calculated using the standard deviation of the time series following the input onset and preceding input offset. Each trial was run for 20 seconds. Figure 5.7a was generated using input amplitudes of $s_i \in \{-3, 0, 3\}$.

To visualize the full dynamics of our single neural population, we visualized the one-dimensional phase space (i.e., flow field on a line) [Strogatz, 1994]. In particular, we calculated the flow field by plotting $\dot{x}$ (i.e., $\frac{dx}{dt}$) as a function of x_i (Equation 5.8). Notably, fixed point attractors (equilibrium states) are defined where $\dot{x} = 0$ (Figure 5.7b).

5.3.13 Model: Two-dimensional minimal network model

To characterize the effects of evoked activity on neural correlations, we use a two-dimensional neural population model. We extended the one-dimensional network model to include two neural populations. The network dynamics obeyed the equations

$$\tau_1 \frac{dx_1}{dt} = -x_1 + f(w_{11}x_1 + w_{21}x_2 + b_1 + s_1 + I_1) \tag{5.10}$$

$$\tau_2 \frac{dx_2}{dt} = -x_2 + f(w_{22}x_2 + w_{12}x_1 + b_2 + s_2 + I_2) \tag{5.11}$$

where x_1 and x_2 describe the activity of each population, and all other variables are a described above. Inter-regional coupling was set to be greater than local coupling, given evidence from previous studies that global coupling is greater than local coupling [Deco et al., 2013a, Cole et al., 2016a, Ito et al., 2017]. Specific network parameters for this network model were: $w_{11} = w_{22} = w_{12} = w_{21} = 4$, $b_1 = b_2 = -3$, $\tau_1 = \tau_2 = 0.1$. I_1 and I_2 were sampled from a Gaussian distribution with mean 0 and standard deviation 1. For this network model, we decreased the slope of the sigmoid $k = 0.5$ to allow for a larger dynamic, linear response range.

To quantify the relationship between evoked activity and neural correlations, we systematically simulated the network under different stimulation states (input strengths). Using the same methods as above, we simulated network activity for 50 seconds. We injected fixed input into both neural populations with amplitudes ranging from $s_i \in [-5, 5]$ in 0.01 increments (Figure 5.8e). Notably, given that the injected stimulation is uncorrelated (due to 0-variance in a fixed input), it is non-trivial that the FC between two nodes would change in response to different inputs. Neural correlations were calculated using a Pearson correlation of the two time series following input onset and preceding input offset.

The use of a minimal model constrained our network to two dimensions. This allowed us to leverage dynamical systems tools to visualize the flow field in the two-dimensional phase plane. To identify the fixed point attractors, we first calculated the nullclines for x_1 and x_2. Nullclines are defined as the values of x_1 and x_2 such that $\dot{x}_2 = 0$ and $\dot{x}_1 = 0$, respectively. The fixed points lie at the intersection of the nullclines. For our particular system of equations, the nullclines of x_1 and x_2 were defined, respectively, as

$$x_2 = \frac{f^{-1}(-x_1) - w_{11}x_1 - b_1 - s_1}{w_{21}} \tag{5.12}$$

$$x_1 = \frac{f^{-1}(-x_2) - w_{22}x_2 - b_2 - s_2}{w_{12}} \tag{5.13}$$

where parameters are identical to those used in equations 5.10 and 5.11. However, the background noise, parameter I, was removed when calculating the nullclines. Fixed point attractors (equilibrium states) are defined where $\dot{x} = 0$ (Figure 5.8b). The full flow field was obtained by applying the system of equations (equations 5.10 and 5.11) to every point in the phase space (e.g., all values of x_1 and x_2).

5.3.14 Model: Evaluating fixed point attractor dynamics and the characteristic time scale

Our models accurately demonstrated that evoked activity decreased the neural variability and correlations from a stochastic dynamical network model. Since our network model was governed by firing rate equations which provided us full access to the system's dynamics, we sought to link dynamical mechanisms (in the absence of spontaneous activity) with changes in the descriptive statistics. Such an analysis would provide us with a mechanistic understanding between descriptive neural statistics used in empirical data analysis and the governing neural dynamics.

To understand how attractor dynamics influenced simulated activity in a network model, we characterized the dynamics around the network's fixed point attractor. Specifically, we performed a linear stability analysis around the fixed point (i.e., the equilibrium level of activity the system is drawn to during a particular state or input, e.g., Figure 5.8b) in both the one-dimensional and two-dimensional network models. In the one-dimensional case, this analysis is equivalent to evaluating the first derivative of equation 5.8 at the fixed point (e.g., the slope of the line at the starred locations in Figure 5.7b). We then calculated the *characteristic time scale* T at the fixed point x^* (in one dimension) with the equation

$$T = \frac{1}{|f'(x^*)|} \tag{5.14}$$

where f represents equation 5.8 [Strogatz, 1994]. The characteristic time scale captures the speed with which the system approaches the fixed point attractor. We calculated the characteristic time scale across the same range of evoked stimulation strengths as in the neural variability analysis. Fixed points were computed numerically by running the network model until it reached a steady state in the absence of noise/spontaneous activity.

The characteristic time scale is an established measure for one-dimensional systems. However, we sought to extend the characteristic time scale beyond a single dimension to evaluate shifting attractor dynamics in higher dimensions. We first performed a linear stability analysis in two dimensions by evaluating the Jacobian matrix for our two-dimensional system at the fixed point (x_1^*, x_2^*)

$$J(x_1^*, x_2^*) = \begin{pmatrix} \frac{df_1}{dx_1} & \frac{df_1}{dx_2} \\ \frac{df_2}{dx_1} & \frac{df_2}{dx_2} \end{pmatrix} \tag{5.15}$$

Where f_1 and f_2 refer to the equations governing neural populations 1 and 2 (equations 5.10 and 5.11, respectively). For our particular system of equations, the Jacobian was calculated as

$$J(x_1^*, x_2^*) = \begin{pmatrix} (-1 + f'(w_{11}x_1 + w_{21}x_2 + b_1 + s_1))\frac{1}{\tau_1} & (f'(w_{11}x_1 + w_{21}x_2 + b_1 + s_1))\frac{1}{\tau_1} \\ (f'(w_{22}x_2 + w_{12}x_1 + b_2 + s_2))\frac{1}{\tau_2} & (-1 + f'(w_{22}x_2 + w_{12}x_1 + b_2 + s_2))\frac{1}{\tau_2} \end{pmatrix} \tag{5.16}$$

For each input strength (i.e., differing evoked states), we evaluated the Jacobian at the fixed point attractor. We then calculated the two eigenvalues (denoted λ_1 and λ_2) and eigenvectors (denoted v_1 and v_2) of the Jacobian using an eigendecomposition. To calculate the generalized characteristic time scale in two dimensions, we first calculated the linear combination of the eigenvectors weighted by the real eigenvalues, and computed the magnitude of the vector, such that

$$v_{sum}(x, y) = re(\lambda_1)v_1 + re(\lambda_2)v_2 \tag{5.17}$$

We the define the two dimensional characteristic time scale T as the reciprocal of the magnitude of $v_{sum}(x, y)$, such that

$$T = \frac{1}{|\sqrt{x^2 + y^2}|} \tag{5.18}$$

We calculated T for a range of values s_1, $s_2 \in [-5, 5]$ in 0.01 increments, and correlated T across all values of s_1 and s_2 with the corresponding neural correlations [Strogatz, 1994].

5.3.15 Model: 300 unit firing rate model

To verify the findings observed in our minimal models would scale to larger networks, we included a 300 region mean-field firing rate model. We chose 300 regions given that most whole-brain human atlases contain 200-400 cortically defined parcels [Power et al., 2011, Glasser et al., 2016a, Schaefer et al., 2018]. Our model followed the same equations as in our minimal model, though inter-area weights were appropriately scaled relative regional self-coupling parameters. Specifically, the network dynamics obeyed the equations

$$\tau_i \frac{dx_i}{dt} = -x_1 + f(w_{ii}x_i + \sum_{j \neq i}^{300} w_{ji}x_j + b_i + s_i + I_i) \tag{5.19}$$

where x_i describes the activity of each population, and all other variables are as described above. Inter-regional coupling was set to be greater than local coupling (2:1 ratio), given evidence from previous studies that global coupling is greater than local coupling [Deco et al., 2013a, Cole et al., 2016a, Ito et al., 2017]. Specific parameters for this network model were specified such that: $w_{ii} = 1$, the mean of the inter-region coupling parameters was $\sum_{j \neq i}^{300} w_{ji} = 2$, $b_i = -2$, $\tau_i = 0.1$. I_i was sampled from a Gaussian distribution with mean 0 and standard deviation 1. $s_i = 0$ during rest state and $s_i = 1$ during task state.

In total, we ran simulations for two classes of network models: a network with random connections and a network with clustered communities (Supplementary Figure C.11). For the random network, we randomly sampled connections with

20% probability rate between all pairs of regions. For the clustered network model, we generated 10 communities of 30 nodes each. Regions within the community had a 20% probability rate for establishing a connection. Between-community connections had a 3% probability rate for establishing a connection.

For each class of network (random or clustered), we weighted connections with either positive weights only (i.e., only E connections) or both positive and negative weights (i.e., both E and I connections. For the E-only network, weights were sampled from a normal distribution with parameters $\mu = 1$, $\sigma = 0.2$. For the network with both E and I weights (80% E, 20% I), weights were sampled from a normal distribution with parameters $\mu = 1$, $\sigma = 1.2$.

Both the rest and task state simulation was run for 10 seconds each and sampled at 100ms. For each group analysis (Supplementary Figure C.11), we simulated 20 subjects worth of data. Model code was written in python 3.7.3.

5.3.16 Model: Simulating fMRI BOLD activity

We used the above model to simulate fMRI BOLD activity to demonstrate that changes in neural variability and correlations would extend to fMRI BOLD dynamics (Supplementary Figures C.6-C.7). Neural activity generated from our model simulations was transformed to fMRI BOLD activity using the Balloon-Windkessel model, a nonlinear transformation from neural activity to the fMRI BOLD signal [Buxton et al., 1998, Friston et al., 2003]. Notably, the transformation assumes a nonlinear transformation of the normalized deoxyhemoglobin content, normalized blood inflow, resting oxygen extraction fraction, and the normalized blood volume. All state equations and biophysical parameters were taken directly from [Friston et al., 2003] (equations 4-5). The Balloon-Windkessel model was implemented in Python (version 2.7.13), and the implementation code has been made publicly available on GitHub (`https://github.com/ito-takuya/HemodynamicResponseModeling`).

5.4 Results

We first show empirically that task-evoked activity suppresses neural correlations and variability across large cortical areas in two highly distinct neural data sets: NHP mean-field spiking and human fMRI data (Figure 5.1). This confirmed previous findings showing quenched neural variability during task states in both NHPs and humans [Churchland et al., 2010, He, 2011, He, 2013, Hennequin et al., 2018], while going beyond those previous studies to report globally quenched inter-area task-state neural correlations. In particular, we focused on neural variability and correlation changes across large cortical areas (mean-field) in our electrophysiology data set (rather than between pairs of neurons) given our focus on large-scale neural interactions, and to facilitate a comparison between different correlation approaches (FC in fMRI data and spike count correlation in electrophysiology data). In addition to spatially downsampling our NHP data to evaluate mean-field spike rates in each cortical area, we also temporally downsampled our NHP data to investigate variability and correlation changes across trials (on the order of hundreds of milliseconds), which appropriately matches the sampling rate of our fMRI data (720ms). Moreover, we limited our inferences to neural interactions between cortical areas to simplify the complexity of analyzing spike count correlations between pairs of local neurons with different receptive fields [Ruff and Cohen, 2014, Ruff and Cohen, 2016]. Following our empirical results, we provide a mechanistic framework using computational simulations and detailed dynamical systems analyses to explain the quenching of neural variability and correlations during task-evoked states.

5.4.1 Task onset reduces neural variability and correlations across spiking populations in NHPs

We estimated the spiking variability and spike count correlations of cortical populations in NHPs following task cue onset (task periods) and during the inter-trial intervals (ITI) (rest periods). We found that across trials, global spiking variability and spike count correlations (r_{sc}) decreased during task as compared to rest (exploratory subject, variance diff = -3.12, t(303)=-10.91, $p < 10e - 22$, r_{sc} diff = -0.04, t(303)=-5.20, $p < 10e - 06$; Figure 5.2c,d; replication subject, variance diff = -5.37, t(807)=-13.45, $p < 10e - 35$; r_{sc} diff = -0.04; t(807)=-9.08, $p < 10e - 17$; Supplementary Figure C.1). Variability reductions were also observed using fano factor (rather than variance) at both the mean-field (averaged across neurons; Supplementary Figure C.14) and for the majority of individual neurons in each cortical area (Supplementary Figure C.12 and Supplementary Figure C.13). Correlated variability reductions were also observed using spike count covariance (rather than correlations) (Figure 5.2h). In addition, we demonstrated that variability and correlation decreased within trial (across time points within a trial, after removing the mean task-evoked response), demonstrating that task state quenching also occurs on a moment-to-moment basis, rather than only on a slower trial-to-trial timescale (Supplementary Figure C.2). We also measured the spiking variability for each cortical area separately, finding that 5/6 cortical areas reduced their spiking variability during task states in the exploratory subject (all areas except for MT, FDR-corrected $p < 0.05$). In the replication subject, all cortical areas, including MT, reduced their spiking variability (FDR-corrected $p < 0.0001$). Similarly, we found that during task states, the spike count correlation significantly decreased between a majority of cortical areas (FDR-corrected $p < 0.05$; Figure 5.2d-g).

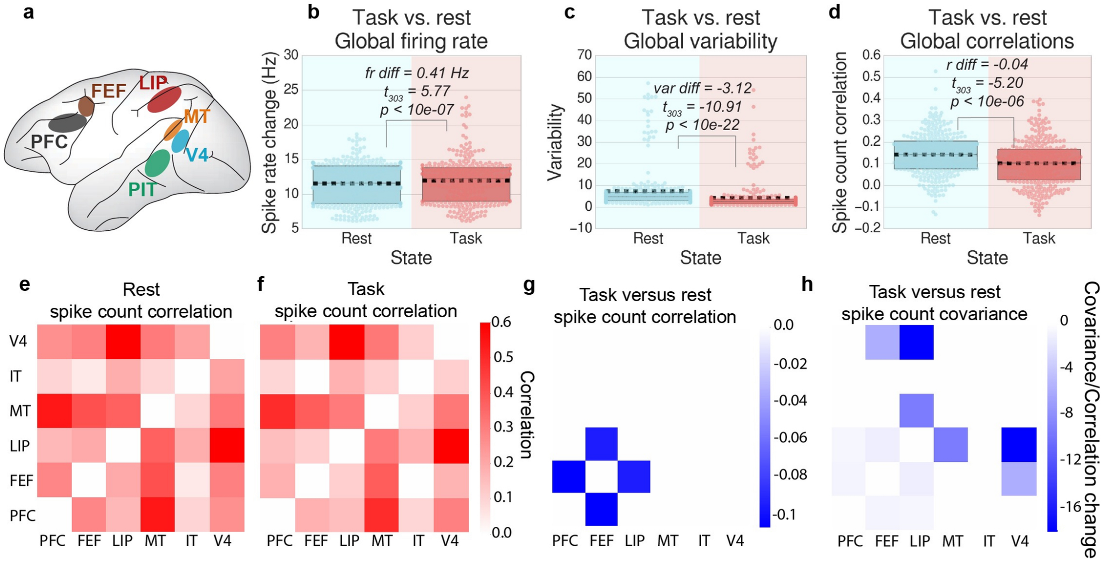


Figure 5.2: Neural variability and correlations decrease during task states relative to rest in spiking data. Results for the replication subject are reported in Supplementary Figure C.1. a) We measured mean-field spike recordings from six different cortical areas during a motion-color categorization task. b) We calculated the average spike rate across all recordings during the rest period (ITI) and task period (task cue), across trials. Each data point reflects the firing rate across 25 consecutive trials. c) We calculated the cross-trial spiking variance for each region during task and rest states, and then averaged across all regions. Each data point reflects the spiking variance across 25 consecutive trials. d) We calculated the average cross-trial neural correlation for task and rest states between all pairs of recorded brain regions. (Spike rates were averaged within each cortical area.) Each data point reflects the correlation across 25 consecutive trials. e-g) For each pair of brain regions, we visualize the correlation matrices between each recording site for the averaged rest period, task period, and the differences between task versus rest state spike count correlations. h) We also observed no increases in covariance (non-normalized correlation) [Siegel et al., 2012, Cole et al., 2016b, Duff et al., 2018]. For panels e-h, plots were thresholded and tested for multiple comparisons using an FDR-corrected $p < 0.05$ threshold. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot. Scatter plot visualizations of b-d can be found in Supplementary Figure C.15

We did not identify any pairwise correlation and covariance increases in our exploratory NHP (Figure 5.2g). However, in our replication NHP we found correlation increases between visual and frontal areas (i.e., MT/IT and PFC/FEF) (Supplementary Figure C.1). When analyzed with covariance (rather than correlation), we found these covariance increases to be weak relative to the observed covariance decreases. (Moreover, the baseline correlation strength between these areas was very low during the ITI period.) Though these correlation increases were only observed in 1 of 2 NHPs, they were generally consistent with our fMRI data (below), which showed that though there were few correlation increases, variability and correlations across cortex were dominated by decreases during task states

To ensure that correlation and variability decreases were associated with increases in the mean activity (rather than just the task period), we estimated the mean spike rate across all regions during the task cue interval and the preceding ITI. Indeed, we found that the mean spike rate during task states was significantly greater than the mean spike rate during rest (exploratory subject, task vs rest firing rate difference = 0.41 Hz, t(303)=5.77, $p < 10e - 06$, replication subject, rate difference = 0.50Hz, t(807)=3.93, $p < 10e - 04$). These findings suggest that task states increase neural activity while quenching spiking variability and spike count correlations across large cortical areas.

Importantly, to accurately dissociate first order statistical effects (mean) from second order effects (variance and covariance/correlation), we removed the cross-trial, mean-evoked response for each task condition. This essential step, which removes the main effect of task, is standard procedure in the spike count (noise) correlation literature [Aertsen et al., 1989]. This procedure isolated the underlying spontaneous/background neural activity during task states, which was subsequently used to infer neural interaction through spike count correlation analysis [Cohen and Kohn, 2011]. To ensure consistency between our spiking and fMRI analysis, it was critical that we also carefully removed the mean-evoked response associated with task blocks in our fMRI data (i.e., the main effect of task; see Methods) [Cole et al., 2019]. To maintain additional consistency between task and rest states in both data sets, we applied the same statistical procedure to our rest data (for both spiking and fMRI data) to control for the possibility that our findings were associated with artifacts related to this procedure (see Methods). (However, we note that the "mean task effects" removed as a result from this step during rest periods were negligible.)

5.4.2 Task-state variability is globally quenched across a wide battery of tasks in human fMRI data

Consistent with the spiking literature, previous work in the fMRI literature has demonstrated that increased activity associated with task-evoked states quenches neural variability [Churchland et al., 2010, He, 2011, He, 2013]. We extended those findings to evaluate variability quenching across seven additional cognitive tasks in humans using data from the Human Connectome Project (HCP) [Van Essen et al., 2013]. We calculated the variability (estimated using time series variance) during task blocks, averaged across tasks and across regions. Consistent with previous reports, we found that the global variability during task blocks was significantly lower than the variability during equivalent periods of resting-state activity (exploratory cohort variance difference = -0.019, t(175)=-23.89, $p < 10e-56$; replication cohort variance difference = -0.019, t(175)=-20.72, $p < 10e-48$; Figure 5.3a). These findings suggest that task states are associated with task-evoked variability reductions.

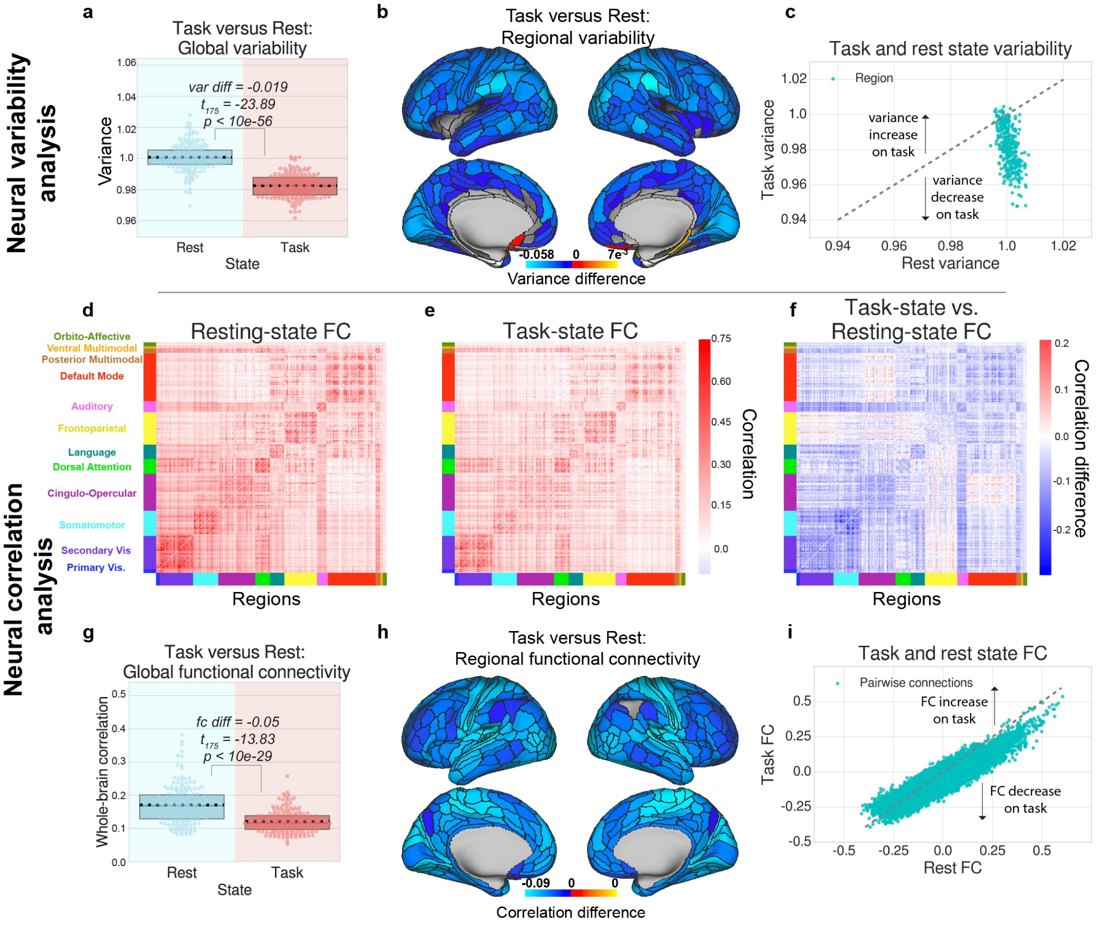


Figure 5.3: Variability and correlations decrease during task states in human fMRI data. Figures for the replication cohort are in Supplementary Figure C.4. Figures for each task separately are shown in Supplementary Figure C.8 and Supplementary Figure C.9. a) We first compared the global variability during task and rest states, which is averaged across all brain regions, and then b) computed the task- versus rest-state variability for each brain region. c) Scatter plot depicting the variance of each parcel during task states (y-axis) and rest states (x-axis). Dotted grey line denotes no change between rest and task states. d) We next compared the correlation matrices for resting state blocks with (e) task state blocks, and (f) computed the task- versus rest-state correlation matrix difference. g) We found that the average FC between all pairs of brain regions is significantly reduced during task state. h) We found that the average correlation for each brain region, decreased for each brain region during task state. i) Scatter plot depicting the FC (correlation values) of each pair of parcels during task states (y-axis) and rest states (x-axis). Dotted grey line denotes no change between rest and task states. For panels b-f, and h, plots were tested for multiple comparisons using an FDR-corrected $p < 0.05$ threshold. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot.

To better understand how global this phenomenon was, we plotted the change in variability from rest to task for each brain region separately. We found that almost all brain regions significantly reduced their variability from rest to task, suggesting that variability reduction occurs across most brain regions (cortical maps are thresholded using an FDR-corrected threshold of $p < 0.05$; Figure 5.3). This finding extends the work of a previous study in human fMRI data during a finger tapping task [He, 2011, Ponce-alvarez et al., 2015], suggesting that task-induced variability reduction is a general phenomenon consistent across most cortical regions, and across a wide variety of cognitive tasks.

Lastly, we evaluated whether variability quenching occurred during task blocks relative to inter-block intervals (rather than comparing task runs to resting-state runs). Since we z-normalized each task run with unit variance, we could evaluate the degree to which variability was quenched during task blocks relative to inter-block intervals by computing the average variance during task blocks relative to 1. (Note that z-normalization of the task time series was performed after removing the mean task-evoked response via a task GLM, such that reduced variability was not an artifact of preprocessing/z-normalizing the time series.) Indeed, we found that the variance during task blocks was reduced relative to the inter-block intervals (exploratory cohort variance - 1 = -0.019, t(175)=-36.58, $p < 10e - 83$; replication cohort variance difference = -0.018, t(175)=-33.01, $p < 10e - 76$). Our findings demonstrate that task-evoked periods quench neural variability relative to both resting-state activity and inter-block intervals.

5.4.3 Task-state FC is globally quenched across a wide battery of tasks in human fMRI data

Despite multiple studies describing task-evoked FC changes [Cole et al., 2014a, Krienen et al., 2014, Gonzalez-Castillo and Bandettini, 2017], the precise mechanisms of how FC can change remain unclear. Our current findings illustrate that mean-field spike count correlations decrease during task-evoked states, consistent with previous literature that focused on local circuits [Churchland et al., 2010, Cohen and Kohn, 2011]. Consistent with the spiking literature's perspective on spike count correlations, and the theoretical evidence suggesting that the correlation of ongoing spontaneous activity should be suppressed during task to facilitate information coding [Averbeck et al., 2006], we hypothesized that FC would also be globally reduced during task states. To ensure consistency in the statistical analysis across spiking and fMRI data, we removed the mean task-evoked response using a finite impulse response (FIR) model. This approach is statistically equivalent to removing the cross-trial mean response of a task condition, and is a critical step when calculating noise correlations in the spiking literature [Cohen and Kohn, 2011]. This step characterizes the correlation of the background spontaneous neural activity (i.e., background connectivity in fMRI), dissociating task-to-neural interactions (main effect of task) from neural-to-neural interactions (FC) [Norman-Haignere et al., 2012].

We first calculated the mean FC across all pairwise correlations across all cortical regions for both task and rest states (Figure 5.3d-f). We found that during task states, the global FC was significantly reduced relative to resting-state fMRI (exploratory cohort FC diff = -0.05, t(175)=-13.83, $p < 10e - 29$; replication cohort FC diff = -0.046, t(175)=-14.00, $p < 10e - 29$; Figure 5.3g). Recent studies have suggested that the use of correlation provides an ambiguous description of how shared variability (relative to unshared variability) change between brain areas [Cole et al., 2016b, Duff et al., 2018]. Thus, to generalize these results, we also calculated FC using covariance rather than correlation, finding that covariance also globally decreases (covariance diff = -192.96, t(351)=-27.30, $p < 10e - 88$; Supplementary Figure C.5). Task-evoked global FC was also reduced in each of the 7 HCP tasks separately (all tasks FDR-corrected $p < 0.0001$; Supplementary Figure C.9). To identify exactly how global this phenomenon was, we plotted the average task versus rest FC change for each brain region (Figure 5.3h,i). We found that nearly all cortical regions significantly reduced their correlation with the rest of cortex during task states. To ensure that correlation differences between rest and task states were not associated with in-scanner head motion, we calculated the average number of motion spikes during rest and task scans using a relative root mean squared displacement threshold of 0.25mm [Ciric et al., 2017]. For both the exploratory and replication cohorts, we found no significant differences in the percentage of motion spikes between rest and task states (exploratory set, average task=0.91% of frames, average rest=0.81% of frames, t(175)=1.08, $p = 0.28$; replication set, average task=0.009% of frames, rest=0.008% of frames, t(175)=1.53, $p = 0.12$).

While we primarily observed global decreases in FC, a small portion of connections increased their FC during task states (exploratory cohort, 7.59% of all connections; replication cohort, 9.07% of all connections; FDR-corrected $p < 0.05$) (Figure 5.3i, Supplementary Figure C.4). However, FC increases were typically limited to cross-network correlations between networks with different functions, where baseline resting-state FC is already quite low (e.g., cingulo-opercular network with the default mode network, or the frontoparietal network with the visual network) (Figure 5.3d-f).

5.4.4 Task state variability and correlation is quenched independently of removing the mean task-evoked response in fMRI data

The above fMRI results employ the use of FIR modeling to remove the mean task-evoked response to compare task- and rest-state correlations/variability. Here we sought to demonstrate that neural variability and correlations are quenched in fMRI data in the absence of any task regression (e.g., FIR modeling). We used an approach that has been previously used to demonstrate variability quenching following task onset, by measuring the cross-trial variance at each time point

[Churchland et al., 2010, He, 2013]. We employ the same general approach, measuring the variance and correlation across blocks for each time point within the block. Moreover, to obtain statistically comparable estimates of resting state variability/correlations, we measured the cross-block variance/correlation during sham blocks during resting state by applying the identical task block structure to resting-state fMRI data. Critically, the removal of the mean task-evoked response was excluded from preprocessing for this analysis, and the time series were not z-normalized.

We found that cross-block variance for time points during task state were significantly reduced relative to resting state (var diff = -1009.56, t(175)=-37.34; $p < 10e{-}84$; Figure 5.4a,b). We also found consistent results for correlations, finding that the cross-block correlation for time points during task state were significantly reduced relative to resting state (r diff = -0.04, t(175)=-10.91, $p < 10e{-}20$; Figure 5.4c,d). These results demonstrate that the quenching of correlations and variability during task states are independent of any potential statistical artifacts that result from removing the mean task-evoked response using FIR task regression.

5.4.5 Task-evoked activity is negatively correlated with neural variability and correlations in human fMRI data

We showed that task states widely reduced neural variability and correlations. We sought to extend this work to directly demonstrate that decreases in neural variability and correlations are associated with changes in task-evoked activity levels. To provide evidence for this hypothesis, we computed the mean task-evoked activity (averaging across all regions). We found that the global activity was significantly greater than baseline across different task states (exploratory

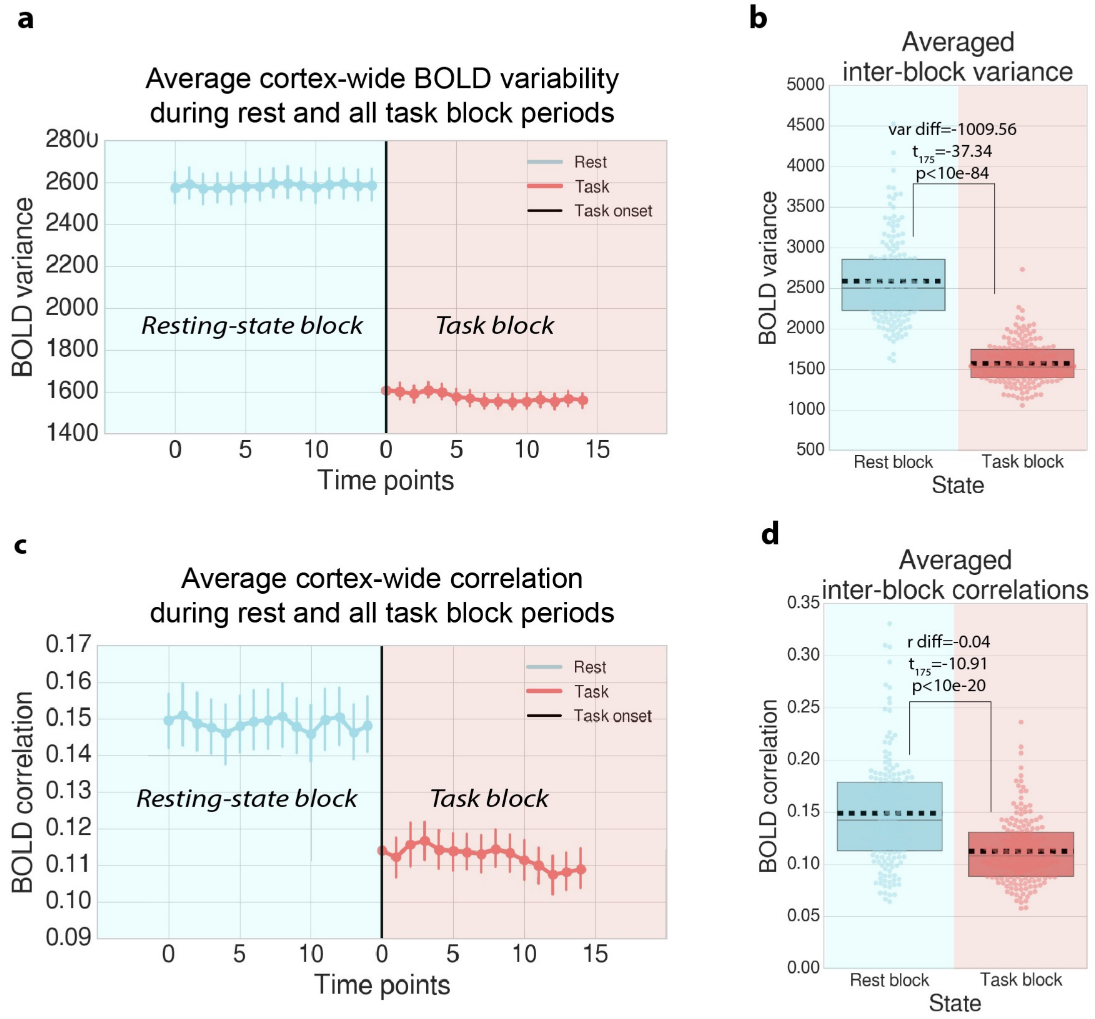


Figure 5.4: Task variability/correlations decrease independently of mean task activity removal step in fMRI data. Instead of computing variance/correlations across time points within task blocks (and removing mean task effects), variance/correlations can be calculated across task blocks (for each time point within a block). This approach isolates ongoing neural activity that is not task-locked, and has been used in both spiking and fMRI data [Churchland et al., 2010, He, 2013]. a) To isolate ongoing spontaneous activity that is not time-locked to the task, we estimated the variance at each time point across task blocks. The variance at each time point was calculated for each ROI and task condition separately, but then averaged across ROIs and task conditions. Note that to obtain an equivalent variance estimate during resting state, we applied an identical block structure to rest data to accurately compare rest to task state variability. Variability across block time points was averaged across brain regions and task conditions. Error bars denote standard deviation across subjects. b) Variance across task block time points was significantly reduced during task blocks relative to identical control blocks during resting-state data. c) We performed a similar procedure for task functional connectivity estimates, correlating across blocks for all pairs of brain regions. Correlations across block time points were averaged for all pairs of brain regions and task conditions. d) Correlations during task state blocks were significantly reduced relative to identical control blocks during resting state. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot.

cohort, t(175)=6.46, $p < 10e-9$; replication cohort, t(175)=12.63, $p < 10e-25$), demonstrating that decreases in neural variability and FC were accompanied by global increases in task-evoked activity.

Previous work has shown that regions that have strong task activations (i.e., the magnitude of the task-induced activation, positive or negative) tend to have greater variability reductions [He, 2013]. (Task activation magnitudes reflect the deflection of the BOLD activity relative to baseline, or the inter-block interval.) We sought to replicate this effect in the current data set, while extending those results to demonstrate that more task-active regions also tend to reduce their FC during task states. We first correlated regional task-evoked activation magnitude with task-evoked variability reduction (task variance minus rest variance) across regions at the group-level. We found that regions with greater task-evoked activation magnitudes (averaged across tasks) exhibited greater variability reductions during task states, confirming previous findings in a finger tapping task (exploratory cohort rho=-0.32, $p < 10e-9$; replication cohort rho=-0.49, $p < 10e-22$; Supplementary Figure C.3a,c) [He, 2013]. This negative relationship was also observed in 6/7 of the HCP tasks when analyzed separately (FDR-corrected $p < 0.01$; Supplementary Figure C.8). To link regional task activation magnitudes with FC decreases, we tested for a correlation between regional task-evoked activation magnitude and the average FC change during task states for each region. Consistent with our hypothesis, we found that regions with greater task-evoked activation magnitudes (averaged across tasks) reduced their average FC more during task states (exploratory cohort rho=-0.25, $p < 10e-05$; replication cohort rho=-0.20, $p = 0.0002$; Supplementary Figure C.3). When tasks were analyzed separately, this negative correlation was observed in 4/7 of the HCP tasks (FDR-corrected $p < 0.05$; Supplementary Figure C.9). Thus, brain areas with higher levels of task-evoked activation magnitudes (i.e., changes in activity relative to baseline) tend to reduce both their task-evoked variability and FC.

5.4.6 The information-theoretic relevance of task state reduction of neural correlations

Results from our empirical data converged across imaging modalities and species, illustrating that task states increased mean activity while reducing neural variability and correlations. However, the theoretical implication of a decreased correlated task state remains unclear. Here we sought to better characterize the information-theoretic implication of a global reduction in neural correlations. In particular, consistent with previous large-scale computational models that have predicted increased dimensionality with stimulus-driven activity [Abbott et al., 2011, Deco et al., 2014], we hypothesized that reductions in neural correlations increase the effective dimensionality across units by suppressing background spontaneous activity/noise. While this increased dimensionality may potentially supports more robust information representations, we acknowledge that a change in neural dimensionality does not necessitate an improvement (or change) in cognitive information representation [Averbeck et al., 2006], and that future studies will need to evaluate the relationship between neural dimensionality and cognitive content. Further, we note that an increase in dimensionality is not trivially implied by decreased global correlations. Because we found that regional time series variance also decreases during task states, the neural data dimensionality would increase only if inter-region covariance decreases more than local regional variance (i.e., off-diagonal is reduced more than the diagonal of the variance-covariance matrix).

We measured the dimensionality using the 'participation ratio' of the neural activity (for human fMRI and NHP spiking data) during rest and task states (see Methods) [Abbott et al., 2011, Litwin-Kumar et al., 2017]. Consistent with our hypothesis, we found that task states increased their overall dimensionality relative to rest states (fMRI task versus rest, exploratory cohort difference =

16.13, t(175) = 19.31, $p < 10e - 44$, replication cohort difference = 15.78, t(175) = 21.66, $p < 10e-50$; NHP task versus rest, exploratory subject difference = 0.13, t(303) = 5.77, $p < 10e - 07$, replication subject difference = 0.26, t(807) = 13.00, $p < 10e-34$) (Figure 5.5). We also found that when analyzing each of the 7 HCP tasks separately, dimensionality increased in all 7 tasks relative to resting state (FDR-corrected $p < 0.0001$; Supplementary Figure C.10). The present results suggest that task states are associated with a decrease in neural variability and correlations, reflecting a suppression of shared and private spontaneous activity, which increases the dimensionality of neural activity.

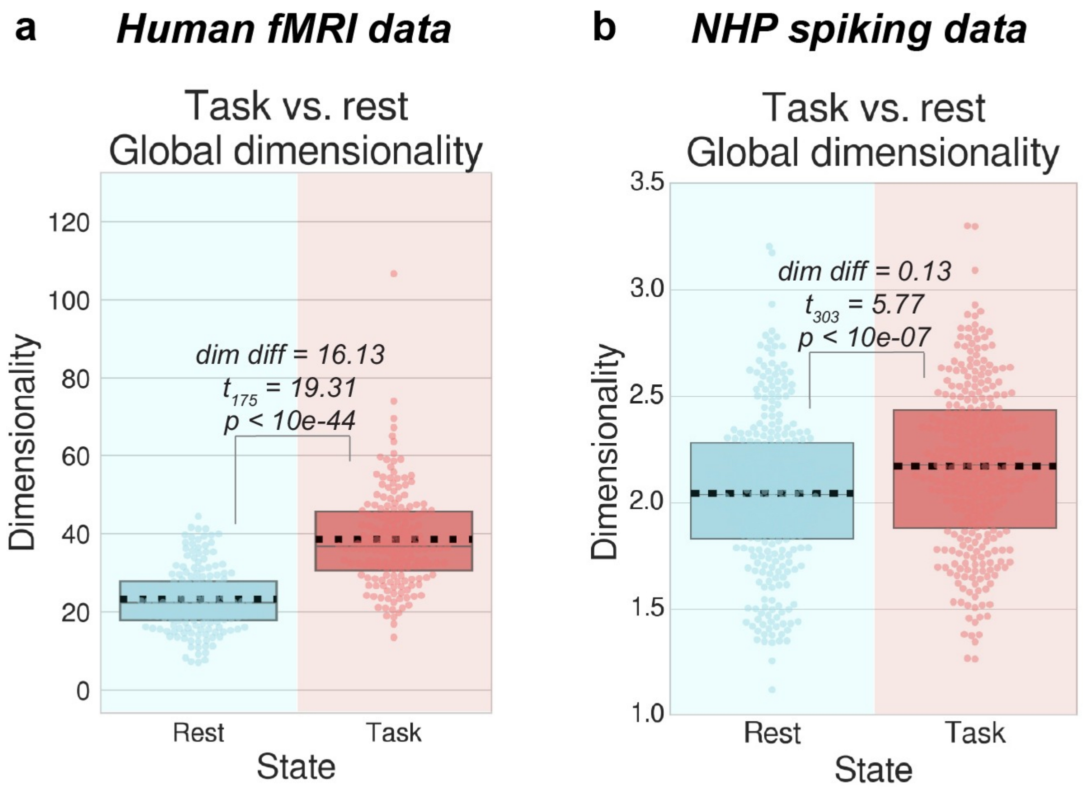


Figure 5.5: Dimensionality increases during task periods relative to resting-state activity. a) For each subject, we calculated the dimensionality using the participation ratio [Abbott et al., 2011, Litwin-Kumar et al., 2017] during task and rest states and found that during task states, dimensionality significantly increased. b) We calculated the dimensionality of spiking activity across trials and found that during task states, dimensionality significantly increased. These findings provide a potential information-theoretic interpretation of neural correlation and variability reduction during task states. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot.

5.4.7 From neurons to neural masses: Modeling neural dynamics of cortical areas

In the previous sections, we provided empirical evidence that task states reduce mean-field inter-area correlations and variability in spike rate and fMRI data. In this section, we construct a biologically plausible model that provides a parsimonious explanation of correlation and variability reductions in mean field spiking networks and cortical BOLD dynamics.

Neurophysiologically, functional brain areas are composed of local circuits with balanced excitatory and inhibitory neural activity (Figure 5.6a). In previous work, local circuits have been demonstrated to have clustered excitatory connections [Song et al., 2005], leading to slow dynamics and high variability in spiking networks simulated in silico [Litwin-Kumar and Doiron, 2012]. Using this previously established model, we systematically perturbed this balanced network under a distribution of inputs (both excitatory and/or inhibitory inputs) to estimate the excitatory output (i.e., mean-field transfer function) of a cortical population. Though most long-range cortical connections are excitatory, we incorporated excitatory and inhibitory stimulation effects on a local population (Figure 5.6b). This is because long-range excitatory afferents may target local inhibitory neurons, producing a net inhibitory effect. Under the presence of inputs, we found that the population transfer function approximated a sigmoid activation function (Figure 5.6b). We note that the upper bound on the sigmoid transfer function (Figure 5.6d) is likely due to inhibitory feedback on excitatory activity rather than the true saturating spiking regime in neurons. This is because excitatory neurons in a local population typically do not reach a saturating spiking regime even for strong visual stimuli [Priebe and Ferster, 2008], and instead reach an upper bound due to strong inhibitory stabilization preventing runaway excitation [Hennequin et al., 2018]. Importantly, simplifying the mean-field transfer function of a cortical area allowed us to focus our modeling efforts on simplified networks across large cortical areas [Joglekar et al., 2018].

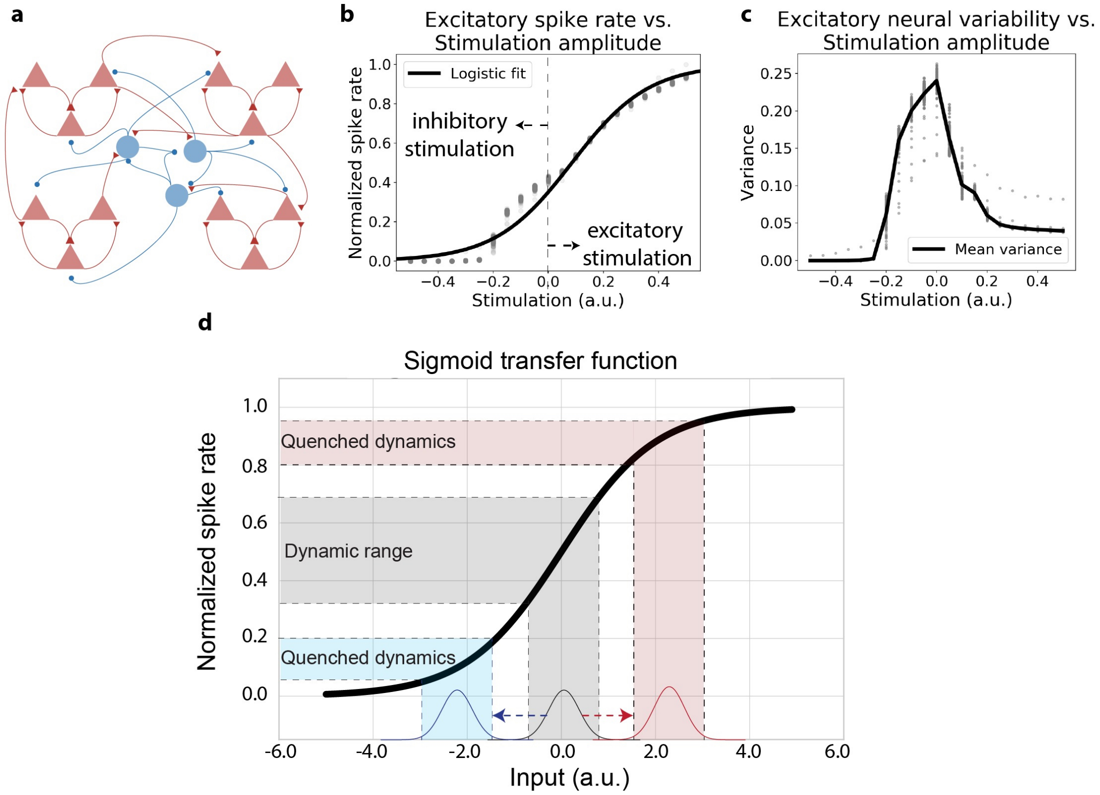


Figure 5.6: Inferring the mean-field transfer function of a neural population with a balanced spiking model with clustered excitatory connectivity. a) Schematic illustration of the balanced spiking model with clustered excitatory connections. Network architecture and parameters are identical to those reported in [Litwin-Kumar and Doiron, 2012]. Red triangles indicate excitatory cells, blue circles indicate inhibitory cells. b) The population spike rate (excitatory cells only) subject to inhibitory regulation. We systematically stimulated a subset of the neural population and measured the corresponding mean excitatory spike rate. Spike rates were normalized between 0 and 1. Excitatory stimulation was implemented by stimulating 400 excitatory neurons, and inhibitory stimulation was implemented by stimulating 400 inhibitory neurons. Spiking statistics were calculated across 30 trials, with each point in the scatter plot indicating a different 50ms time bin. c) Population neural variability (excitatory cells only), as a function of input stimulation. d) Based on panel b, we approximated the mean field neural transfer function as a sigmoid. A sigmoid transfer function produces optimal input-output dynamics for a narrow range of inputs (gray). The same input distribution mean shifted by some excitatory/inhibitory stimulation produces a quenched dynamic range.

In this balanced spiking network, any evoked stimulation, excitatory or inhibitory, would result in reduced variability (Figure 5.6c). Specifically, the magnitude of stimulation was negatively correlated with spiking variability in the balanced spiking model (rho = -0.92; $p < 0.0001$). While previous studies have suggested that the mean and variance of the spike rate may be independent of each other, those studies focused on mean-matching the spike rate of individual neurons within the same local population [Churchland et al., 2010, Litwin-Kumar and Doiron, 2012]. However, in this study, we focus exclusively on the mean-field level rather than individual neurons. We found a highly negative association between mean and variance under experimental perturbation, suggesting that at the mean-field level, mean and variance cannot be mechanistically dissociated. Based on these considerations, we hypothesized that during periods in which global neural activity levels are elevated, such as task states, both neural variability and correlations would be globally quenched.

5.4.8 Neural variability is quenched during task-evoked states in a neural mass model

Here we rigorously ground the intuition that task-evoked activity reduces output variability using neural mass modeling and dynamical systems theory. A recent study provided evidence that an evoked stimulus drives neural populations in sensory cortex around a stable fixed point attractor [Hennequin et al., 2018]. We first extended these findings using a simplified neural mass model, which allows for a comprehensive dynamical systems analysis that is mathematically difficult in higher dimensions. Additionally, this enabled a simpler theoretical approach to investigating changes in neural dynamics that are generalizable across mean-field neural cortical areas (i.e., populations with sigmoidal transfer functions).

We first characterized the relationship between task-evoked and spontaneous activity in a large neural population using a single neural mass unit. We simulated the neural population's dynamics across a range of fixed input strengths (Figure 5.7a), finding a nonlinear relationship between stimulus strength and the observed variability of the neural population (Figure 5.7c).

We found that variability was highest when there was no stimulation, while variability decreased for any type of evoked stimulation (e.g., negative or positive input amplitudes). Despite the model's simplicity, these findings are consistent with our (and others') empirical and model results demonstrating that task states quench time series variability in both human and animal data [Churchland et al., 2010, He, 2011, Hennequin et al., 2018]. We also generalized the findings from our minimal (single region) model to large-scale firing rate models (with 300 regions), where we found variability decreases during task-evoked states in both network models with random structural connections and clustered structural connections (Supplementary Figure C.11). We demonstrated this for network models with excitatory connections only, as well as networks with both excitatory and inhibitory connections. However, due to the large number of possible network models when scaling to n-dimensions, we constrained our analyses to only four network architectures, leaving a more complete analysis to future studies.

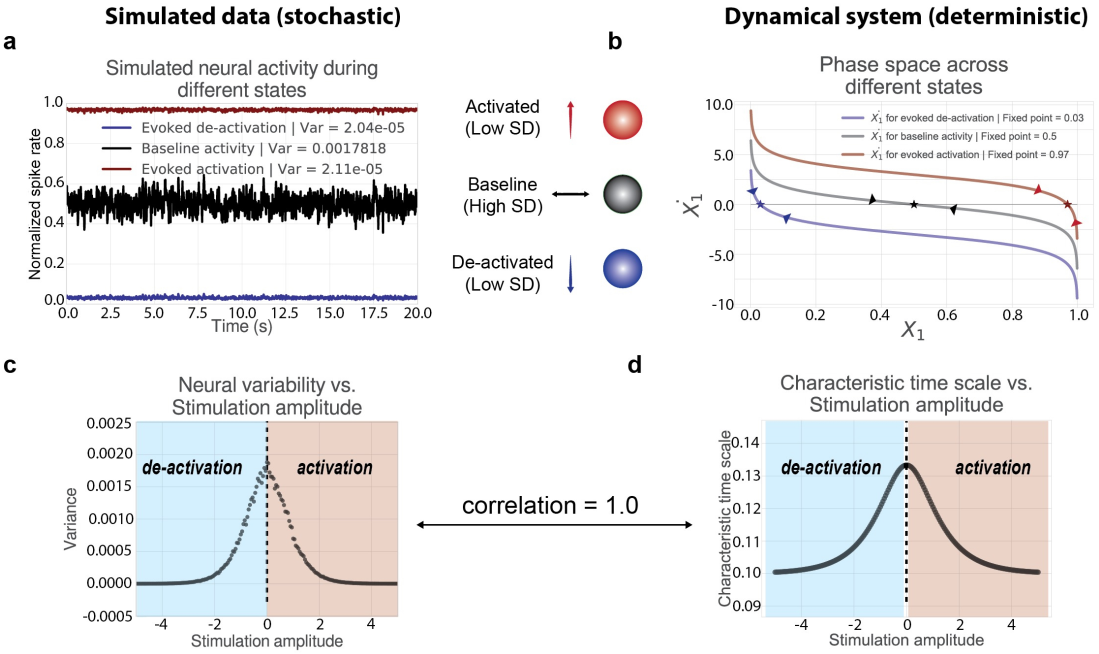


Figure 5.7: Task-evoked activity induces changes in neural variability and the underlying attractor dynamics. Our minimal modeling approach directly links descriptive statistics (e.g., time series variability) with rigorous dynamical systems analysis (e.g., attractor dynamics). a) During different evoked states (i.e., fixed inputs), there is a reduction in the observed time series variability (measured by variance across time). This is directly related to how input-output responses change due to the changing slope in the sigmoid transfer function. b) We visualized the phase space for each of the neural populations according to state by plotting the derivative of X_1 denoted by $\dot{X}_1$. For each state, we estimated the fixed point attractor (plotted as a star), denoting the level of mean activity the system is drawn to given some fixed input (or absence thereof). Arrows denote the direction/vector toward each fixed point, which specify the characteristic time scale (i.e., the speed) the system approaches the fixed point. c) We ran simulations across a range of stimulation amplitudes, calculating the variance across time at each amplitude. d) We characterized the shifting attractor dynamics for each stimulus by computing the characteristic time scale at the fixed point for each stimulation amplitude. The characteristic time scale across all fixed points are nearly perfectly correlated with the neural variability of the simulated time series across all fixed inputs (rank correlation = 0.9996).

We sought to leverage the model's simplicity to characterize dynamical systems properties governing the observed neural variability. This would provide rigorous evidence that shifting the underlying attractor dynamics alters the observed neural signals. We first performed a state space analysis in one dimension to identify the stable fixed point attractor (i.e., the equilibrium level of activity the system is drawn to during a particular state) for the intrinsic and evoked states (Figure 5.7b). The state space view enabled visualization of the system's full dynamics across different evoked states (Figure 5.7b). For example, dynamics around the fixed point attractor in the intrinsic baseline (rest) state appeared to approach equilibrium slowly. This can be identified by observing the angle where the curve intersects 0 on the y-axis (i.e., when $\dot{x} = 0$; Figure 5.7b). The angle of this curve corresponds to the characteristic time scale, a dynamical property characterizing the speed with which the system approaches the attractor (a higher value reflects slower dynamics; see Methods) [Strogatz, 1994].

To quantify this more rigorously, we performed a linear stability analysis around the fixed point attractor of the system across the same range of stimulation amplitudes. For each input, we analytically calculated the characteristic time scale at each fixed point. Again, we found a nonlinear relationship between the amplitude of the stimulus and the characteristic time scale of the neural population (Figure 5.7d), and found that the characteristic time scale explained nearly 100% of the variance (rho=0.9996) of the simulated stimulus-evoked variability (Figure 5.7c). These results demonstrate that changes in observed neural variability can be directly attributed to changes in the underlying attractor dynamics.

To ensure that the model explanation would generalize to data obtained on a slower time scale (e.g., fMRI BOLD data), we transformed the simulated neural activity into fMRI BOLD activity using the Balloon-Windkessel model [Friston et al., 2003]. The Balloon-Windkessel is a nonlinear transformation of neural activity to model the BOLD signal that takes into account the normalized blood volume, blood inflow, resting oxygen extraction fraction, and the normalized deoxyhemoglobin content. Consistent with previous accounts [He, 2013], we found that the characteristic time scale around the fixed point attractor was still strongly correlated with BOLD variability (rho=0.97; $p < 0.0001$; Supplementary Figure C.6).

5.4.9 Neural correlations are quenched during task states in a network model

We generalized the dynamical systems analysis in one dimension to two dimensions, allowing us to focus on correlations across cortical areas. We show illustrations of the state space for intrinsic and task-evoked states (Figure 5.8b,d), as well as the corresponding time series (Figure 5.8a,c) for our model. Induced negative activity produced qualitatively similar results to the activated state (Figure 5.8d) due to subthreshold levels of activity rather than saturating levels of activity.

The state space analysis (Figure 5.8b,d) allowed us to track the simultaneous evolution of the two neural masses, providing a geometric interpretation of the system. We observed qualitatively that shifts in the attractor dynamics (i.e., changes to the flow field) due to stimulation were directly associated with changes to the correlation between the two neural masses. Specifically, we observed that intrinsic state dynamics supported slower, elongated trajectories along a diagonal axis, consistent with correlated neural activity between the two masses (Figure 5.8b). This was due to a slower characteristic timescale near the fixed point attractor, which corresponds mathematically to eigenvalues with smaller magnitudes. In contrast, during evoked states, the system approached the fixed point attractor at a faster speed, quenching trajectories in state space that supported correlated variability (Figure 5.8d). Thus, the visualization of the state space demonstrated that changes in neural correlations were associated with changes to the flow field around the fixed point attractor.

To more carefully test the relationship between state-dependent neural correlations, we simulated our network model across a range of fixed input amplitudes. Despite no changes to the network's connectivity structure, we found that neural correlations systematically changed (decreased) as a function of evoked stimulation (Figure 5.8e). Further, using dynamical systems analysis, we found

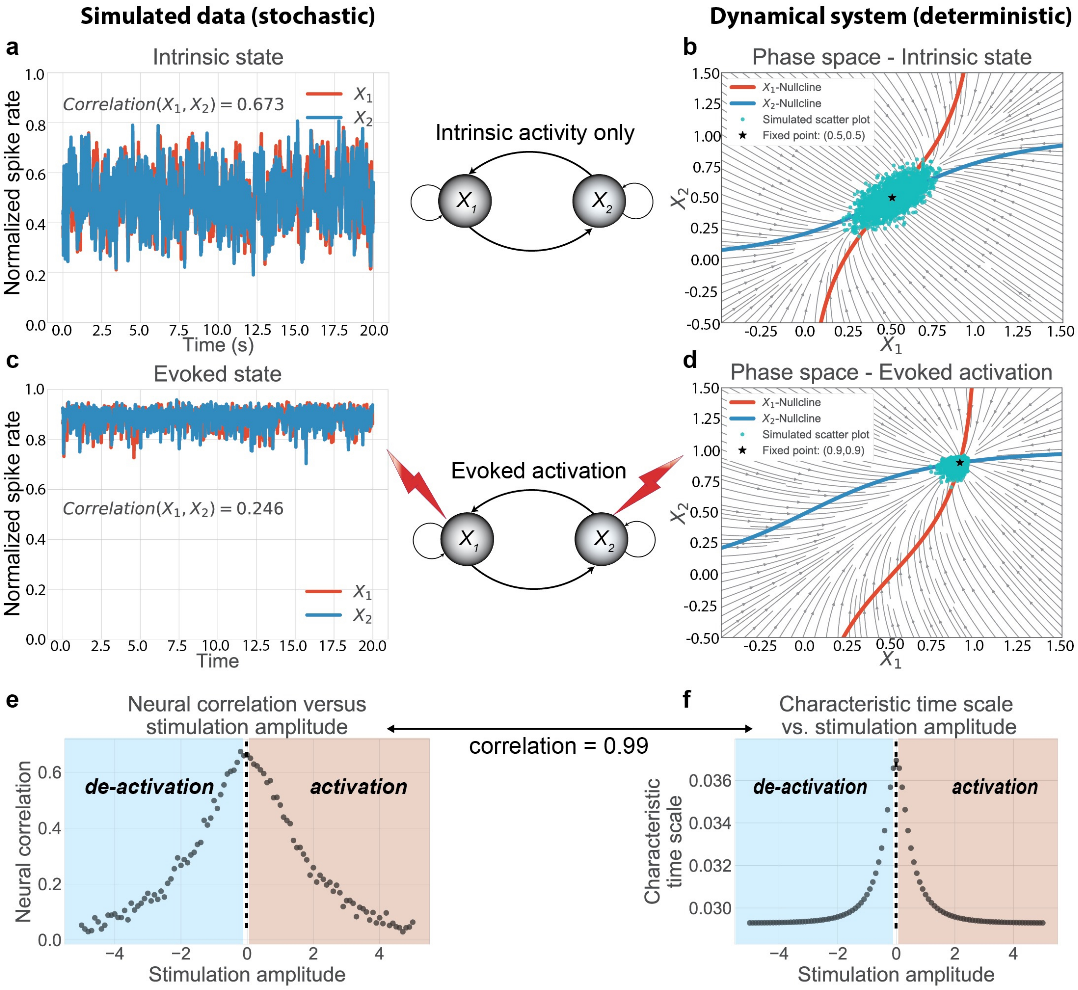


Figure 5.8: Task-evoked activity quenches neural correlations by altering the underlying attractor dynamics. We used a two unit network model, the minimal model necessary to study dynamic changes in neural correlations. a) At baseline, we observed slow, high amplitude fluctuations and high neural correlations. b) To characterize the underlying attractor dynamics, we visualized the two-dimensional state space, visualizing the flow field and the nullclines (blue and red curves, where the rate of change is 0) for each unit. The intersection of the two nullclines denote the fixed point attractor. We overlaid the simulated scatter plot (cyan dots) to illustrate the correspondence between the attractor dynamics and simulation. c) We injected a fixed input stimulation, shifting the network to an 'evoked' state, which caused a decrease in neural variability and correlation. d) The external input transiently moved the fixed point, altering the attractor dynamics and the corresponding scatter plot. e) We systematically injected a range of fixed inputs into the network. We found that neural correlations were optimal with no external stimulation, and decreased with any external stimulation. f) Across stimulation strengths, we found that the generalized characteristic time scale (see Methods) near the fixed point explained 98% of the neural correlation variance, providing a direct association between the network's attractor dynamics and observed neural correlations.

that a generalization of the characteristic time scale in higher dimensions accounted for changes in neural correlations (rho=0.99; $p < 0.0001$; Figure 5.8f). In other words, we analytically determined that evoked stimulation shifted the attractor dynamics, changing the neural correlations in a network model with fixed synaptic connections. We found consistent results after transforming the neural activity to fMRI BOLD activity using the Balloon-Windkessel model [Friston et al., 2003], finding that changes to the characteristic time scale accounted for changes in BOLD correlations (i.e., FC) (rho=0.97; $p < 0.0001$; Supplementary Figure C.7). These results were reproduced using mutual information (rho=0.94; $p < 0.0001$), a nonlinear measure of statistical dependence [MacKay, 2003], and non-parametric rank correlation (rho=0.99; $p < 0.0001$). This suggests that the quenching of shared variance encompasses both parametric and non-parametric linear and nonlinear measures of statistical dependencies.

To ensure that our findings in simplified two node networks would scale to large-scale network models, we simulated large-scale firing rate models (with 300 regions). We found correlation decreases during task-evoked states in both network models with random structural connections and clustered structural connections (Supplementary Figure C.11), suggesting that the mechanisms we identified in these minimal models likely scale to the larger networks. We demonstrated this for network models with excitatory connections only, as well as networks with both E and I connections.

5.5 Discussion

The present results suggest that task-evoked neural activity globally quenches neural time series variability and correlations. We showed this in NHP spiking and human fMRI data, illustrating the generality of the phenomena. This supports the hypothesis that during task states, decreases in neural variability and correlations suppress ongoing spontaneous activity, better supporting information coding [Averbeck et al., 2006]. We subsequently provided a dynamical systems model to demonstrate that evoked activity strengthened the system's fixed point attractor, quenching neural variability and correlations. This provided a mechanistic framework to interpret the empirical results. Importantly, the use of a sigmoid transfer function to model mean-field cortical dynamics revealed a simple interpretation underlying neural variability and correlation suppression widely applicable to many types of neural data. During task states, the slope of the neural transfer function decreases, reducing the dynamic range of input-output responses. This results in reduced overall output variability, as well as reduced shared variability (e.g., correlations) from connected neural populations. The collective empirical and theoretical results provide strong evidence that observed neural variability and correlations are state-dependent, and these changes emerge from the activity dynamics governed by the transfer functions of large neural masses.

The relationship between neural correlations and neural communication (or FC) is complex. For example, it appears that a decrease in the neural correlation between a pair of brain regions does not simply imply a reduction in communication. In the spiking literature, this interpretation is attributed to a reduction of shared spontaneous activity (or, neural noise). This is because the cross-trial mean evoked response (i.e., "the signal" associated with the task or stimulus) is removed prior to calculating the correlation, leaving only "neural noise" or spontaneous, moment-to-moment activity [Cohen and Kohn, 2011]. Notably, this "neural noise" can still be important, since some portion of it drives trial-by-trial variability in cognition and behavior. In the fMRI literature, this is equivalent to regressing out the cross-trial mean task-evoked activity associated with the task/stimulus prior to calculating FC [Cole et al., 2019]. (This type of FC is also referred to as "background connectivity" [Norman-Haignere et al., 2012])

The primary reason for this is to target neural-to-neural correlations, rather than task-to-neural associations.

Our theoretical and empirical results clarify the interpretation of correlation changes from rest to task states in large-scale neural systems. Though empirical studies in large-scale functional networks using fMRI have reported FC increases during task states [Fiebach et al., 2006], we recently found that task-evoked activity inappropriately inflates FC estimates if the mean-evoked activity response is not properly accounted for [Cole et al., 2019]. Indeed, when properly accounting for the mean-evoked response, we found that FC changes from rest to task states were dominated by FC decreases (see Figure 5.3i). The correction of the mean-evoked response in our paradigm brought the empirical results in line with our modeling results, suggesting a counterintuitive interpretation of FC changes during tasks: task co-activation in the presence of neural correlation quenching is consistent with task-related signal communication with background noise suppression. This can be understood from an information-theoretic perspective: during task communication, ongoing spontaneous activity will be suppressed (i.e., neural variability and correlations), increasing the fidelity of the task signal (mean task-evoked response). Our results were consistent using both correlation and covariance measures, suggesting that these decreases were due to reductions in shared variance rather than changes in unshared variance [Fegen, 2012, Siegel et al., 2012, Cole et al., 2016b, Duff et al., 2018]. Furthermore, the present results do not rely on regressing out the task; correlation and variability quenching were also observed independent of this preprocessing step (see Figure 5.4). This was achieved by isolating cross-trial variance, which is similar to computing FC with a beta series correlation [Rissman et al., 2004].

Though we largely focused on FC decreases during task states in both data sets, we identified a small number of correlations that increased during task state. Most of these correlation increases were primarily between regions belonging to different functional networks, such as frontal and visual areas, which is consistent with previous literature [Cole et al., 2014a, Krienen et al., 2014, Gonzalez-Castillo and Bandettini, 2017]. Some correlation increases have also been reported in the NHP spiking literature, where spike count correlations between units with similar task receptive fields tend to decrease, while spike count correlations between units with dissimilar task receptive fields increase [Ruff and Cohen, 2014]. This appears to be conceptually consistent with the present findings, where we focus on mean-field correlation changes between functionally distinct cortical areas (rather than between pairs of individual neurons). Specifically, we found that regions in the same network tend to decrease their correlations, while regions across functionally distinct areas/networks marginally increase their correlations. In our fMRI data, we found marginal task-state correlation increases between different networks, such as the frontoparietal and visual networks (Figure 5.3f). Correlation increases were also observed in one of the two NHPs between frontal (PFC and FEF) and visual (MT and IT) areas (Supplementary Figure C.1). (However, we note we did not observe any correlation increases in our exploratory NHP.) While these correlation increases appear to be statistically reliable, the spontaneous (resting-state) correlations between these areas were very low and the correlation increases marginal. Thus, it will be important for future investigations to directly evaluate the functional relevance of these marginal correlation increases.

We propose that a sigmoid transfer function is an effective model of the activation dynamics of large cortical areas. This was based on causally manipulating a locally balanced E-I circuit with clustered excitatory connectivity [Litwin-Kumar and Doiron, 2012]. The sigmoid transfer function provides a simplification of mean-field neural dynamics that captures excitatory dynamics subject to inhibitory regulation, where feedback inhibition is implicitly modeled by the saturation of the sigmoid function. (We note that the saturation of the sigmoid function does not represent the saturating spiking regime, since spiking saturation does not typically occur in vivo [Priebe and Ferster, 2008].) Moreover, the implementation of the sigmoid transfer function is consistent with prior computational studies demonstrating that resting-state activity corresponds to dynamic regimes with large amplitude, slow fluctuations [Deco et al., 2013b]. In contrast, during task-evoked states, the output dynamics of the sigmoid transfer function are reduced, which correspond to evoked states (e.g., cortical "Up" or asynchronous states) that exhibit quenched variability [Renart et al., 2010, Harris and Thiele, 2011]. The quenching of output variability can be explained by different biological mechanisms, such as clustered excitatory connectivity in local circuits, tightening of E-I balance due to inhibitory feedback, neural adaptation, and/or irregular synaptic vesicle release [Deco and Hugues, 2012, Litwin-Kumar and Doiron, 2012, Rosenbaum et al., 2012, Tetzlaff et al., 2012, Doiron et al., 2016, Hennequin et al., 2018]. The manifestation of these biological mechanisms can be summarized at the mean-field by the reduction of the response variability due to the decreased slope in the sigmoid transfer function during highly active or inactive states. Though other detailed spiking models have offered biophysical mechanisms for inter-area spiking correlations [Huang et al., 2019], we focused here on simplified dynamical systems explanations for correlation and variability changes at the mean-field. It will be important for future work to directly investigate how lower-level biophysical mechanisms map onto model descriptions at the mean-field.

The present results may appear to contradict some reports that task engagement increases (rather than decreases) overall neural communication. Yet there are several key differences between those previous findings and the present results. First, many of the previous results focused on communication through coherence, which often involves frequency-specific coupling of neural signals [Fries, 2005].

This involves the phase-alignment of neural activity on faster timescales, which relates only indirectly to the slower correlation measures of spike rate activity and metabolic demand focused on here [Kahn et al., 2013, Pesaran et al., 2018]. A second key difference between the present and most previous results is our emphasis on the absolute amount of correlation change from rest to task, rather than changes in network organization. Previous studies have also acknowledged that global signal regression, a common fMRI preprocessing step, shifts the baseline of rest- and task-state correlations and artificially induces negative correlations [Fox et al., 2009, Murphy et al., 2009]. This preprocessing step confounds the comparison of the magnitude of correlation changes during independent rest- and task-state fMRI. In the present study, we rigorously preprocessed our fMRI data while ensuring to not remove the global signal. Along with the recent finding that incorrect removal of the mean-evoked response can inflate FC estimates, we suggest that rest to task correlation increases in previous fMRI studies should be interpreted with care.

Despite converging results, there are several key differences in our two empirical data sets. First, the time scale of fMRI BOLD activity is much slower than the NHP spiking activity. However, these differences were mitigated by measuring spiking variability across trials, which is comparable to the time scale of fMRI's sampling rate (in the hundreds of milliseconds). (We note, however, that downsampling spike data does not make it statistically equivalent to fMRI data.) In addition, our computational model results demonstrated that reductions in neural variability and correlations were preserved after nonlinearly transforming the spike rate signal to the fMRI BOLD signal with the Balloon-Windkessel model [Friston et al., 2003], suggesting that the observed signal changes are likely due to the BOLD signal changes rather than MRI artifacts. Despite the computational model demonstrating that statistical properties of BOLD dynamics can be directly caused by spiking dynamics (via the hemodynamic transformation), it is difficult to rule out other possible in vivo explanations in the present study. Another key difference in the data sets is the lack of a true resting-state data set in our NHP data. However, to better compare these two data sets, we demonstrated in the human fMRI data set that the task block periods showed reduced variability relative to inter-block intervals, which is a more analogous comparison to the NHP data set. Despite replication of results across species and task paradigms, our conclusions are based on independently obtained data sets from two species and across task designs. Thus, it will be important for future work to more thoroughly investigate the differences in variability and correlation quenching using experimental designs that simultaneously record both BOLD signal and spiking activity across multiple cortical areas.

In conclusion, we propose a mechanistic framework for interpreting changes in neural variability and correlations by investigating the effects of task-state activity on the underlying neural attractor dynamics. Using empirical data analysis across two highly distinct neural data sets and theoretical modeling, we demonstrated convergent evidence suggesting that task states quench neural variability and correlations due to strengthening neural attractor dynamics across large-scale neural systems. Our work extends previous research establishing similar attractor mechanisms in sensory cortex [Hennequin et al., 2018] to characterize the role of attractor dynamics across large-scale cortical areas. We expect these findings to spur new investigations to better understand how we can interpret neural variability and correlations during task states, providing a deeper understanding of dynamic processes in the brain.

Chapter 6
General Discussion

6.1 Overview and significance

Inspired by connectionist theory, this thesis combines network neuroscience with traditional cognitive mapping techniques to characterize network computations in empirical brain data. Typical network neuroscience approaches facilitate the characterization of either the structural or functional connectomes, providing a way to describe the brain's components [Bassett and Sporns, 2017]. However, these studies often investigate properties of network organization without relating them to the cognitive task activations associated with cognitive processes. In contrast, traditional cognitive mapping approaches aim to characterize which brain areas activate during different cognitive processes. Yet such studies often fail to address exactly how these local brain activations emerge from a mechanistic point of view. Connectionist theory provides a unique opportunity to bridge task-related activations with network organization. This is because connectionist theory addresses how distributed connectivity in a network can perform cognitive computations by propagating activations within the network [Ito et al., 2020b]. Thus, by combining network and cognitive mapping with connectionist theory, this thesis provides a framework to understand cognitive processes by simulating task activation flow through empirically-estimated functional brain networks.

I presented three specific scientific aims that address how cognitive processes reflected in task-evoked activations emerge from distributed network computations. In **Aim 1** (Chapter 2) I focused on how distributed information reflected in highly decodable task activations across cortex were linked to each other through resting-state network connectivity. Specifically, activity in a target brain region could be predicted as the task-evoked activity mapped from a source region. This "activity flow mapping" onto the target region could be described as a linear transformation of its resting-state FC pattern with a source region's activation. This provided an explicit network computation (formalized by the inter-region, voxel-to-voxel resting-state FC matrix) that describe inter-area activity flow pathways.

In **Aim 2** (Chapters 3 and 4), I investigated how distributed information during different task intervals (i.e., rule encoding, stimulus presentation, and behavioral responses) were related to each other within the brain. In Chapter 3, using the same fMRI data set obtained in Chapter 2, we found that behavioral response information during a cognitive control task could be predicted as a nonlinear activity flow transformation from brain regions containing task rule and sensory stimulus information. Critically, this activity flow mapping could be estimated directly from resting-state FC data, suggesting an important role of intrinsic FC in shaping cognitive computations.

In Chapter 4, we extended the concepts and activity flow techniques we tested in fMRI data and applied them to neural spike recordings. We used a previous published data set [Siegel et al., 2015] that obtained multi-unit spike recordings from six cortical areas during a flexible sensorimotor task. Consistent with our report in Chapter 3, we found that the spiking activity during the response (saccade) period in the FEF could be predicted as a nonlinear spiking activity flow transformation from sensory and frontoparietal regions during the task rule and stimulus intervals. Critically, we used stimulus-free spiking activity during ITIs to construct network models of inter-unit FC to predict spiking activity. Though neural interactions have long been studied using noise correlation analyses in the electrophysiology literature [Cohen and Kohn, 2011], few electrophysiology studies have applied the commonly used large-scale network estimation techniques in fMRI to predict spiking activity across cortical ares (however, see [Semedo et al., 2019]). Thus, Aim 2 provided us with the unique opportunity to show that network-based cognitive computations could be identified with both large-scale fMRI and multi-unit spike recordings, bridging concepts and techniques across the human fMRI and non-human animal electrophysiology literatures.

Finally, in **Aim 3** (Chapter 5), I addressed the uses and interpretations of neural correlations in humans and NHPs during intrinsic and task states. Even though the signals measured in fMRI and electrophysiology data come from different sources, in Aims 1 and 2 we were able to construct functional network models that could predict task-evoked activations by estimating inter-unit statistical dependencies. Thus, Aim 3 focused on identifying the principles that govern statistical relationships in neural data.

We began by observing that inter-area mean-field correlation changes were consistently reduced during task states in both data modalities. (Mean-field refers to the average activity of a large neural population.) We subsequently constructed a dynamical systems model that could parsimoniously explain why correlations were reduced during task-evoked states. In particular, in a network with fixed synaptic connectivity, changes in correlated activity can be explained by nonlinear (sigmoidal) activation functions of local neural populations. This suggested to us that the incorporation of nonlinearities in network modeling (i.e., nonlinear transfer functions at each unit) could 1) account for previously reported changes in task-state FC dynamics and 2) improve future activity flow models for better task activation predictions. Indeed, we have recently showed in a follow-up study that accounting for state-dependent network changes (which implicitly accounts for nonlinearities) can improve network models of activity flow [Cole et al., 2020].

6.2 General limitations

Though this thesis aimed to provide an account of cognitive computations through brain network computations, there are several key limitations.

6.2.1 Oversimplification of network computations

The first primary limitation is that the network computations that we estimated in this thesis are likely oversimplified relative to their actual biological implementations. For example, we primarily use activity flow mapping to predict task activations. As presently construed, activity flow mapping takes the linear weighted sum of source regions' activations weighted by each functional connection to a target area (or neuron). Though we show this approximation typically works well in both fMRI and multi-unit data, this simplification ignores the complex biological implementations of synaptic efficacies between neurons, which are highly heterogeneous and nonlinear. Our approach aims to use the simplest possible network computations to approximate this complex process, providing opportunity for future work to build more sophisticated and detailed models of network computations.

6.2.2 Causal and mechanistic interpretability

In Aims 1 and 2, we simulated activity flow processes as if they were modeled over actual connections. However, current FC estimation procedures are limited in their causal and mechanistic interpretations. This is because current FC methods rely on statistical dependencies, such as correlation and or regression-based techniques, rather than causal relations [Reid et al., 2019]. Moreover, these statistical dependencies do not account for directionality. Instead, the FC methods employed in this thesis typically optimize for prediction of a brain region. This is because we use linear regression-based methods (e.g., principal components and/or ridge regression), which optimizes for prediction on a target given a set of inputs. (Linear regression minimizes the mean-squared error of the target variable.) Though it will be important for future models to use more causally-validated methods [Sanchez-Romero and Cole, 2020, Friston et al., 2003], the approaches used in this thesis are still capable of addressing questions that relate intrinsic functional network organization with cognitive task activations.

Another important limitation is that much of the work presented in this thesis used the fMRI BOLD signal, which is only indirectly related to the neural activity thought to underlie behavior. However, recent studies have illustrated a strong correspondence between BOLD activity and neural activity [Ma et al., 2016, Lake et al., 2020]. And even though BOLD activity likely does not *cause* behavior, BOLD activity reflects the neural activity that in turn likely causes behavior. Importantly, in Aims 2 and 3, we showed that the methods and approaches that we employed on fMRI data were validated using neural spiking data. This correspondence suggests that mechanistic inferences made with fMRI data may also generalize to inferences made on electrophysiology data.

We primarily focus on macro- and mesoscale network computations associated with cognitive processes. Previous theoretical frameworks have suggested that this level of organization may be appropriate for characterizing higher level cognition [Craver, 2007, Craver and Bechtel, 2007]. Nevertheless, it is worth noting that there are likely other levels of organization that can produce similarly mechanistic accounts of cognitive processes. For example, some mechanistic theories suggest that balanced amplification of signals from lower-to-higher order areas is critical for conscious perception [Joglekar et al., 2018, Deco and Kringelbach, 2017], and that this "balanced amplification" is governed by slight changes in excitatory-inhibitory balance [Murphy and Miller, 2009, Ahmadian and Miller, 2019]. Thus, though the current thesis focused mostly on network-level computations of cognitive manipulations, there are many other possible mechanisms at different levels of organization that remain to be explored.

6.3 Potential future directions

This thesis laid the groundwork for future investigations into the role of network computations in cognitive processes. Below are several potential future directions that follow from the results presented here.

6.3.1 What are the computational properties of network transformations?

We used empirically-estimated FC patterns to predict information transfer and transformation. But what were the mathematical and computational properties of these mappings? For example, did some network mappings project from a high-dimensional input space to a low-dimensional output space (or vice-versa)? Did some network mappings preserve aspects of representational invariance (e.g., translation invariance) that are commonly found in some deep neural networks [Richards et al., 2019, Yamins et al., 2014]? Thus, it will be interesting to evaluate the mathematical properties of network-based transformations, as they would shed light on exactly how the brain implements information transformation during cognitive tasks.

6.3.2 What is the role of local processes?

A central concept to connectionist modeling is that a target unit's activity can be predicted as a function of propagating activity from other units. This takes an extreme view that neural computations are purely distributed, and that local intrinsic properties, such as operating timescales or the shape of input-output transfer functions, play a minimal role in shaping computations. However, most models of large-scale neural resting-state dynamics suggest the existence of hierarchical heterogeneity [Wang et al., 2019, Demirtaş et al., 2019]. Indeed, though not included in this thesis, we showed in a follow-up study that different regions across the cortical hierarchy were more easily predicted by distributed activity flow modeling (in fMRI data) than other regions [Ito et al., 2020c]. Specifically, we found that unimodal brain regions were harder to predict that transmodal (association) regions via activity flow modeling. We inferred from this that unimodal regions may process information more locally, since their task activations were not as well predicted by network connectivity. However, exactly how this can be accurately modeled and accounted for remains an open question. Thus, it will be important for future studies to better account for local intrinsic properties and cortical heterogeneity to produce more accurate models of distributed network computations.

6.3.3 Applications to artificial intelligence design?

Our findings in Aim 2 have the potential to inform the machine learning and artificial intelligence fields. Chapters 3 and 4 provided evidence that task-performing, connectionist models can be directly parameterized from empirical data. This provides a proof-of-principle that network principles identified in neuroscience could potentially inform AI architectures. For example, we provided evidence that the intrinsic functional network organization could be leveraged to parameterize a connectionist model of sensorimotor transformations during flexible behavior. Thus, what principles and characteristics of the brain's network organization could inform connectionist architectures and optimization principles? Could optimizing for specific network properties, such as hub-related properties (e.g., rich-club or modular organization) provide more efficient learning principles for ANNs than techniques that exclusively focus on optimizing task performance? Thus, investigating the role of brain network organization could potentially inform and constrain ANN models to produce more efficient and generalizable models.

6.3.4 Application of network modeling to other data modalities

Recent technological advances have increased the ability to invasively record thousands of neurons from many different areas simultaneously [Steinmetz et al., 2018]. Related advances have enabled imaging of the entire dorsal cortex using wide-field calcium imaging [Pinto et al., 2019]. Chapter 4 of this thesis only began to scratch the surface of how network-based methods developed in fMRI (e.g., activity flow mapping) can translate to other types of neural recordings. Thus, as data acquisition methods such as wide-field calcium imaging and neuropixel technology become more available, whole-brain network modeling techniques that were previously reserved for fMRI data provide a unique opportunity to bridge human and non-human neuroscience literatures.

6.4 Conclusion

In conclusion, this thesis aimed to leverage both traditional cognitive brain mapping with network neuroscience to address how intrinsic network organization produce task-related cognitive activations. In Aim 1 (Chapter 2), I showed that cognitive task information in a target brain region could be predicted using FC patterns from a source region. In Aim 2 (Chapters 3 and 4), I demonstrated that cognitive information transformation during flexible behavior can be predicted as a nonlinear transformation of sensory stimulus to motor behavior activations using empirically estimated inter-unit FC weights. This showed that functional network organization formed the functional backbone from which cognitive computations can emerge. Finally, in Aim 3 (Chapter 5), I demonstrated that a local nonlinear activation function (sigmoid) could account for network changes from spontaneous to task-evoked states, suggesting that accounting for nonlinear relationships among neural units may provide a way forward in improving network models of cognitive processes. Together, the work presented in this thesis provides a framework for understanding cognitive processes in empirical brain data in terms of connectionist principles.

References

[Abbott et al., 2011] Abbott, L. F., Rajan, K., and Sompolinsky, H. (2011). Interactions between Intrinsic and Stimulus-Evoked Activity in Recurrent Neural Networks. *The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance*, pages 1–16.

[Advani and Saxe, 2017] Advani, M. S. and Saxe, A. M. (2017). High-dimensional dynamics of generalization error in neural networks. *arXiv:1710.03667 [physics, q-bio, stat]*. arXiv: 1710.03667.

[Aertsen et al., 1989] Aertsen, A. M., Gerstein, G. L., Habib, M. K., and Palm, G. (1989). Dynamics of neuronal firing correlation: modulation of" effective connectivity". *Journal of neurophysiology*, 61(5):900–917.

[Ahmadian and Miller, 2019] Ahmadian, Y. and Miller, K. D. (2019). What is the dynamical regime of cerebral cortex? *arXiv:1908.10101 [q-bio]*. arXiv: 1908.10101.

[Anzellotti et al., 2016] Anzellotti, S., Fedorenko, E., Caramazza, A., and Saxe, R. (2016). Measuring and Modeling Transformations of Information Between Brain Regions with fMRI. *bioRxiv*, pages 1–13.

[Averbeck et al., 2006] Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7:358.

[Barch et al., 2013] Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., and Van Essen, D. C. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–89.

[Barlow, 1992] Barlow, H. B. (1992). Single Cells versus Neuronal Assemblies. In Aertsen, A. and Braitenberg, V., editors, *Information Processing in the Cortex*, pages 169–173, Berlin, Heidelberg. Springer.

[Barral and Reyes, 2016] Barral, J. and Reyes, A. D. (2016). Synaptic scaling rule preserves excitatory–inhibitory balance and salient neuronal network dynamics. *Nature Neuroscience*, (October).

[Bashivan et al., 2019] Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436.

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

[Behzadi et al., 2007] Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1):90–101.

[Biswal et al., 1995] Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10008(10):6.

[Bolt et al., 2017] Bolt, T., Nomi, J. S., Rubinov, M., and Uddin, L. Q. (2017). Correspondence between evoked and intrinsic functional brain network configurations. *Human Brain Mapping*, 38(4):1992–2007.

[Brette, 2019] Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, pages 1–44.

[Brincat et al., 2018] Brincat, S. L., Siegel, M., von Nicolai, C., and Miller, E. K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. *Proceedings of the National Academy of Sciences*, 115(30):E7202 LP – E7211.

[Burden and Faires, 2001] Burden, R. L. and Faires, J. D. (2001). Numerical Analysis (seven edition). *PWS. Boston.*

[Button et al., 2013] Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5):365–76.

[Buxton et al., 1998] Buxton, R. B., Wong, E. C., and Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864. Publisher: Wiley Online Library.

[Carandini and Heeger, 2012] Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62.

[Castelo-Branco et al., 1998] Castelo-Branco, M., Neuenschwander, S., and Singer, W. (1998). Synchronization of visual responses between the cortex, lateral geniculate nucleus, and retina in the anesthetized cat. *Journal of neuroscience*, 18(16):6395–6410.

[Churchland et al., 2010] Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., Bradley, D. C., Smith, M. A., Kohn, A., Movshon, J. A., Armstrong, K. M., Moore, T., Chang, S. W., Snyder, L. H., Lisberger, S. G., Priebe, N. J., Finn, I. M., Ferster, D., Ryu, S. I., Santhanam, G., Sahani, M., and Shenoy, K. V. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13:369.

[Ciric et al., 2017] Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., and Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174–187.

[Cocuzza et al., 2020] Cocuzza, C. V., Ito, T., Schultz, D., Bassett, D. S., and Cole, M. W. (2020). Flexible coordinator and switcher hubs for adaptive task control. *Journal of Neuroscience*. Publisher: Society for Neuroscience Section: Research Articles.

[Cohen et al., 2004] Cohen, J. D., Aston-Jones, G., and Gilzenrat, M. S. (2004). A Systems-Level Perspective on Attention and Cognitive Control: Guided Activation, Adaptive Gating, Conflict Monitoring, and Exploitation versus Exploration. In *Cognitive neuroscience of attention*, pages 71–90. The Guilford Press, New York, NY, US.

[Cohen et al., 1990] Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review*, 97(3):332–61.

[Cohen and Kohn, 2011] Cohen, M. R. and Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 14(7):811–819.

[Cohen and Maunsell, 2009] Cohen, M. R. and Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12:1594.

[Cole et al., 2011a] Cole, M. W., Anticevic, A., Repovs, G., and Barch, D. (2011a). Variable global dysconnectivity and individual differences in schizophrenia. *Biological Psychiatry*, 70(1):43–50.

[Cole et al., 2010a] Cole, M. W., Bagic, A., Kass, R., and Schneider, W. (2010a). Prefrontal Dynamics Underlying Rapid Instructed Task Learning Reverse with Practice. *Journal of Neuroscience*, 30(42):14245–14254.

[Cole et al., 2014a] Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., and Petersen, S. E. (2014a). Intrinsic and task-evoked network architectures of the human brain. *Neuron*, 83(1):238–251.

[Cole et al., 2017] Cole, M. W., Braver, T. S., and Meiran, N. (2017). The task novelty paradox: Flexible control of inflexible neural pathways during rapid instructed task learning. *Neuroscience & Biobehavioral Reviews*, 81:4–15.

[Cole et al., 2011b] Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W., and Braver, T. S. (2011b). Rapid Transfer of Abstract Rules to Novel Contexts in Human Lateral Prefrontal Cortex. *Frontiers in Human Neuroscience*, 5:142.

[Cole et al., 2016a] Cole, M. W., Ito, T., Bassett, D. S., and Schultz, D. H. (2016a). Activity flow over resting-state networks shapes cognitive task activations. *Nature Neuroscience*.

[Cole et al., 2015] Cole, M. W., Ito, T., and Braver, T. S. (2015). The Behavioral Relevance of Task Information in Human Prefrontal Cortex. *Cerebral cortex (New York, N.Y. : 1991)*, pages bhv072–.

[Cole et al., 2020] Cole, M. W., Ito, T., Cocuzza, C., and Sanchez-Romero, R. (2020). The functional relevance of task-state functional connectivity. *bioRxiv*, page 2020.07.06.187245. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[Cole et al., 2019] Cole, M. W., Ito, T., Schultz, D., Mill, R., Chen, R., and Cocuzza, C. (2019). Task activations produce spurious but systematic inflation of task functional connectivity estimates. *NeuroImage*, 189:1–18.

[Cole et al., 2010b] Cole, M. W., Pathak, S., and Schneider, W. (2010b). Identifying the brain's most globally connected regions. *NeuroImage*, 49(4):3132–3148.

[Cole et al., 2014b] Cole, M. W., Repov, G., and Anticevic, A. (2014b). The Frontoparietal Control System: A Central Role in Mental Health. *The Neuroscientist*, 20(6):652–664.

[Cole et al., 2013] Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., and Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, 16(9):1348–1355.

[Cole et al., 2016b] Cole, M. W., Yang, G. J., Murray, J. D., Repovš, G., and Anticevic, A. (2016b). Functional connectivity change as shared signal dynamics. *Journal of Neuroscience Methods*, 259:22–39.

[Cole et al., 2012] Cole, M. W., Yarkoni, T., Repovs, G., Anticevic, A., and Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(26):8988–99.

[Collins and Loftus, 1975] Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

[Corbetta and Shulman, 2002] Corbetta, M. and Shulman, G. L. (2002). Control of Goal-Directed and Stimulus-Driven Attention in the Brain. *Nature Reviews Neuroscience*, 3(3):215–229.

[Coutanche and Thompson-Schill, 2013] Coutanche, M. N. and Thompson-Schill, S. L. (2013). Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in human neuroscience*, 7:15.

[Cowan et al., 2016] Cowan, J. D., Neuman, J., and van Drongelen, W. (2016). Wilson-Cowan Equations for Neocortical Dynamics. *Journal of mathematical neuroscience*, 6(1):1.

[Craver, 2007] Craver, C. F. (2007). *Explaining the brain: mechanisms and the mosaic unity of neuroscience*. Clarendon Press, Oxford : New York : Oxford University Press.

[Craver and Bechtel, 2007] Craver, C. F. and Bechtel, W. (2007). Top-down causation without top-down causes. *Biology & philosophy*, 22(4):547–563. Publisher: Springer.

[da Silveira and Berry, 2014] da Silveira, R. A. and Berry, M. J. (2014). High-Fidelity Coding with Correlated Neurons. *PLoS Computational Biology*, 10(11):e1003970.

[De-Wit et al., 2016] De-Wit, L., Alexander, D., Ekroll, V., and Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, pages 1–14.

[Deco and Hugues, 2012] Deco, G. and Hugues, E. (2012). Neural Network Mechanisms Underlying Stimulus Driven Variability Reduction. *PLOS Computational Biology*, 8(3):e1002395.

[Deco et al., 2013a] Deco, G., Jirsa, V. K., and McIntosh, A. R. (2013a). Resting brains never rest: Computational insights into potential cognitive architectures. *Trends in Neurosciences*, 36(5):268–274.

[Deco and Kringelbach, 2017] Deco, G. and Kringelbach, M. L. (2017). Hierarchy of Information Processing in the Brain: A Novel 'Intrinsic Ignition' Framework. *Neuron*, 94(5):961–968.

[Deco et al., 2014] Deco, G., Ponce-Alvarez, A., Hagmann, P., Romani, G. L., Mantini, D., and Corbetta, M. (2014). How local excitation-inhibition ratio impacts the whole brain dynamics. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(23):7886–98.

[Deco et al., 2013b] Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G. L., Hagmann, P., and Corbetta, M. (2013b). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(27):11239–52.

[Demirtaş et al., 2019] Demirtaş, M., Burt, J. B., Helmer, M., Ji, J. L., Adkinson, B. D., Glasser, M. F., Van Essen, D. C., Sotiropoulos, S. N., Anticevic, A., and Murray, J. D. (2019). Hierarchical Heterogeneity across Human Cortex Shapes Large-Scale Neural Dynamics. *Neuron*, 101(6):1181–1194.e13.

[Diedrichsen and Kriegeskorte, 2017] Diedrichsen, J. and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508.

[Doiron et al., 2016] Doiron, B., Litwin-Kumar, A., Rosenbaum, R., Ocker, G. K., and Josić, K. (2016). The mechanics of state-dependent neural correlations. *Nature Neuroscience*, 19:383.

[Dosenbach et al., 2007] Dosenbach, N. U. F., Fair, D. a., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. a. T., Fox, M. D., Snyder, A. Z., Vincent, J. L., Raichle, M. E., Schlaggar, B. L., and Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 104(26):11073–11078.

[Dosenbach et al., 2006] Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., Burgund, E. D., Grimes, A. L., Schlaggar, B. L., and Petersen, S. E. (2006). A Core System for the Implementation of Task Sets. *Neuron*, 50(5):799–812.

[Duff et al., 2018] Duff, E. P., Makin, T., Cottaar, M., Smith, S. M., and Woolrich, M. W. (2018). Disambiguating brain functional connectivity. *NeuroImage*, 173:540–550.

[Ecker et al., 2010] Ecker, A. S., Berens, P., Keliris, G. A., Bethge, M., Logothetis, N. K., and Tolias, A. S. (2010). Decorrelated neuronal firing in cortical microcircuits. *Science (New York, N.Y.)*, 327(5965):584–7.

[Eliasmith, 2007] Eliasmith, C. (2007). How to build a brain: from function to implementation. *Synthese*, 159(3):373–388.

[Eliasmith et al., 2012] Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., Dewolf, T., Tang, Y., and Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, 338(NOVEMBER):1202–1205.

[Fegen, 2012] Fegen, D. (2012). *Cortical mechanisms underlying verbal working memory*. PhD thesis, UC Berkeley.

[Fiebach et al., 2006] Fiebach, C. J., Rissman, J., and D'Esposito, M. (2006). Modulation of Inferotemporal Cortex Activation during Verbal Working Memory Maintenance. *Neuron*, 51(2):251–261.

[Finn et al., 2015] Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., and Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671.

[Fox et al., 2005] Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678.

[Fox et al., 2009] Fox, M. D., Zhang, D., Snyder, A. Z., and Raichle, M. E. (2009). The global signal and observed anticorrelated resting state brain networks. *Journal of neurophysiology*, 101(6):3270–3283. Publisher: American Physiological Society.

[Fries, 2005] Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9(10):474–480.

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

[Friston et al., 1994] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210.

[Fusi et al., 2016] Fusi, S., Miller, E. K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74.

[Genon et al., 2018] Genon, S., Reid, A., Langner, R., Amunts, K., and Eickhoff, S. B. (2018). How to Characterize the Function of a Brain Region. *Trends in Cognitive Sciences*, 22(4):350–364.

[Genovese and Wasserman, 2002] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3):499–517.

[Glasser et al., 2018] Glasser, M. F., Coalson, T. S., Bijsterbosch, J. D., Harrison, S. J., Harms, M. P., Anticevic, A., Van Essen, D. C., and Smith, S. M. (2018). Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. *NeuroImage*, 181:692–717.

[Glasser et al., 2016a] Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016a). A multi-modal parcellation of human cerebral cortex. *Nature*, pages 1–11.

[Glasser et al., 2016b] Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L. R., Auerbach, E. J., Behrens, T. E. J., Coalson, T. S., Harms, M. P., Jenkinson, M., Moeller, S., Robinson, E. C., Sotiropoulos, S. N., Xu, J., Yacoub, E., Ugurbil, K., and Van Essen, D. C. (2016b). The Human Connectome Project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–87.

[Glasser et al., 2013] Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.

[Gonzalez-Castillo and Bandettini, 2017] Gonzalez-Castillo, J. and Bandettini, P. A. (2017). Task-based dynamic functional connectivity: Recent findings and open questions. *NeuroImage*, (May):1–8.

[Gordon et al., 2014] Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., and Petersen, S. E. (2014). Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral Cortex*.

[Gratton et al., 2016] Gratton, C., Laumann, T. O., Gordon, E. M., Adeyemo, B., and Petersen, S. E. (2016). Evidence for Two Independent Factors that Modify Brain Networks to Meet Task Goals. *Cell Reports*, 17(5):1276–1288.

[Haak et al., 2013] Haak, K. V., Winawer, J., Harvey, B. M., Renken, R., Dumoulin, S. O., Wandell, B. A., and Cornelissen, F. W. (2013). Connective field modeling. *NeuroImage*, 66:376–384.

[Hafting et al., 2005] Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806.

[Hagmann et al., 2008] Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Van Wedeen, J., and Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, 6(7):1479–1493.

[Harris and Thiele, 2011] Harris, K. D. and Thiele, A. (2011). Cortical state and attention. *Nature Reviews Neuroscience*, 12:509.

[Hartman et al., 1990] Hartman, E. J., Keeler, J. D., and Kowalski, J. M. (1990). Layered Neural Networks with Gaussian Hidden Units as Universal Approximations. *Neural Computation*, 2(2):210–215.

[Haxby et al., 2014] Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual review of neuroscience*, (June):435–456.

[Haxby et al., 2006] Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2006). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. 2425(2001):2425–2431.

[He, 2011] He, B. J. (2011). Scale-Free Properties of the Functional Magnetic Resonance Imaging Signal during Rest and Task. *The Journal of Neuroscience*, 31(39):13786 LP – 13795.

[He, 2013] He, B. J. (2013). Spontaneous and Task-Evoked Brain Activity Negatively Interact. 33(11):4672–4682.

[Heinzle et al., 2011] Heinzle, J., Kahnt, T., and Haynes, J. D. (2011). Topographically specific functional connectivity between visual field maps in the human brain. *NeuroImage*, 56(3):1426–1436.

[Hennequin et al., 2018] Hennequin, G., Ahmadian, Y., Rubin, D. B., Lengyel, M., and Miller, K. D. (2018). The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability. *Neuron*, 98(4):846–860.e5.

[Henson, 2005] Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology Section A*, 58(2):193–233. Publisher: Taylor & Francis.

[Honey et al., 2009] Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., and Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040. Publisher: National Academy of Sciences Section: Biological Sciences.

[Huang et al., 2019] Huang, C., Ruff, D. A., Pyle, R., Rosenbaum, R., Cohen, M. R., and Doiron, B. (2019). Circuit Models of Low-Dimensional Shared Variability in Cortical Networks. *Neuron*, 101(2):337–348.e4.

[Huth et al., 2016] Huth, A. G., Heer, W. A. D., Griffiths, T. L., Theunissen, F. E., and Jack, L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

[Huth et al., 2012] Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, 76(6):1210–1224.

[Ito et al., 2020a] Ito, T., Brincat, S. L., Siegel, M., Mill, R. D., He, B. J., Miller, E. K., Rotstein, H. G., and Cole, M. W. (2020a). Task-evoked activity quenches neural correlations and variability across cortical areas. *PLOS Computational Biology*, 16(8):e1007983. Publisher: Public Library of Science.

[Ito et al., 2020b] Ito, T., Hearne, L., Mill, R., Cocuzza, C., and Cole, M. W. (2020b). Discovering the Computational Relevance of Brain Network Organization. *Trends in Cognitive Sciences*, 24(1):25–38.

[Ito et al., 2020c] Ito, T., Hearne, L. J., and Cole, M. W. (2020c). A cortical hierarchy of localized and distributed processes revealed via dissociation of task activations, connectivity changes, and intrinsic timescales. *NeuroImage*, 221:117141.

[Ito et al., 2017] Ito, T., Kulkarni, K. R., Schultz, D. H., Mill, R. D., Chen, R. H., Solomyak, L. I., and Cole, M. W. (2017). Cognitive task information is transferred between brain regions via resting-state network topology. *Nature Communications*.

[Jacobs et al., 2018] Jacobs, E. A. K., Steinmetz, N. A., Carandini, M., and Harris, K. D. (2018). Cortical state fluctuations during sensory decision making. *bioRxiv*.

[Jbabdi et al., 2013] Jbabdi, S., Sotiropoulos, S. N., and Behrens, T. E. (2013). The topographic connectome. *Current Opinion in Neurobiology*, 23(2):207–215.

[Ji et al., 2019] Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., and Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, 185:35–57.

[Joglekar et al., 2018] Joglekar, M. R., Mejias, J. F., Yang, G. R., and Wang, X.-J. (2018). Inter-areal Balanced Amplification Enhances Signal Propagation in a Large-Scale Circuit Model of the Primate Cortex. *Neuron*, 98(1):222–234.e8.

[Jutla et al., 2011] Jutla, I. S., Jeub, L. G. S., and Mucha, P. J. (2011). A generalized Louvain method for community detection implemented in MATLAB. *URL http://netwiki. amath. unc. edu/GenLouvain.*

[Kahn et al., 2013] Kahn, I., Knoblich, U., Desai, M., Bernstein, J., Graybiel, A. M., Boyden, E. S., Buckner, R. L., and Moore, C. I. (2013). Optogenetic drive of neocortical pyramidal neurons generates fMRI signals that are correlated with spiking activity. *Brain research*, 1511:33–45. Publisher: Elsevier.

[Kanwisher, 2010] Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170.

[Kanwisher et al., 1997] Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, 17(11):4302 LP – 4311.

[Khaligh-Razavi and Kriegeskorte, 2014] Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11).

[Kikumoto and Mayr, 2020] Kikumoto, A. and Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences*. Publisher: National Academy of Sciences Section: Biological Sciences.

[Kingma and Ba, 2017] Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

[Kong et al., 2019] Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2019). Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. *Cerebral Cortex*, 29(6):2533–2551. Publisher: Oxford Academic.

[Kotter and Sommer, 2000] Kotter, R. and Sommer, F. T. (2000). Global relationship between anatomical connectivity and activity propagation in the cerebral cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1393):127–134. Publisher: Royal Society.

[Kriegeskorte and Kievit, 2013] Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412.

[Kriegeskorte et al., 2008] Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(November):4.

[Krienen et al., 2014] Krienen, F. M., Yeo, B. T. T., and Buckner, R. L. (2014). Reconfigurable task-dependent functional coupling modes cluster around a core functional architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653):20130526–20130526.

[Lake et al., 2020] Lake, E. M. R., Ge, X., Shen, X., Herman, P., Hyder, F., Cardin, J. A., Higley, M. J., Scheinost, D., Papademetris, X., Crair, M. C., and Constable, R. T. (2020). Simultaneous cortex-wide fluorescence Ca 2+ imaging and whole-brain fMRI. *Nature Methods*, pages 1–10. Publisher: Nature Publishing Group.

[Lashley, 1931] Lashley, K. S. (1931). Mass Action in Cerebral Function. *Science*, 73(1888):245–254. Publisher: American Association for the Advancement of Science Section: Articles.

[Litwin-Kumar and Doiron, 2012] Litwin-Kumar, A. and Doiron, B. (2012). Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature Neuroscience*, 15(11):1498–1505.

[Litwin-Kumar et al., 2017] Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H., and Abbott, L. (2017). Optimal Degrees of Synaptic Connectivity. *Neuron*, 0(0):1153–1164.e7.

[Litwin-Kumar and Turaga, 2019] Litwin-Kumar, A. and Turaga, S. C. (2019). Constraining computational models using electron microscopy wiring diagrams. *Current Opinion in Neurobiology*, 58:94–100.

[Luck, 2014] Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

[Ma et al., 2016] Ma, Y., Shaik, M. A., Kozberg, M. G., Kim, S. H., Portes, J. P., Timerman, D., and Hillman, E. M. C. (2016). Resting-state hemodynamics are spatiotemporally coupled to synchronized and symmetric neural activity in excitatory neurons. *Proceedings of the National Academy of Sciences*, 113(52):E8463–E8471.

[MacKay, 2003] MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Mante et al., 2013] Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84.

[Marrelec et al., 2016] Marrelec, G., Messé, A., Giron, A., and Rudrauf, D. (2016). Functional Connectivity's Degenerate View of Brain Computation. *PLOS Computational Biology*, 12(10):e1005031.

[Mars et al., 2018] Mars, R. B., Passingham, R. E., and Jbabdi, S. (2018). Connectivity Fingerprints: From Areal Descriptions to Abstract Spaces. *Trends in Cognitive Sciences*, 22(11):1026–1037.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[McNaughton et al., 2006] McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., and Moser, M.-B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, 7(8):663–678.

[Medaglia et al., 2015] Medaglia, J. D., Lynall, M.-E., and Bassett, D. S. (2015). Cognitive Network Neuroscience. *Journal of Cognitive Neuroscience*, 27(8):1471–1491.

[Mill et al., 2017] Mill, R. D., Ito, T., and Cole, M. W. (2017). From connectome to cognition: The search for mechanism in human functional brain networks. *NeuroImage*, (January):0–1.

[Miller and Buschman, 2015] Miller, E. K. and Buschman, T. J. (2015). Working Memory Capacity: Limits on the Bandwidth of Cognition. *Daedalus*, 144(1):112–122.

[Miller and Cohen, 2001] Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202.

[Muhle-Karbe et al., 2016] Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., and Brass, M. (2016). Neural Coding for Instruction-Based Task Sets in Human Frontoparietal and Visual Cortex. *Cerebral Cortex*, page bhw032.

[Mur et al., 2009] Mur, M., Bandettini, P. A., and Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, 4(1):101–109. Publisher: Oxford Academic.

[Murphy and Miller, 2009] Murphy, B. K. and Miller, K. D. (2009). Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns. *Neuron*, 61(4):635–648.

[Murphy et al., 2009] Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., and Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage*, 44(3):893–905.

[Nichols and Holmes, 2001] Nichols, T. E. and Holmes, A. P. (2001). Nonparametric Permutation Tests for Functional Neuroimaging Experiments: A Primer with examples. *Human Brain Mapping*, 15(1):1–25.

[Nishimoto et al., 2011] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646.

[Norman et al., 2006] Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–30.

[Norman-Haignere et al., 2012] Norman-Haignere, S. V., McCarthy, G., Chun, M. M., and Turk-Browne, N. B. (2012). Category-Selective Background Connectivity in Ventral Visual Cortex. *Cerebral Cortex*, 22(2):391–402.

[Oosterhof et al., 2011] Oosterhof, N. N., Wiestler, T., Downing, P. E., and Diedrichsen, J. (2011). A comparison of volume-based and surface-based multivoxel pattern analysis. *NeuroImage*, 56(2):593–600.

[Passingham et al., 2002] Passingham, R. E., Stephan, K. E., and Kötter, R. (2002). The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience*, 3(8):606–616.

[Pesaran et al., 2018] Pesaran, B., Vinck, M., Einevoll, G. T., Sirota, A., Fries, P., Siegel, M., Truccolo, W., Schroeder, C. E., and Srinivasan, R. (2018). Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nature Neuroscience*, 21(7):903.

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

[Pinto et al., 2019] Pinto, L., Rajan, K., DePasquale, B., Thiberge, S. Y., Tank, D. W., and Brody, C. D. (2019). Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron*, 0(0).

[Poldrack et al., 2009] Poldrack, R. A., Halchenko, Y. O., and Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, 20(11):1364–1372.

[Poldrack et al., 2011] Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis.* Cambridge University Press.

[Ponce-alvarez et al., 2015] Ponce-alvarez, A., He, B. J., Hagmann, P., and Deco, G. (2015). Task-Driven Activity Reduces the Cortical Activity Space of the Brain : Experiment and Whole-Brain Modeling. pages 1–26.

[Power et al., 2012] Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154.

[Power et al., 2011] 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., and Petersen, S. E. (2011). Functional Network Organization of the Human Brain. *Neuron*, 72(4):665–678.

[Power et al., 2014] Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84:320–341.

[Power and Petersen, 2013] Power, J. D. and Petersen, S. E. (2013). Control-related systems in the human brain. *Current Opinion in Neurobiology*, 23(2):223–228.

[Power et al., 2018] Power, J. D., Plitt, M., Gotts, S. J., Kundu, P., Voon, V., Bandettini, P. A., and Martin, A. (2018). Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data. *Proceedings of the National Academy of Sciences*, 115(9):E2105–E2114.

[Power et al., 2013] Power, J. D., Schlaggar, B. L., Lessov-Schlaggar, C. N., and Petersen, S. E. (2013). Evidence for hubs in human functional brain networks. *Neuron*, 79(4):798–813.

[Priebe and Ferster, 2008] Priebe, N. J. and Ferster, D. (2008). Inhibition, Spike Threshold, and Stimulus Selectivity in Primary Visual Cortex. *Neuron*, 57(4):482–497.

[Rabbitt, 1997] Rabbitt, P. (1997). *Methodology of frontal and executive function*. Psychology Press, Hove, East Sussex, U.K.

[Raichle, 2010] Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, 14(4):180–190.

[Raichle et al., 2001] Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):676–82.

[Reid et al., 2019] Reid, A. T., Headley, D. B., Mill, R. D., Sanchez-Romero, R., Uddin, L. Q., Marinazzo, D., Lurie, D. J., Valdés-Sosa, P. A., Hanson, S. J., Biswal, B. B., Calhoun, V., Poldrack, R. A., and Cole, M. W. (2019). Advancing functional connectivity research from association to causation. *Nature Neuroscience*, 22(11):1751–1760.

[Renart et al., 2010] Renart, A., Rocha, J. D., Bartho, P., Hollender, L., Parga, N., Reyes, A., and Harris, K. D. (2010). in Cortical Circuits. *Science*, 327(January):587–591.

[Reverberi et al., 2012] Reverberi, C., Görgen, K., and Haynes, J.-D. (2012). Compositionality of Rule Representations in Human Prefrontal Cortex. *Cerebral Cortex*, 22(6):1237–1246.

[Richards et al., 2019] Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., Berker, A. d., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

[Rigotti et al., 2013] Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–90.

[Rissman et al., 2004] Rissman, J., Gazzaley, A., and D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, 23(2):752–763.

[Rosenbaum et al., 2012] Rosenbaum, R., Rubin, J. E., and Doiron, B. (2012). Short-term synaptic depression and stochastic vesicle dynamics reduce and shape neuronal correlations. *Journal of Neurophysiology*, 109(2):475–484.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

[Ruff and Cohen, 2014] Ruff, D. A. and Cohen, M. R. (2014). Attention can either increase or decrease spike count correlations in visual cortex. *Nature Neuroscience*, 17(11):1591–1597.

[Ruff and Cohen, 2016] Ruff, D. A. and Cohen, M. R. (2016). Stimulus Dependence of Correlated Variability across Cortical Areas. *Journal of Neuroscience*, 36(28):7546–7556.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., McClelland, J. L., and Others (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1:45–76.

[Russo et al., 2018] Russo, A. A., Bittner, S. R., Perkins, S. M., Seely, J. S., London, B. M., Lara, A. H., Miri, A., Marshall, N. J., Kohn, A., Jessell, T. M., Abbott, L. F., Cunningham, J. P., and Churchland, M. M. (2018). Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. *Neuron*, 97(4):953–966.e8.

[Sadaghiani et al., 2015] Sadaghiani, S., Poline, J.-B., Kleinschmidt, A., and D'Esposito, M. (2015). Ongoing dynamics in large-scale functional connectivity predict perception. *Proceedings of the National Academy of Sciences of the United States of America*, 112(27):8463–8468.

[Sanchez-Romero and Cole, 2020] Sanchez-Romero, R. and Cole, M. W. (2020). Combining Multiple Functional Connectivity Methods to Improve Causal Inferences. *Journal of Cognitive Neuroscience*, pages 1–15. Publisher: MIT Press.

[Saygin et al., 2012] Saygin, Z. M., Osher, D. E., Koldewyn, K., Reynolds, G., Gabrieli, J. D. E., and Saxe, R. R. (2012). Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience*, 15(2):321–327.

[Saygin et al., 2016] Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D. E., and Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*, 19(9):1250–1255.

[Schaefer et al., 2018] Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114. Publisher: Oxford University Press.

[Schneider et al., 2002] Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime: User's guide*. Psychology Software Incorporated.

[Schultz and Cole, 2016] Schultz, D. H. and Cole, M. W. (2016). Higher Intelligence Is Associated with Less Task-Related Brain Network Reconfiguration. *Journal of Neuroscience*, 36(33):8551–8561.

[Schultz et al., 2019] Schultz, D. H., Ito, T., Solomyak, L. I., Chen, R. H., Mill, R. D., Anticevic, A., and Cole, M. W. (2019). Global connectivity of the fronto-parietal cognitive control network is related to depression symptoms in the general population. *Network Neuroscience*, 3(1):107–123.

[Semedo et al., 2019] Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., and Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. *Neuron*, 102(1):249–259.e4.

[Shannon et al., 2011] Shannon, B. J., Raichle, M. E., Snyder, A. Z., Fair, D. a., Mills, K. L., Zhang, D., Bache, K., Calhoun, V. D., Nigg, J. T., Nagel, B. J., Stevens, A. a., and Kiehl, K. a. (2011). Premotor functional connectivity predicts impulsivity in juvenile offenders. *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11241–11245.

[Shannon, 1948] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

[Siegel et al., 2015] Siegel, M., Buschman, T. J., and Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, 348(6241):1352–55.

[Siegel et al., 2012] Siegel, M., Donner, T. H., and Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience.*

[Simony et al., 2016] Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., and Hasson, U. (2016). Dynamical reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(May 2015):1–13.

[Smith et al., 2013] Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., Kelly, M., Laumann, T., Miller, K. L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A. Z., Vu, A. T., Woolrich, M. W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D. C., and Glasser, M. F. (2013). Resting-state fMRI in the Human Connectome Project. *NeuroImage*, 80:144–168.

[Smith et al., 2015] Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., J Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., and Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(September):1–7.

[Smith et al., 2006] Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J., and Jarvis, E. D. (2006). Computational inference of neural information flow networks. *PLoS Computational Biology*, 2(11):1436–1449.

[Song et al., 2016] Song, H. F., Yang, G. R., and Wang, X.-J. (2016). Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework. *PLOS Computational Biology*, 12(2):e1004792.

[Song et al., 2005] Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits. *PLoS Biology*, 3(3):e68.

[Sporns et al., 2005] Sporns, O., Tononi, G., and K"otter, R. (2005). The human connectome: a structural description of the human brain.

[Steinmetz et al., 2018] Steinmetz, N. A., Koch, C., Harris, K. D., and Carandini, M. (2018). Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Current opinion in neurobiology*, 50:92–100. Publisher: Elsevier.

[Stern et al., 2014] Stern, M., Sompolinsky, H., and Abbott, L. F. (2014). Dynamics of random neural networks with bistable units. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 90(6):1–7.

[Strogatz, 1994] Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering (Cambridge, MA.* Westview Press.

[Szucs and Ioannidis, 2016] Szucs, D. and Ioannidis, J. P. A. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *bioRxiv*.

[Tavor et al., 2016] Tavor, I., Jones, O. P., Mars, R. B., Smith, S. M., Behrens, T. E., and Jbabdi, S. (2016). Task-free MRI predicts individual differences in brain activity during task performance. *Science*, 352(6282):1773–1776.

[Tetzlaff et al., 2012] Tetzlaff, T., Helias, M., Einevoll, G. T., and Diesmann, M. (2012). Decorrelation of Neural-Network Activity by Inhibitory Feedback. *PLoS Computational Biology*, 8(8):e1002596.

[Timme et al., 2016] 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):e1004858.

[Tomasi et al., 2014] Tomasi, D., Wang, R., Wang, G.-J., and Volkow, N. D. (2014). Functional Connectivity and Brain Activation: A Synergistic Approach. *Cerebral Cortex*, 24(10):2619–2629.

[Tononi et al., 1999] Tononi, G., Sporns, O., and Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, 96(6):3257–3262.

[Traud et al., 2011] Traud, A. L., Kelsic, E. D., Mucha, P. J., and Porter, M. a. (2011). Comparing Community Structure to Characteristics in Online Collegiate Social Networks. *SIAM Review*, 53(3):17.

[Tsao et al., 2006] Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A Cortical Region Consisting Entirely of Face-Selective Cells. *Science*, 311(5761):670 LP – 674.

[Tschopp et al., 2018] Tschopp, F. D., Reiser, M. B., and Turaga, S. C. (2018). A Connectome Based Hexagonal Lattice Convolutional Network Model of the Drosophila Visual System. *arXiv:1806.04793 [cs, q-bio]*. arXiv: 1806.04793.

[Turing, 1948] Turing, A. M. (1948). *Intelligent machinery*. NPL. Mathematics Division.

[van den Heuvel et al., 2009] van den Heuvel, M. P., Mandl, R. C. W., Kahn, R. S., and Hulshoff Pol, H. E. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Human Brain Mapping*, 30(10):3127–3141.

[van den Heuvel and Sporns, 2011] van den Heuvel, M. P. and Sporns, O. (2011). Rich-Club Organization of the Human Connectome. *Journal of Neuroscience*, 31(44):15775–15786.

[van den Heuvel and Sporns, 2013] van den Heuvel, M. P. and Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12):683–696.

[Van Essen et al., 2013] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., and WU-Minn HCP Consortium (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79.

[Varoquaux, 2018] Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77.

[Von Neumann and Kurzweil, 2012] Von Neumann, J. and Kurzweil, R. (2012). *The computer and the brain*. Yale University Press.

[Wallis, 2018] Wallis, J. D. (2018). Decoding Cognitive Processes from Neural Ensembles. *Trends in Cognitive Sciences*, 22(12):1091–1102.

[Wang et al., 2019] Wang, P., Kong, R., Kong, X., Liégeois, R., Orban, C., Deco, G., Heuvel, M. P. v. d., and Yeo, B. T. T. (2019). Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain. *Science Advances*, 5(1):eaat7854.

[Waskom and Kiani, 2018] Waskom, M. L. and Kiani, R. (2018). Decision Making through Integration of Sensory Evidence at Prolonged Timescales. *Current Biology*, 28(23):3850–3856.e9.

[Waskom et al., 2014] Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., and Wagner, A. D. (2014). Frontoparietal representations of task context support the flexible control of goal-directed cognition. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(32):10743–55.

[Wen et al., 2018] Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2018). Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28(12):4136–4160.

[Widrow and Lehr, 1990] Widrow, B. and Lehr, M. A. (1990). 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442.

[Wilson and Cowan, 1972] Wilson, H. R. and Cowan, J. D. (1972). Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. *Biophysical Journal*, 12(1):1–24.

[Wong et al., 2013] Wong, C. W., Olafsson, V., Tal, O., and Liu, T. T. (2013). The amplitude of the resting-state fMRI global signal is related to EEG vigilance measures. *NeuroImage*, 83:983–990.

[Wu et al., 2016] Wu, Y., Zhang, S., Zhang, Y., Bengio, Y., and Salakhutdinov, R. R. (2016). On Multiplicative Integration with Recurrent Neural Networks. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2856–2864. Curran Associates, Inc.

[Yamins et al., 2014] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

[Yang et al., 2019] Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, page 1.

[Yang and Wang, 2020] Yang, G. R. and Wang, X.-J. (2020). Artificial neural networks for neuroscientists: A primer. *arXiv:2006.01001 [cs, q-bio]*. arXiv: 2006.01001.

[Yeo et al., 2011] Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zollei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165.

[Yokoi and Diedrichsen, 2018] Yokoi, A. and Diedrichsen, J. (2018). Parcellation of motor sequence representations in the human neocortex. *bioRxiv*, page 419754.

[Zhang et al., 2013] Zhang, J., Kriegeskorte, N., Carlin, J. D., and Rowe, J. B. (2013). Choosing the rules: distinct and overlapping frontoparietal representations of task rules for perceptual decisions. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(29):11852–62.

Appendix A

Appendix – Chapter 2

A.1 Supplementary Figures

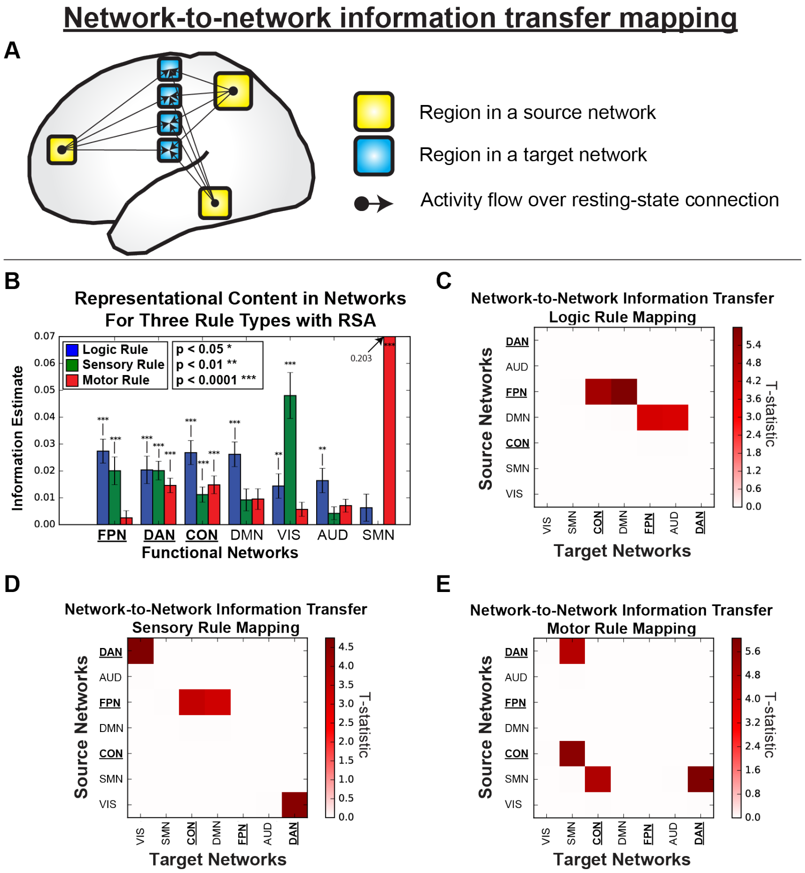


Figure A.1: Supplementary Figure 1. Task information persist in functional networks and are transferred between networks via network-to-network information transfer mapping. All reported results were statistically significant $p < 0.05$ (FWE-corrected). A) Network-to-network information transfer mapping uses the network-level activation pattern (using the mean activations of brain regions) and the region-to-region resting-state FC topology to predict the network-level activation pattern of another functional network. B) Information estimates of task-rule information across three rule domains prior to performing information transfer mapping. The seven networks contain statistically significant decodable representations of at least one rule domain using a cross-validated representational similarity analysis approach. In particular, the SMN contains the highest information estimate for motor task rules. In addition, most networks contain logic rule information, suggesting that abstract rule representations were highly distributed across cortical networks. C) Network-to-network information transfer mapping of logic rules. As in Figure 2.6, functional networks along the rows indicate the activation patterns that were projected to the networks indicated along the columns. Colors indicate the T-statistic from a one-sided t-test against 0. The transfer of logic rule information was distributed among other domain-general networks, such as the CON and DMN. D) Network-to-network information transfer mapping of the sensory rules. Sensory rule information is transferred between the FPN and other domain-general networks (DMN, CON), as well as from VIS and the DAN. E) Network-to-network information transfer of motor rules. Information transfer mapping of motor rule representations occurs between the DAN and SMN, CON and SMN.

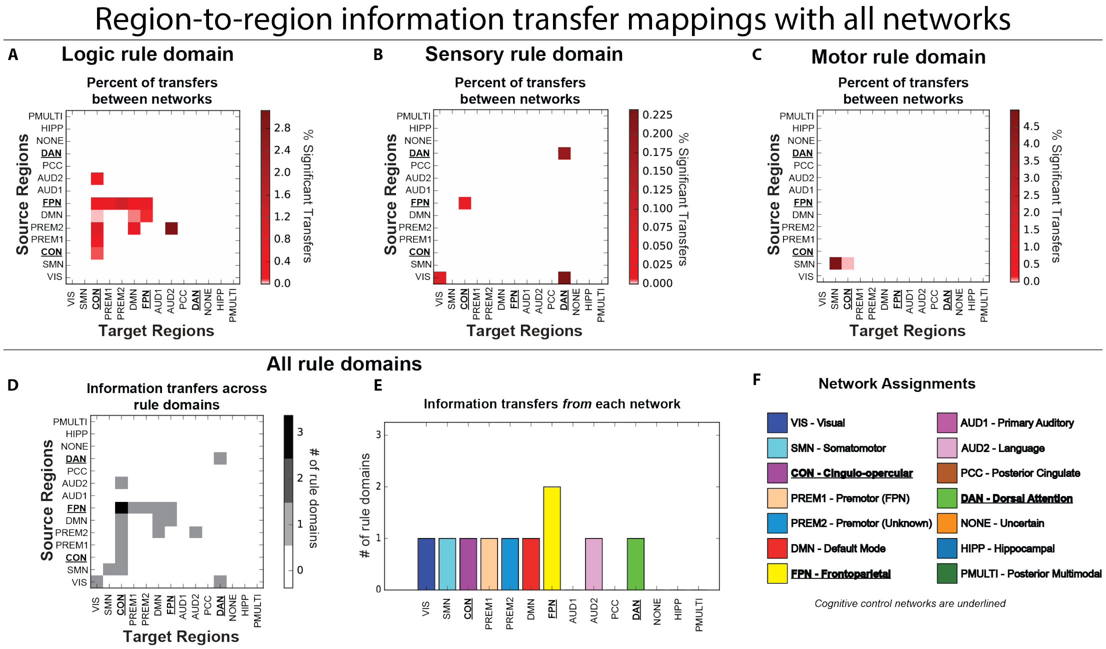


Figure A.2: Supplementary Figure 2. Information transfer mappings between all pairs of regions for all defined functional networks. All reported results were statistically significant at $p < 0.05$ (FWE-corrected). Here, we show a superset of the summary results shown in Figure 2.6b,d,f, including significant information transfer results for all 14 functional networks. While results for all regions (belonging to all 14 functional networks) are shown in panels Figure 2.6a,c,e, functional networks that were not well-defined by previous network partitions were not included in Figure 2.6b,d,f. Note that the 7 functional networks not included in Figure 2.6 are the seven smallest networks, each consisting of fewer than 20 parcels. A) Percent of significant region-to-region information transfers for all 14 network definitions for the logic rule domain. B) Percent of significant region-to-region information transfers for all 14 network affiliations for the sensory rule domain. C) Percent of significant region-to-region information transfers for all 14 network affiliations for the motor rule domain. D) Significant information transfers between regions for all 14 network affiliation across rule domains, derived in the same way as data in Figure 2.6g. Despite including all functional networks, we found that transfers between the FPN and the CON were still the only transfers between a pair of networks that consistently transferred information across two rule domains. E) We assessed whether a network was consistently involved in sending task rule information (as a source region) across each rule domain. We found that regions in the FPN consistently transferred information across two rule domains. F) Network assignments and color definitions for all 14 functional networks. Here, we attribute functional names for all 14 networks. Color schemes are consistent with colorings shown on the anatomical surface in Figure 2.3a

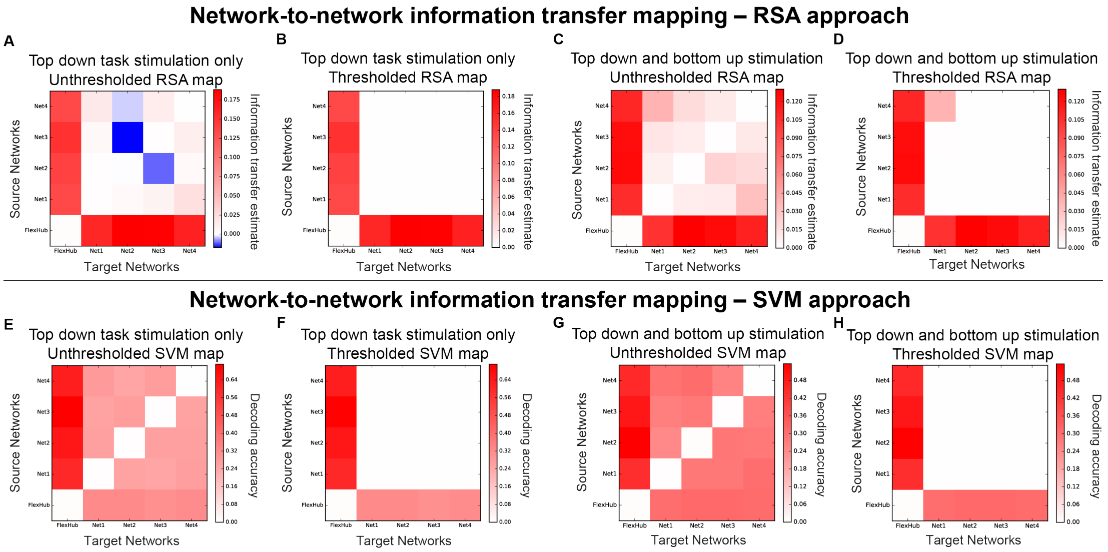


Figure A.3: Supplementary Figure 3. Computational validation of information transfer mapping with different task stimulation patterns and decoding approaches. We simulated an additional 35 subjects to test whether information transfer mapping would be consistent across different task stimulation patterns and decoding approaches. A) We performed the same analysis as depicted in Figure 2.4, where we simulated cognitive control task rules by stimulating regions in the hub network for four distinct task-rule conditions. We depict the uncorrected information transfer estimates for every network-to-network configuration using the information transfer mapping procedure described in Figure 2.1c. B) Thresholded map of panel Supplementary Figure 2.3a. For every network-to-network information transfer mapping, we performed an across subject, one-sided t-test against 0. Statistical significance is assessed using FWE-corrected p-values of $p < 0.05$. C) We simulated a task that combined both top down stimulation (e.g., mimicking task-rule encoding) and bottom up stimulation in local networks (e.g., mimicking stimuli presentations). This task also included four distinct task conditions, where each condition stimulated a subset of regions in the hub network along with a subset of regions in a local network simultaneously. Each task condition stimulated a subset of regions in a different local network. D) Thresholded map of panel C, using FWE-corrected p-values of $p < 0.05$. While the pattern of information transfer was largely the same, information transfer of both top down and bottom up stimulation was more disperse than top down stimulation only. E-H) We performed the group analyses on the same exact data (see Supplementary materials) as Supplementary Figures A.3a-d but instead of using an RSA approach (i.e., predicted-to-similarity analysis; Figure 2.1c), we used SVMs (training on predicted activation patterns and testing on held-out actual activation patterns). Note that panel F has qualitatively identical results as in our computational validation results (Figure 2.4e) using representational similarity analysis. For panels F and H, statistical significance was assessed using one-sided t-tests against chance (25% chance) for a four-way task condition classification. Thresholds were applied using an FWE-corrected $p < 0.05$. All panels show the raw effect sizes.

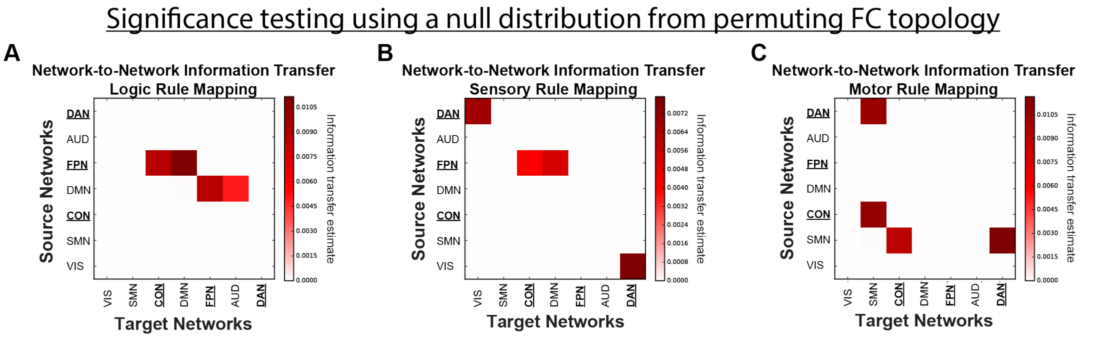
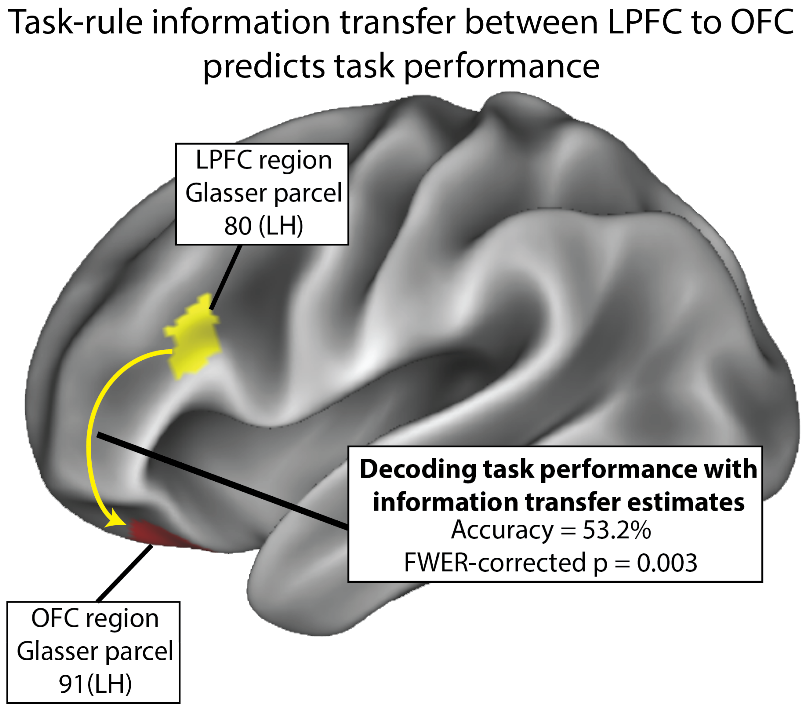


Figure A.4: Supplementary Figure 4. Network-to-network information transfer mapping depends on precise FC topology between pairs of networks. All reported results were statistically significant at $p < 0.05$ (FWE-corrected). To ensure that information transfer mapping between networks depended on the precise FC topology between pairs of networks, we generated a null distribution of information transfers by permuting the inter-region FC patterns between pairs of networks prior to performing the network-to-network information transfer procedure. For each network-to-network information transfer mapping, 1000 FC permutations were conducted. Significant results demonstrate that the information transfer depended on the precise network-to-network FC topology. This analysis demonstrates that the results obtained using parametric statistical testing (Supplementary Figure A.1) depend on the precise inter-region FC patterns between pairs of networks, as results from the parametric and non-parametric tests are virtually identical. Color maps represent the group averaged information transfer estimate, since no t-statistic is available in the null distribution.



Task-rule information transfer between LPFC to OFC predicts task performance

Figure A.5: Supplementary Figure 5. The behavioral relevance of cognitive task information transfer. We found that task-rule information transfer between two FPN regions could decode miniblock task performance significantly above chance. We constructed a decoding model using multiple logistic regression to decode task performance in a held-out miniblock by fitting to the logic, sensory, and motor information transfer estimates across miniblocks. When transformed into the OFC region's spatial dimensions, task-rule information in the LPFC region could predict a miniblock's task performance significantly above chance, suggesting that the transfer of task-rule information between these regions is relevant for task performance.

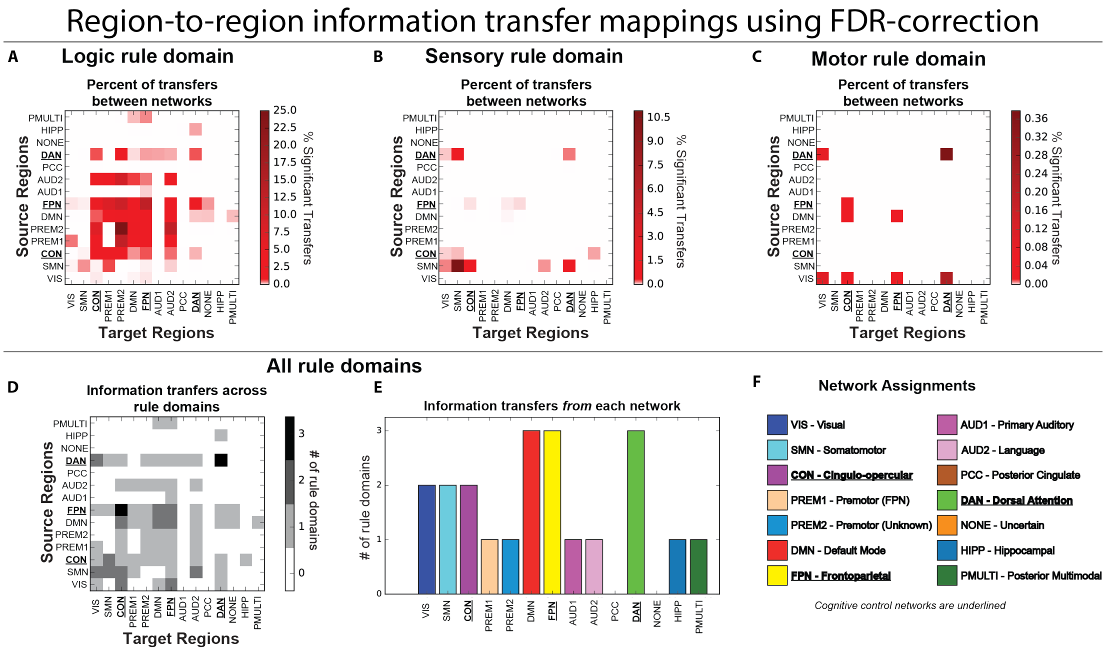


Figure A.6: Supplementary Figure 6. Information transfer mappings between all pairs of regions using an FDR-corrected threshold for all defined functional networks. Due to the conservative nature of FWE correction for multiple comparisons correction, we also report the same results from Figure 2.6 and Supplementary Figure A.2 using an FDR-corrected p-value of $p < 0.05$. Using FDR-correction, we found that statistically significant task-rule information transfers were much more distributed than with FWE-correction, particularly with logic rule transfers. A) Percent of significant region-to-region information transfers for all 14 network definitions for the logic rule domain. B) Percent of significant region-to-region information transfers for all 14 network affiliations for the sensory rule domain. C) Percent of significant region-to-region information transfers for all 14 network affiliations for the motor rule domain. D) Significant information transfers between regions for all 14 network affiliation across rule domains, derived in the same way as data in Figure 2.6G. E) We assessed whether a network was consistently involved in sending task rule information (as a source region) across the three rule domains. We find that with an FDR-corrected threshold of $p < 0.05$, the FPN, DAN, and DMN all contain regions that transfer information across all three rule domains. F) Network assignments and color definitions for all 14 functional networks. Here, we attribute functional names for all 14 networks. Color schemes are consistent with colorings shown on the anatomical surface in Figure 2.3A.

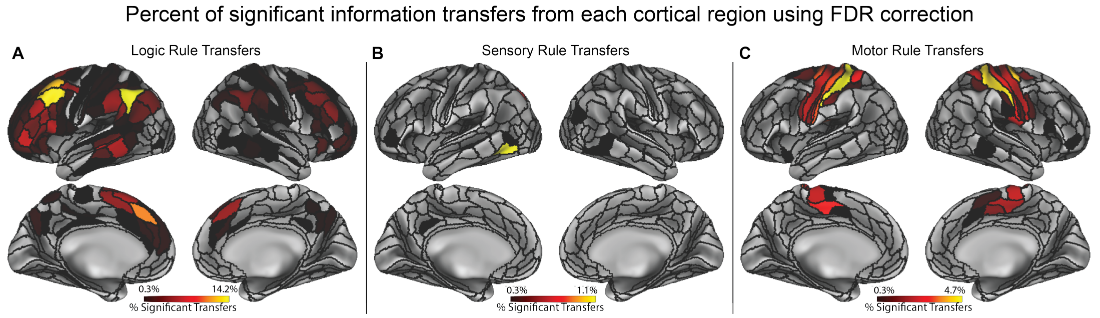


Figure A.7: Supplementary Figure 7. Percent of significant information transfers from each cortical region using an FDR-corrected threshold. Due to the conservative nature of FWE correction for multiple comparisons correction, we also report results from Figure 2.7 using an FDR-corrected threshold of $p < 0.05$. A) Percent of statistically significant information transfers from each region for the logic rule domain. Percentages were computed by taking the number of significant transfers from each region, and dividing it by the total number of possible transfers from that region (359 other regions). B) Percent of statistically significant information transfers from each region for the sensory rule domain. C) Percent of statistically significant information transfers from each region for the motor rule domain.

A.2 Supplementary Methods

We provide Supplementary Methods for several of our Methods subsections below. For completeness, we have included redundant text for the Methods subsections that contain additional information. However, subsections for which there is no additional information (e.g., the "Participants" subsection) are not included in the Supplementary Methods.

A.2.1 Behavioral paradigm

We used the Concrete Permuted Rule Operations (C-PRO) paradigm (Figure 2.2), which is a modified version of the original PRO paradigm introduced in [Cole et al., 2010a]. Briefly, the C-PRO cognitive paradigm permutes specific task rules from three different rule domains (logical decision, sensory semantic, and motor response) to generate dozens of novel and unique task sets. This creates a condition-rich dataset in the task configuration domain akin in some ways to movies and other condition-rich datasets used to investigate visual and auditory domain [Huth et al., 2016, Nishimoto et al., 2011, Simony et al., 2016]. The primary modification of the C-PRO paradigm from the PRO paradigm was to use concrete, sensory (simultaneously presented visual and auditory) stimuli, as opposed to the abstract, linguistic stimuli in the original paradigm. Visual stimuli included either horizontal or vertical oriented bars with either blue or red coloring. Simultaneously presented auditory stimuli included continuous (constant) or non-continuous (non-constant, i.e., "beeping") tones beeps presented at high (3000Hz) or low (300Hz) frequencies. Figure 2.2 demonstrates two example task-rule sets for "Task 1" and "Task 64". The paradigm was presented using E-Prime software version 2.0.10.353 [Schneider et al., 2002].

Each rule domain (logic, sensory, and motor) consisted of four specific rules, while each task set was a combination of one rule from each rule domain (Figure

2.2). The sensory rules specified the audiovisual features to attend to (e.g., "is it vertical?" for visual decisions, or "is it high-pitch?" for auditory decisions). The logic rules specified how to respond based on the pair of stimuli presentations (e.g., "if both are vertical" or "if either are vertical"). Finally, the motor rules specified which button to press, which depended on the answer to the logic rule. For "true" outcomes, subjects were asked to respond with the motor rule presented in the task-rule set; for "false" outcomes, subjects were asked to respond with the other finger on the same hand.

A total of 64 unique task sets (4 logic rules x 4 sensory rules x 4 motor rules) were possible, and each unique task set was presented twice for a total of 128 task miniblocks. Identical task sets were not presented in consecutive blocks. Each task miniblock included three trials, each consisting of two sequentially presented instances of simultaneous audiovisual stimuli. A task block began with a 3925ms instruction screen (5 TRs), followed by a jittered delay ranging from 1570ms to 6280ms (2 – 8 TRs; randomly selected). Following the jittered delay, three trials were presented for 2355ms (3 TRs), each with an inter-trial interval of 1570ms (2 TRs). A second jittered delay followed the third trial, lasting 7850ms to 12560ms (10-16 TRs; randomly selected). A task block lasted a total of 28260ms (36 TRs). Subjects were trained on four of the 64 task-rule sets for 30 minutes prior to the fMRI session. The four practiced rule sets were selected such that all 12 rules were equally practiced. There were 16 such groups of four task sets possible, and the task sets chosen to be practiced were counterbalanced across subjects. Subjects' mean performance across all trials performed in the scanner was 85% (median=86%) with a standard deviation of 8% (min=66%; max=96%). All subjects performed statistically above chance (25%).

A.2.2 Network assignment of Glasser et al. (2016) parcels

Partitioning of the parcels (regions) into networks was based on the procedure used in Cole et al. (2014; see Supplemental Information). Specifically, we used the Louvain locally-greedy algorithm [Blondel et al., 2008, Jutla et al., 2011] for community detection. Data from the publically available Washington University-Minnesota Human Connectome Project "HCP100" dataset were used (N=100). Similar preprocessing procedures as used for the primary dataset were applied to the HCP100 dataset. Specifically, in addition to minimal preprocessing [Glasser et al., 2013], we ran a GLM nuisance regression using white matter, ventricles, and motion regressors (and their first derivatives). Global signal regression, motion scrubbing, and temporal filtering were not used. For each subject, all four resting state runs were concatenated and FC was estimated using standard Pearson correlations. The FC matrices were averaged across subjects to generate a group-mean resting-state FC matrix. We searched over two free parameters to find a community partition for the group-mean resting-state FC matrix. The first parameter was the density threshold, whereby weak connections (based on the absolute value of FC strengths) were removed prior to running the community detection algorithm. The second parameter was the structural resolution parameter, which can be used to tune the number of communities identified in the FC matrix. The parameter search was conducted across combinations of these two parameters (density of 40% to 100% in increments of 5%, and resolution of 0.8 to 3 in increments of 0.05), with two criteria: 1) there should be a peak of partition similarity (z-score of the Rand coefficient) [Traud et al., 2011] among adjacent locations in this two-dimensional parameter space, and 2) there should be distinct communities corresponding to visual, auditory, dorsal attention, default-mode, and motor/tactile systems (given decades of neuroscience research demonstrating their existence). Approximate locations of these systems were based on standard neuroscientific knowledge of these systems (given their strong establishment in the literature), in addition to their identification using resting-state FC in previous reports [Power et al., 2011, Gordon et al., 2014, Yeo et al., 2011]. A five-community partition had the highest nearest-neighbor similarity in parameter space, but this did not separate out the auditory system. The next-highest nearest-neighbor similarity peak (density = 100%, resolution = 1.2) with distinct communities corresponding to auditory, visual, dorsal attention, default-mode, and motor/tactile systems was a 14-community partition. This partition was then visualized using Connectome Workbench software (Figure 2.3a). Labels were assigned to the seven most replicated networks identified using resting-state FC [Power et al., 2011, Gordon et al., 2014, Yeo et al., 2011]. Colors were assigned to networks based on the colors used by Power et al. (2011).

A.2.3 Neural network model

To validate our information transfer estimation approach we constructed a simple dynamical neural network model with similar network topological properties identified in our empirical fMRI data. We constructed a neural network with 250 regions, each of which were clustered into one of five network communities (50 regions per community). Regions within the same community had a 35% probability of connecting to another region (i.e., 35% connectivity density), and regions not assigned to the same community were assigned a connectivity probability of 5% (i.e., 5% out-of-network connectivity density). We selected one community to act as a “network hub”, and increased the out-of-network connectivity density of those regions to 20% density. We then applied Gaussian weights on top of the underlying structural connectivity to simulate mean-field synaptic excitation between regions. These mean-field synaptic weights were set with a mean of $\frac{1}{\sqrt{K}}$ with a standard deviation of $\frac{0.2}{\sqrt{K}}$, where K is the number of synaptic inputs into a region such that synaptic input scales proportionally with the number of inputs.

This approach was recently shown to be a plausible rule in real-world neural systems based on in vitro estimation of between-neuron synaptic-weight-setting rules [Barral and Reyes, 2016].

To simulate network-level firing rate dynamics, as similar to Stern et al. (2014), region x_i's dynamics for $i = 1, 2, ..., 250$ obeyed the equation

$$\frac{dx_i}{dt}\tau_i = -x_i(t) + s\phi(x_i(t)) + g\big(\sum_{j \neq i}^{N} W_{ij}\phi(x_j(t))\big) + I_i(t) \tag{A.1}$$

We define the transfer function ϕ as the hyperbolic tangent, x_j the dynamics of region $j = 1, 2, ..., 250$ for $i \neq j$, $I_i(t)$ the input function (e.g., external spontaneous activity alone or both spontaneous activity and task stimulation) for $i \in [1, 250]$, W the underlying synaptic weight matrix, s the local coupling (i.e., recurrent) parameter, g the global coupling parameter, and τ_i the region's time constant. For simplicity, we set $s = g = 1$ and $\tau_i = 10ms$, though we show in a previous study [Cole et al., 2016a] that the activity flow mapping breaks down for parameter regimes $s >> g$.

We first simulated spontaneous activity in our model by injecting Gaussian noise (parameter $I_i(t)$; mean of 0.0, standard deviation 1.0). Numerical simulations were computed using a Runge-Kutta second order method with a time step of $dt = 10ms$. We ran our simulation for 600 seconds (10 minutes). To simulate resting-state fMRI, we then convolved our time series with the SPM canonical hemodynamic response function and down sampled to a 1 second TR, resulting in 600 time points. We then computed resting-state FC using multiple linear regression. To replicate the empirical data, we computed the BGC of the resting-state data (as in the empirical data; see equation 2) to validate that widespread out-of-network connectivity was preserved from synaptic to FC.

To model task-evoked activity, we simulated four distinct task conditions by injecting stimulation into four randomly selected but distinct sets of twelve regions in the hub network. Stimulation to the hub network was chosen to mimic four distinct top-down, cognitive control task rules. Task stimulation coincided with spontaneous activity (e.g., for time points t during a task, $I(t)$ = spontaneous activity at $t + 0.5$ constant task stimulation). We ran each task for 20 blocks, where each block lasted for 100 seconds. Each block contained five trials, each lasting for five seconds with an inter-trial interval of 15 seconds. In total, each task condition contained 100 task trials, with 500 seconds per task total. We then convolved these task time series with the SPM canonical hemodynamic response function and down sampled to 1-second TRs, as in the resting-state simulation. We simulated 30 subjects worth of data, and generated figures using group t-tests and controlled for multiple comparisons using FWE-correction permutation tests [Nichols and Holmes, 2001].

We validated the usefulness of the model for characterizing hub-related dynamics by testing whether estimated resting-state FC preserved the hub network's higher out-of-network intrinsic FC (specified by its underlying synaptic connectivity) by computing each network's BGC. BGC was computed in the same way as in the empirical data (see equation 2) for each of the network model's communities. For each of the five networks, we compared the BGC between each network using a cross-subject t-test. We corrected for multiple comparisons using FWE permutation tests [Nichols and Holmes, 2001] and significance was assessed with an FWE-corrected $p < 0.05$ threshold.

To perform network-to-network information transfer mapping in the model, we used the task-evoked activity (estimated by standard GLM beta estimates), and performed the information transfer mapping procedure between networks of regions using the resting-state FC matrix obtained via multiple linear regression. Network-to-network information transfer mapping is computationally identical to region-to-region information transfer mapping, and is described below. The information transfer mapping matrix (Figure 2.4e) was obtained using an FWE-corrected threshold of $p < 0.05$.

We primarily focused on stimulating the hub network to mimic top-down processes, since our empirical results focused on task-rule manipulations irrespective of stimuli presentations and motor responses. However, to demonstrate the generality with which information transfer can occur, we performed an additional set of simulations that focused on demonstrating that information transfer occurs with simultaneous top-down (hub network) and bottom-up (local network) stimulation. Using the same parameters as in the original simulation, we first replicated the same results as in Figure 2.4e with hub network stimulation only (i.e., top-down control). To simulate top-down and bottom-up activation we simulated four task conditions by injecting activity into four sets of regions. For each task condition, we simultaneously injected two sets of 12 regions; one set of 12 regions in the hub network (mimicking top-down activity), and one set of 12 regions in a local network (mimicking bottom-up activity). Each task condition stimulated a set of regions belonging to a different local network and a distinct set of regions in the hub network. Aside from task stimulation, all other model and simulation parameters were kept the same from the simulation result in Figure 2.4.

Our results were highly similar to the previous results, demonstrating that in both the top-down-only task and the simultaneous top-down and bottom-up task, information transfers between the hub and local networks were the strongest (Supplementary Figure A.3a,c). However, statistical testing demonstrated that some local-network-to-local-network information transfers were significant (after correcting for multiple comparisons; Supplementary Figure A.3d,h). We believe these effects are likely due to the existence of random (albeit sparse) connections between local networks. We also show that the predicted-to-actual similarity analysis portion of the information transfer procedure (described below) can be substituted with support vector machine (SVM) classification (Supplementary Figure A.3e-h; see below for details).

A.2.4 Computing baseline information estimates for regions and networks

To compute the baseline (i.e., unrelated to FC) information content at the region level (Figure 2.5), we performed a within-subject, cross-validated multivariate pattern analysis using representational similarity analysis for every [Glasser et al., 2016a] parcel (using the vertex-level multivariate activation pattern within each parcel). We estimated task-activation beta coefficients separately for each vertex within a region, and separately for each miniblock. Note that each miniblock was associated with a specific task-rule condition for each rule domain. Mathematically, we defined IE_B, the information estimate of region B, as

$$IE_B = Match_B - Mismatch_B \tag{A.2}$$

where $Match_B$ and $Mismatch_B$ correspond to the averaged Spearman rank correlation for matched and mismatched conditions, respectively. Specifically, we define $Match_B$ and $Mismatch_B$ as

$$Match_B = \frac{\sum_{k=1}^{K} scorr(B_k, B_{match})}{K} \tag{A.3}$$

$$Mismatch_B = \frac{\sum_{k=1}^{K}[\sum_{n=1}^{N}(scorr(B_k, B_{mismatch_n})/N]}{K} \tag{A.4}$$

where K corresponds to the total number of miniblocks (in this paradigm, 128 miniblocks), $scorr$ corresponds to a Fisher z-transformed Spearman's rank correlation between two activation vectors, B_k is the activation pattern in region B during block k, B_{match} is the task-rule condition prototype (obtained by averaging across blocks of the same condition, holding out block k) of region B's activation pattern for which block k's condition matches the condition prototype, and $B_{mismatch_n}$ as the task-rule condition prototypes for which block k's condition does not match. (In the present study $N = 3$, since each rule dimension has four task-rule conditions, and for a given miniblock there's one match and three mismatched conditions.) To avoid circularity, we performed a leave-four-out cross-validation scheme, holding out a miniblock of each task-rule. This ensured that miniblock B_k was not included in constructing the condition prototype B_{match} and that condition prototypes were each constructed using the same number of miniblocks. Prior to running the representational similarity analysis, all blocks were spatially demeaned to increase the likelihood that the representations we were identifying was a multivariate regional pattern (rather than a change in region-level mean activity). Use of Spearman's rank correlation also reduced the likelihood that the identified multivariate representation patterns were driven by mean activity changes or a small number of outlier values.

Statistical significance was assessed by taking a one-sided group t-test against 0 for each region's information estimate across subjects, since a greater than 0 difference of matches versus mismatches indicated significant representation of specific task rules. All p-values were corrected for multiple comparisons across the 360 parcels using FWE-correction with permutation tests15, and significance was assessed using an FWE-corrected threshold of $p < 0.05$.

For network-level information estimates (Supplementary Figure A.1b), the same cross-validated representational similarity analysis procedure was conducted for the seven functional networks separately across the three rule domains, using region-level representations within each of the networks. Region-level beta estimates were obtained for every block by fitting the same GLM model as described above to every region separately. All p-values were FWE-corrected for multiple comparisons across seven networks with permutation tests15, and significance was assessed using an FWE-corrected $p < 0.05$.

A.2.5 Region-to-region information transfer mapping

We extended the original activity flow mapping procedure as defined in [Cole et al., 2016a] (Figure 2.1a) to investigate transfer of task-related information between pairs of brain regions using vertex-wise activation patterns (i.e., region-to-region activity flow mapping; Figure 2.1b). This involved predicting the activity of the vertices of a held-out target region based on the vertices within a source region. Mathematically, we define region-to-region activity flow mapping between regions A and B as

$$\bar{B}_k = A_k \cdot W_{RSFC} \tag{A.5}$$

where $\bar{B}_k$ corresponds to the predicted activation pattern vector for the target region B, A_k corresponds to region A's activation pattern vector (i.e., the source region), W_{RSFC} corresponds to the vertex-to-vertex resting-state FC between regions A and B, and the operator $\cdot$ refers to the dot product. This formulation allowed us to map activation patterns in one region's spatial dimension to the spatial dimension of another region.

To test the extent that task representations are preserved in the region-to-region multivariate predictions, we quantified how much information transfer occurred between the two regions. Briefly, information transfer mapping comprises three steps, illustrated in Figure 2.1c: (1) Region-to-region (or network-to-network) activity flow mapping; (2) A cross-validated representational similarity analysis between predicted activation patterns and actual, held-out activation patterns; (3) Information classification/decoding by computing the difference between matched condition similarities and mismatched condition similarities. This final step produces an information transfer estimate.

Mathematically, our information transfer estimate was derived using almost the exact formulation as our information estimate formula. Specifically, we defined information transfer between regions A and B, or ITE_{AB}, as

$$ITE_{AB} = Match_B - Mismatch_B \tag{A.6}$$

where $Match_{AB}$ and $Mismatch_{AB}$ correspond to the averaged Spearman rank correlation for matched and mismatched conditions using the source region A, respectively. Similarly to equations A.3 and A.4, we define $Match_{AB}$ and $Mismatch_{AB}$ as

$$Match_{AB} = \frac{\sum_{k=1}^{K} scorr(\bar{B}_k, B_{match})}{K} \tag{A.7}$$

$$Mismatch_{AB} = \frac{\sum_{k=1}^{K}[\sum_{n=1}^{N}(scorr(\bar{B}_k, B_{mismatch_n})/N]}{K} \tag{A.8}$$

where K corresponds to the total number of miniblocks, *scorr* corresponds to a Fisher z-transformed Spearman's rank correlation between two vectors, $\bar{B}_k$ as the predicted activation pattern in the target region B (using region A's activation pattern) for block k, B_{match} as the condition prototype (obtained by averaging across blocks of the same condition, holding out block k) of the target region B's actual activation pattern for which block k's condition matches, and $B_{mismatch_n}$ as the condition prototypes for which block k's condition does not match. (In the present study N=3, since each rule dimension has four task-rule conditions.) As with the previously defined information estimate, we performed a leave-four-out cross-validation scheme, holding out a miniblock of each task-rule. This ensured that the actual activation pattern B_k of the predicted miniblock $\bar{B}_k$ was not included in constructing the condition prototype B_{match}. Prior to running the representational similarity analysis, all blocks were spatially demeaned to increase the likelihood that the representation we were identifying was a multivariate regional pattern (rather than a change in region-level mean activity). This formulation allowed us to quantify how much "information transfe" occurred between two regions by comparing the predicted activation pattern in the target region to the actual activation pattern in the

We also demonstrate that the predicted-to-actual similarity analysis in our information transfer mapping procedure can be substituted with an SVM decoding scheme. Specifically, we show in our computational model that we could train a linear classifier on the target region's predicted activation patterns that could decode the actual, activation patterns in that target region (Supplementary Figures A.3e,f). We used the same leave-four-out cross-validation scheme as above to obtain these results, and we find that the information transfer mapping results with SVM decodings (Supplementary Figure A.3f) are identical to using representational similarity analysis (Figure 2.4e).

Note that information decoding was performed on the cortical surface, using vertices rather than voxels. This vertex-wise approach has been shown to provide better multivariate classifications than voxel-wise information decoding [Oosterhof et al., 2011], likely because surface analyses better reflect the underlying cortical anatomy.

Information transfer mapping was performed within subject between every pair of regions in the [Glasser et al., 2016a] atlas (360 regions in total). The results of this approach between all region pairs were then visualized via a 360-by-360 matrix (a total of 129,240 region-to-region mappings), where the regions along rows (source regions) indicated the activation patterns used to map onto a target region's activation pattern, which was indicated along the columns (Figures 2.6b,d,f). Statistical tests were performed using a group one-sided t-test ($t > 0$) for every pair-wise mapping. A one-sided t-test was appropriate here given that our hypotheses were implicitly one-sided, since any significant deviation above 0 indicated a significantly higher matched versus mismatched correlation between predicted-to-actual activation patterns (i.e., the information transfer estimate).

Our use of mismatched correlations as a baseline ensured that any positive information transfer estimates was a result of a task-rule-specific representation, rather than a task-general effect. Any information estimate that was not significantly greater than 0 indicated that the predicted-to-actual similarity was at chance (akin to chance decoding using classifiers). We tested for multiple comparisons using permutation testing [Nichols and Holmes, 2001] for every region-to-region mapping, and significance was assessed using FWE-corrected p-values with $p < 0.05$. Note that to avoid circularity for region-to-region information transfer mapping, any vertices in a source region that fell within a 10mm radius of the to-be-predicted target region (e.g., an adjacent region) would not contribute any activity flow to the to-be-predicted target region (see FC estimation Methods section for details).

Given the visual sparsity of the region-to-region information transfer mapping visualization, we opted to down sample our matrix to provide a simpler visualization to assess how pairs of regions transfer information between and within functional networks (Figures 2.6c,e,g). Thus, we computed the percent of statistically significant transfers for every pair of networks. This allowed us to better visually assess how region-to-region information transfer mappings may have been influenced by underlying network organization. To compute the percent of statistically significant transfers, we counted the number of significant transfers between every pair of networks and divided that by the total number of possible transfers within that network-to-network configuration. To characterize the generality with which information transfer mappings occurred between specific network configurations, we computed the number of rule domains in which each network configuration contained at least one region-to-region transfer (Figure 2.6h). In other words, we took the matrices in Figures 2.6c,e,g and binarized them with a 1 if a cell had a greater than 0 percentage of transfers, and a 0 otherwise. We then summed these matrices element-wise to obtain the number of rule domains each network configuration had a successful information transfer in. To assess the number of rule domains each network contained at least one successful source region, we took the percent of significant transfers from each network to any other region in the brain (a 7-element array) and then binarized the array for each rule domain. We then summed across the three arrays (one for each rule domain) to obtain the number of rule domains each network had at least one successful source region used for information transfer (Figure 2.6i).

Lastly, to visualize the anatomical locations of the source regions for information transfer, we computed the percent of significant transfers from each cortical region for each rule domain (Figure 2.7). Percentages were obtained by taking the number of successful transfers from a region, and dividing it by total number of possible transfers (i.e., 359 other regions). We then plotted each of these percentages on the cortical surface using Connectome Workbench software (version 1.2.3) for each rule domain [Glasser et al., 2016b].

A.2.6 Network-to-network information transfer mapping

Network-to-network information transfer mapping in both the computational model (Figure 2.4e) and empirical data (Supplementary Figures A.1c,d,e) was performed in the same computational framework as above, though instead of predicting region-level activation patterns using vertex-level activation patterns, network-level activation patterns were predicted using region-level activations (averaging across vertices within a given region). In other words, when predicting a target network B's region-level activation pattern, we computed the dot product between a source network A's region-level activity vector and the region-to-region resting-state FC matrix between regions in network A and B. We then submitted our 128 task block predictions for network B to our information transfer mapping procedure, as described above. This was repeated for every pair of the seven functional networks defined by our community-detection algorithm, resulting in 7-by-7 network-to-network mappings which were visualized as a 7x7 matrix (Supplementary Figures A.1c,d,e). We tested for multiple comparisons using FWE-correction for every network-to-network mapping within a rule domain, and significance was assessed using the FWE-corrected p-values of $p < 0.05$.

A.2.7 Permutation testing of FC topology

We hypothesized that the precise topology of resting-state FC described the baseline architecture of information processing during task states. Thus, to ensure that our information transfer mapping procedure depended on resting-state FC topology, we performed permutation testing, shuffling the network-to-network FC topology prior to performing information transfer mapping. Due to computational cost, we limited this control analysis to network-to-network information transfer mapping.

For each subject, we permuted the network-to-network resting-state FC prior to applying the information transfer mapping procedure for every pair of networks. More specifically, each network's connectivity was permuted within-network, such that no FC values from one network was ever moved to another network. This helped ensure that the permutations only altered the network-to-network FC topology, such that (for example) the overall mean level of FC between the networks was never altered across the permutations. To correct for multiple comparisons, a single permutation cycle involved permuting the FC topology for every pair of networks, for all subjects. We then performed a group t-test for every pair of network-to-network information transfers, extracting the maximal t-statistic across all network-to-network comparisons. We ran 1000 of these permutation cycles, obtaining the maximal t-statistic for each permutation. This formed a null distribution of the maxima across the family of tests (i.e., all possible network-to-network information transfers), thus controlling for FWE [Nichols and Holmes, 2001]. Using our permutation distribution, we computed

FWE-corrected p-values with a one-tailed test, i.e., $p = P(X > ite)$, where ite corresponds to the true information transfer estimate, and X as the null distribution of maximal t-statistics. Statistical significance was then assessed using a FWE-corrected threshold of $p < 0.05$.

A.2.8 Behavioral relevance of information transfers

To characterize the behavioral relevance of information transfers, we performed a within-subject analysis to decode task performance using miniblock-by-miniblock information transfer estimates. We first sought to ensure that baseline miniblock information estimates could decode miniblock task performance within subjects prior to the information transfer mapping procedure. We defined miniblock information estimates as

$$IE_{X_k} = match_{X_k} - mismatch_{X_k} \tag{A.9}$$

where IE_{X_k} corresponds to the information estimate of rule domain X during miniblock k, $match_{X_k}$ corresponds to the matched task-rule condition similarity of rule domain X during miniblock k, and $mismatch_{X_k}$ corresponds to the averaged rank correlation of miniblock k's activation pattern to the mismatched task-rule conditions.

To perform a given task, knowledge of all three rule domains (i.e., logic, sensory, and motor rule domains) is required. Thus, we constructed a decoding model with logistic regression, training the model to decode the task performance of a given miniblock using the information estimates of a given brain region across all three rule domains. The model was tested using cross-validation in MATLAB using the glmfit function (with the logit link function), and was formulated as

$$\overrightarrow{y}_{accuracy} = f(\beta_0 + \beta_1 X_L + \beta_2 X_S + \beta_3 X_M) \tag{A.10}$$

where $\overrightarrow{y}_{accuracy}$ corresponds to the vector containing task accuracy for all miniblocks, X_L, X_S, X_M correspond to the regressors for logic, sensory, and motor information estimates, respectively, β_0 corresponds to the training bias (which accounts for the imbalance of the correct:error trial ratio), and β_1, β_2, β_3 correspond to the estimated model coefficients for the logic, sensory, and motor information estimates, respectively. The link function f corresponds to the sigmoid function, defined as

$$f(x) = \frac{1}{1 + e^{-x}} \tag{A.11}$$

Miniblocks with over 50% of trials performed correctly were characterized as 1, and 0 otherwise.

To test our model, we used cross-validation to predict the binarized accuracy of held-out data. However, to account for the imbalanced training data (on average, subjects performed 85% of trials correctly), we removed the intercept term β_0 to center our predictions (as computed by a sigmoid function) at 0.5. Thus, our predictions on held-out data were computed as probabilities by the equation

$$g(x_L, x_S, x_M) = f(\beta_1 X_L + \beta_2 X_S + \beta_3 X_M) \tag{A.12}$$

where $g \in (0, 1)$ and accuracies were predicted/classified by the equation

$$G_{decoder}(g(x)) = \begin{cases} 1 & g(x) > 0.5 \\ 0 & g(x) < 0.5 \end{cases} \tag{A.13}$$

where $G_{decoder}$ generates predictions for miniblocks with greater than 50% task performance as 1, and 0 otherwise.

Given that region-to-region information transfers consistently occurred between regions in the FPN and CON across all three rule domains (Figure 2.6h), we constrained our search to those networks. We applied our decoding model to all regions within the FPN and CON across subjects. For each region, we applied one-sided t-tests against chance (50%), and corrected for multiple comparisons using FWE-correction permutation tests [Nichols and Holmes, 2001]. We identified a single FPN region in the LPFC (LH region 80 in the Glasser et al. atlas; Supplementary Figure A.5) whose baseline information estimates predicted miniblock task performance.

We subsequently tested whether information transfer estimates from the LPFC region could predict task performance. We applied the decoding model to information transfer estimates across all rule domains (instead of baseline information estimates) for all information transfers from the LPFC region to all other FPN and CON regions. (We used the LPFC region here as the "source" region, obtaining decoding accuracies from that region to all other FPN/CON regions.) We performed one-sided t-tests against chance (50%) for each information transfer, and corrected for multiple comparisons using FWE-correction permutation tests [Nichols and Holmes, 2001]. We identified a single information transfer from the LPFC to the OFC (LH region 91; both FPN regions) that survived multiple corrections with an FWE-corrected $p < 0.05$. Surface visualizations for Supplementary Figure A.5 were made using Connectome Workbench software (version 1.2.3) [Glasser et al., 2016b].

Appendix B

Appendix – Chapter 3

B.1 Supplementary Figures

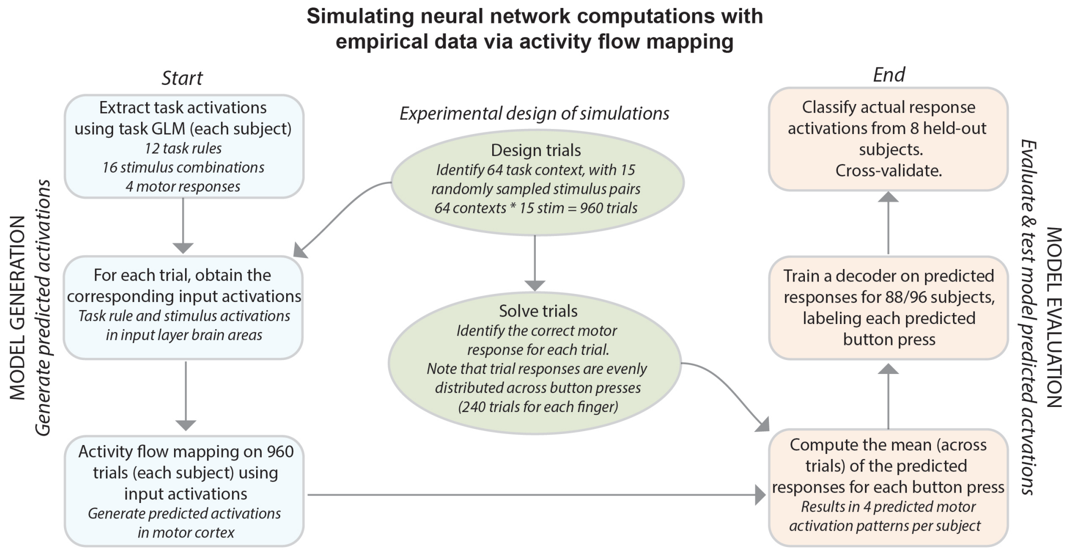

Simulating neural network computations with empirical data via activity flow mapping
Start
Extract task activations using task GLM (each subject)
12 task rules
16 stimulus combinations
4 motor responses
For each trial, obtain the corresponding input activations
Task rule and stimulus activations in input layer brain areas
Activity flow mapping on 960 trials (each subject) using input activations
Generate predicted activations in motor cortex
MODEL GENERATION
Generate predicted activations
Experimental design of simulations
Design trials
Identify 64 task context, with 15 randomly sampled stimulus pairs
64 contexts * 15 stim = 960 trials
Solve trials
Identify the correct motor response for each trial.
Note that trial responses are evenly distributed across button presses
(240 trials for each finger)
End
Classify actual response activations from 8 held-out subjects.
Cross-validate.
Train a decoder on predicted responses for 88/96 subjects, labeling each predicted button press
Compute the mean (across trials) of the predicted responses for each button press
Results in 4 predicted motor activation patterns per subject
MODEL EVALUATION
Evaluate & test model predicted actvations

Figure B.1: Flow chart describing neural network simulations with empirical data via activity flow mapping. We generate a subject's predicted motor response activations using only task rule and sensory stimulus activation patterns as inputs. We then test these predictions against the actual motor response activations of held-out subjects.

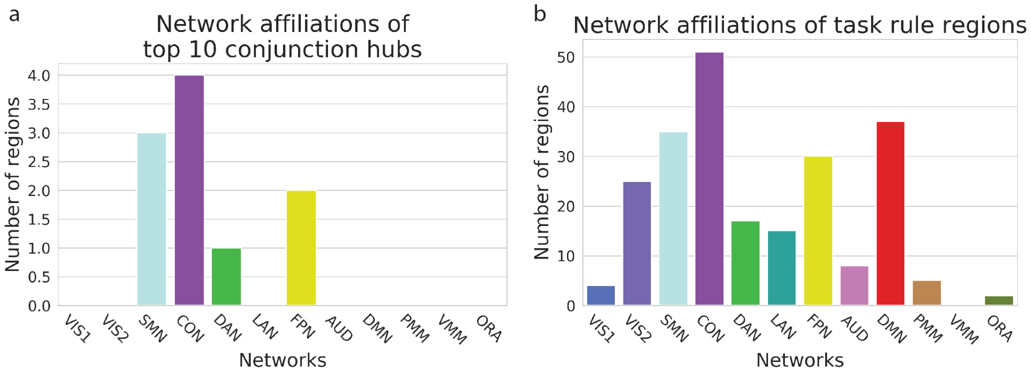


Figure B.2: Network affiliations of conjunction hubs and the task rule input layer using a previously defined multimodal atlas and network partition [Glasser et al., 2016a, Ji et al., 2019]. a) The network affiliations of the 10 conjunction hub brain areas. b) Network affiliations of the 228 brain regions that contained decodable task rule information.

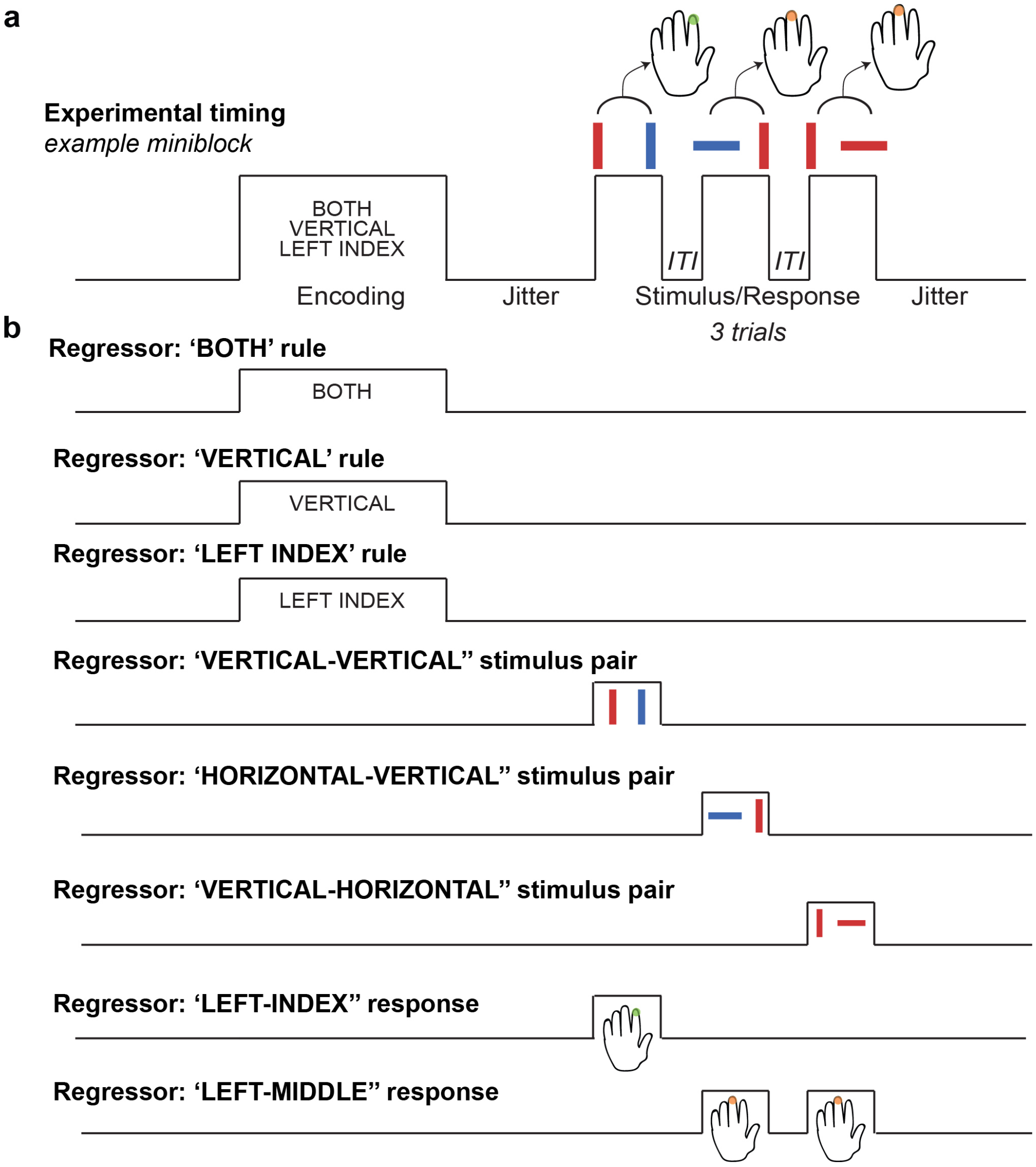

a
Experimental timing
example miniblock
BOTH
VERTICAL
LEFT INDEX
Encoding
Jitter
ITI
ITI
Stimulus/Response
3 trials
Jitter
b
Regressor: 'BOTH' rule
BOTH
Regressor: 'VERTICAL' rule
VERTICAL
Regressor: 'LEFT INDEX' rule
LEFT INDEX
Regressor: 'VERTICAL-VERTICAL" stimulus pair
Regressor: 'HORIZONTAL-VERTICAL" stimulus pair
Regressor: 'VERTICAL-HORIZONTAL" stimulus pair
Regressor: 'LEFT-INDEX" response
Regressor: 'LEFT-MIDDLE" response

Figure B.3: Example of task GLM approach to obtain task activation estimates. a) An example miniblock containing one encoding block (task rule set) and three trials. Note that while stimulus presentation and response periods overlap, they are not collinear. b) The regressors for the relevant task conditions in the example miniblock. We obtain regressors (estimated across all 128 miniblocks) for all task rule, sensory stimuli, and motor response conditions. Altogether there are 32 different task conditions (12 task rules, 16 sensory stimuli pairs, and four motor response periods). Note that task rule regressors (logic, sensory, and motor rule examples) appear collinear in this example, but that across all 128 miniblocks task rule conditions are properly counterbalanced to avoid collinearity. Regressors shown here are illustrated without convolution with SPM's canonical HRF.

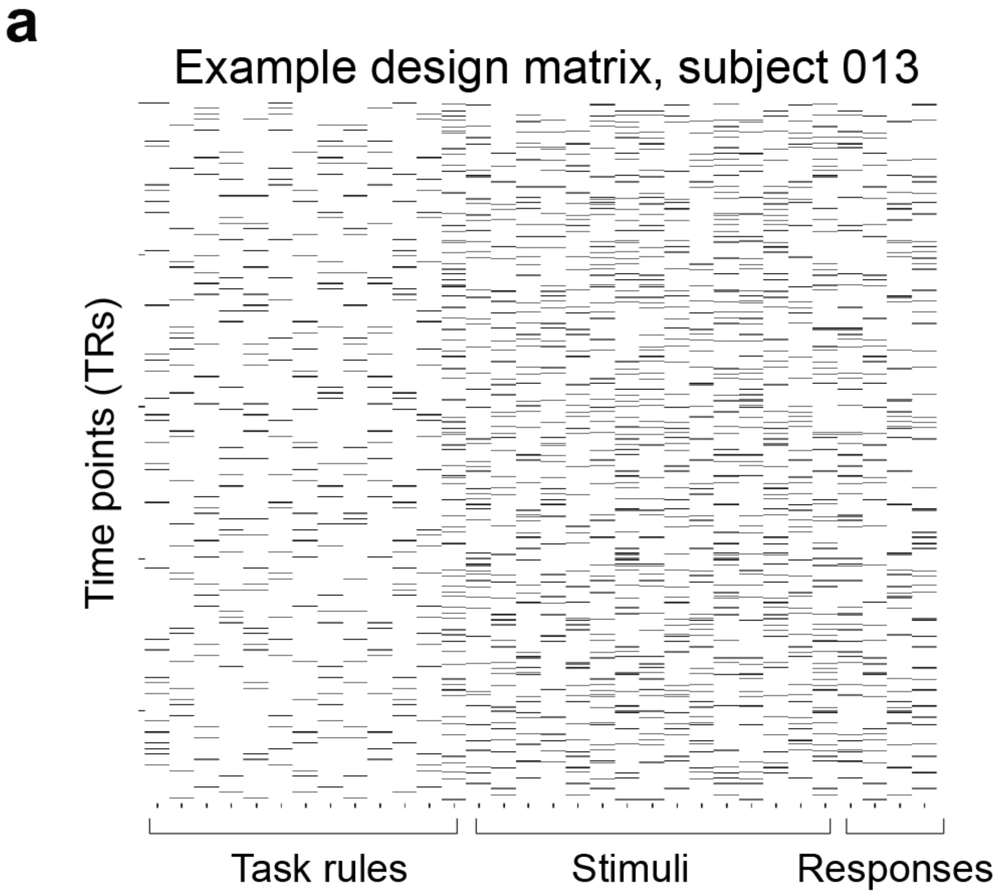


Figure B.4: A task GLM design matrix for an example subject.

Appendix C

Appendix – Chapter 5

This chapter contains the supplementary materials for [Ito et al., 2020a]

C.1 Supplementary Figures

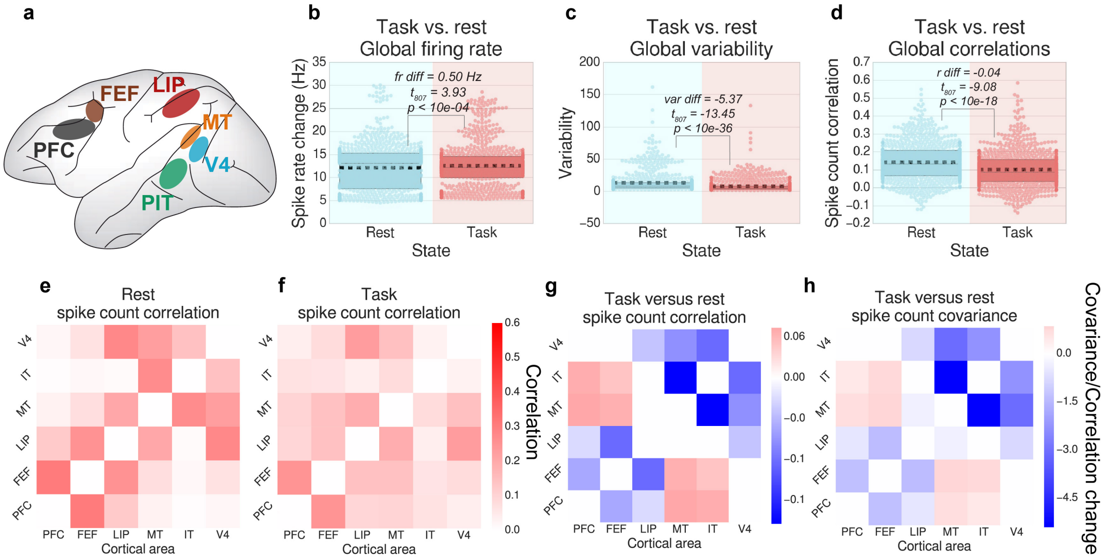


Figure C.1: Supplementary Figure 1. Replication analysis for the excluded NHP subject. This figure is organized identically to Fig 5.2, but using data from a replication subject. We find nearly identical patterns between the exploratory and replication subjects, with the exception that we did not replicate any correlation increases. a) Mean-field spike recordings from six different cortical regions fed into our analyses. b) As in our empirical fMRI data set, we calculated the global variability across task and rest states (estimated using the standard deviation across trials). c) We then calculated the global neural correlation (i.e., the spike count correlation across trials) for task and rest states between all pairs of recorded brain regions. (Spike rates were averaged within each cortical area.) d-f) For each pair of brain regions, we visualized the correlation matrices between each recording site for the averaged rest, task, and the differences between task versus rest state spike count correlations. For panels d-f, plots were thresholded and tested for multiple comparisons using an FDR-corrected $p < 0.05$ threshold. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot.

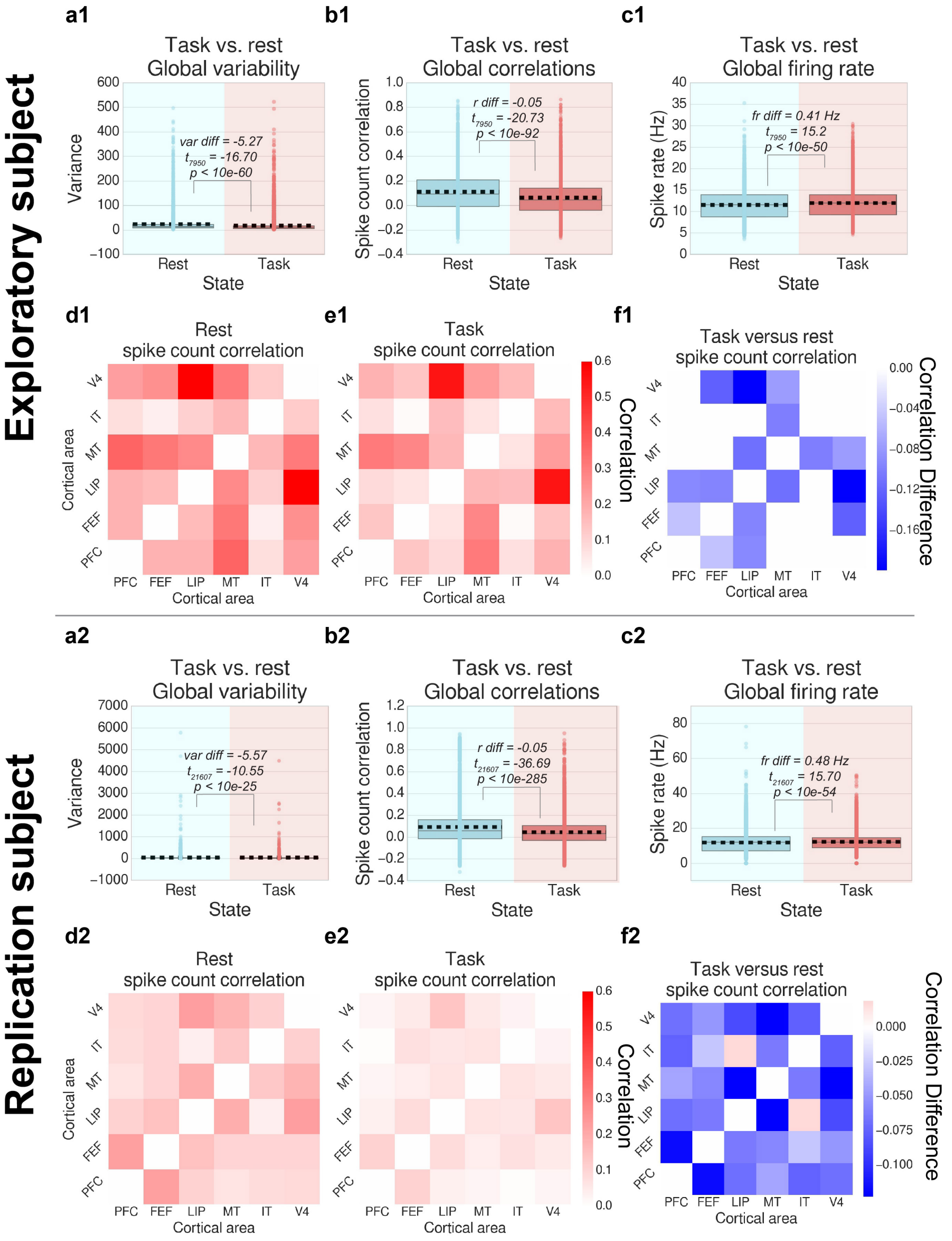


Figure C.2: Supplementary Figure 2. Neural correlations and variability are quenched within trials from rest to task intervals. We analyzed the variability across time points (within trial) during ITIs and task cue periods to evaluate whether correlation and variability quenching also occurred on a moment-to-moment basis (i.e., faster timescale). Task cue intervals and ITIs were matched to have equivalent time points on a trial-by-trial basis. a1,a2) Global variability across the two states (estimated using the variance across time points) between task and rest state windows. b1,b2) We then calculated the global spike count correlation between the exact same task cue intervals with equivalent rest intervals between all pairs of recorded brain regions. (Spike rates were averaged within each cortical area.) c1,c2) We also calculated the global firing rate (averaged across all recording areas) during the task interval and rest interval. d1-f1,d2-f2) For each pair of brain regions, we visualize the spike count correlation matrices between each recording site for the averaged rest, task, and the differences between task versus rest state spike count correlation. For panels d-f, plots were thresholded and tested for multiple comparisons using an FDR-corrected p¡0.05 threshold. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a strip plot.

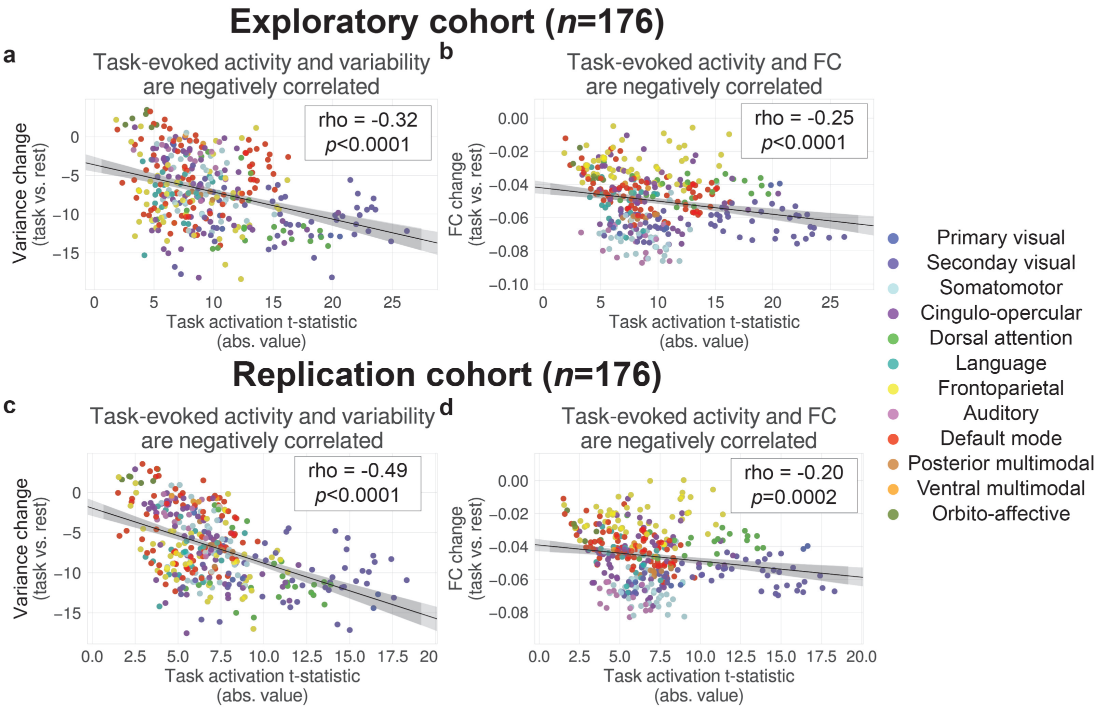


Figure C.3: Supplementary Figure 3. Task-evoked activity is negatively correlated with variability and correlations across regions in fMRI data. a) We replicated a previous result [He, 2013], demonstrating that regions that activated more during tasks tend to decrease their BOLD variability more during task states. b) We extended those results to evaluate the relationship between task-evoked activity and FC across regions. We found that regions that activated more during tasks tend to decrease their global functional FC accordingly during task states. Scatter plots reflect each parcel in the Glasser atlas [Glasser et al., 2016a], and are colored according to network affiliation [Ji et al., 2019]. Best fit lines were estimated using linear regression, but correlations were calculated using a non-parametric rank correlation. c,d) Replication of panels a,b, respectively using the replication cohort of subjects. Statistics were calculated using the same steps as in [He, 2013]. To calculate the averaged regional task activation, we first performed a group t-test for each task against 0, took the absolute value of the t-statistic, and then averaged across tasks. To calculate the averaged regional FC and SD, we performed a group t-test against 0 for each region. We then correlated these values across regions to measure the relationship between activity and FC, and activity and SD.

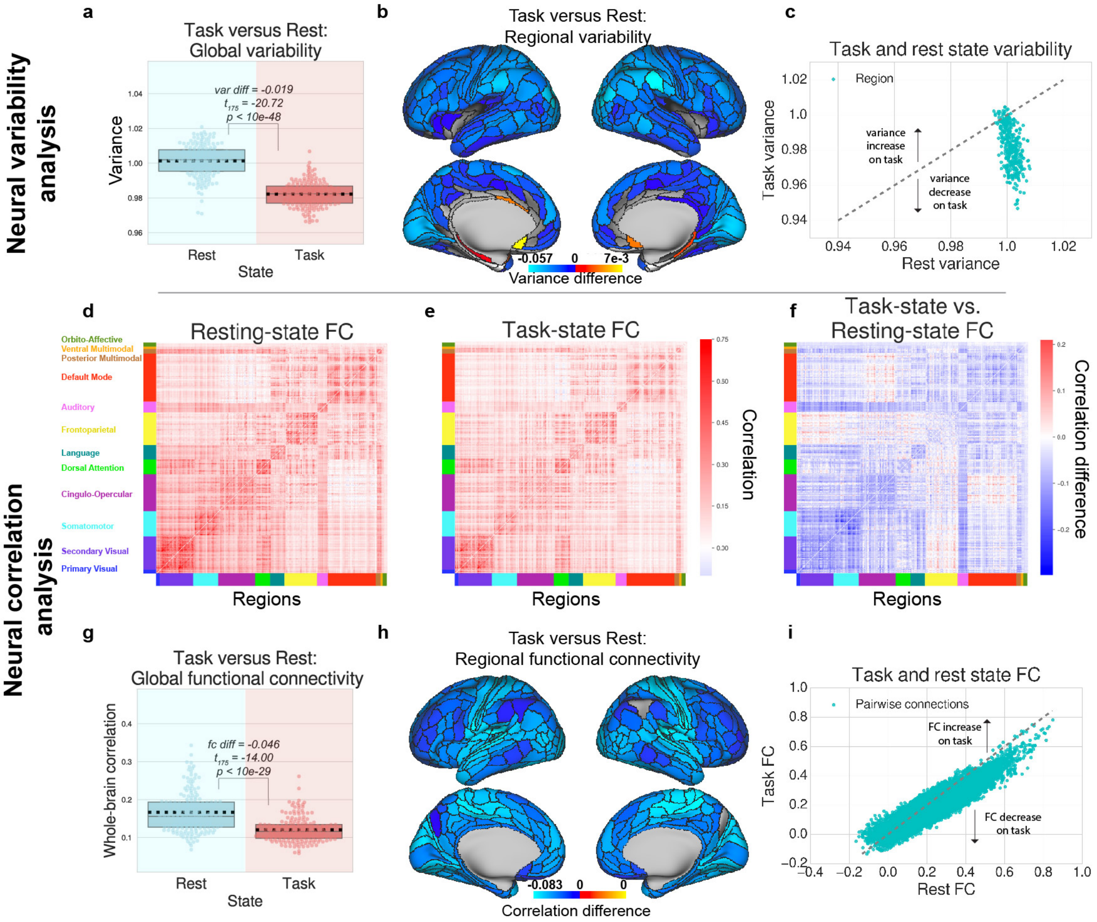


Figure C.4: Supplementary Figure 4. Replication data set: Variability and correlations decrease during task states in human fMRI data. We successfully replicated results from Fig 5.3 using our held-out cohort of 176 subjects. a) We first compared the global variability during task and rest states, which is averaged across all brain regions, and then b) computed the task- versus rest-state variability for each brain region. c) Scatter plot depicting the variance of each parcel during task states (y-axis) and rest states (x-axis). Dotted grey line denotes no change between rest and task states. d) We next compared the correlation matrices for resting state blocks with (e) task state blocks, and (f) computed the task- versus rest-state correlation matrix difference. g) We found that the average FC between all pairs of brain regions is significantly reduced during task state. h) We found that the average correlation for each brain region, decreased for each brain region during task state. i) Scatter plot depicting the FC (correlation values) of each pair of parcels during task states (y-axis) and rest states (x-axis). Dotted grey line denotes no change between rest and task states. For panels b-f, and h, plots were tested for multiple comparisons using an FDR-corrected $p < 0.05$ threshold. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot.

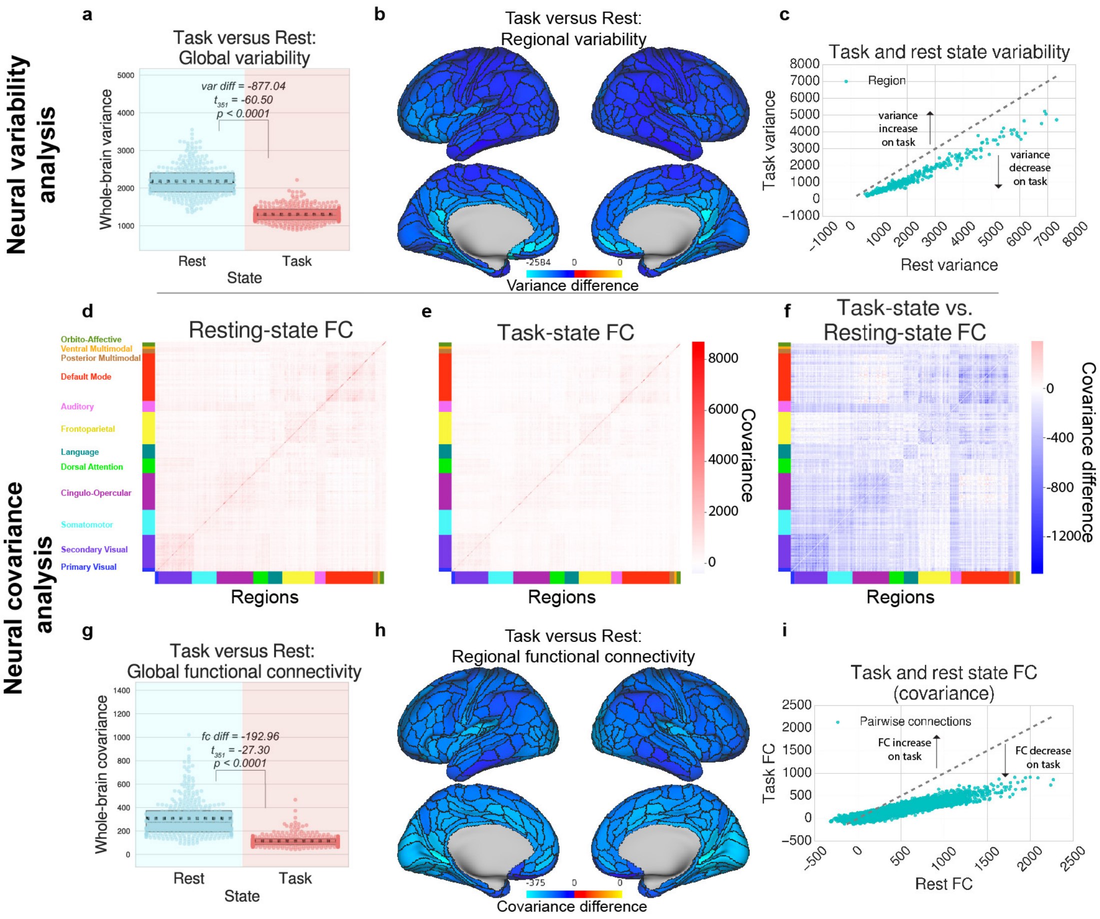


Figure C.5: Supplementary Figure 5. Non-normalized data using variance and covariance, using the full set of 352 subjects. Variance and covariance decreased during task states in human fMRI data. We successfully replicated results from Fig 5.3 using, but without z-normalizing the time series (and using covariance instead of correlation). The combination of reduced correlations (Fig 5.3) and covariance measures suggested that shared signal dynamics is reduced from task to rest [Duff et al., 2018, Cole et al., 2016b, Siegel et al., 2012]. a) We first compared the global variability during task and rest states, which is averaged across all brain regions, and then b) computed the task- versus rest-state variability for each brain region. c) Scatter plot depicting the variance of each parcel during task states (y-axis) and rest states (x-axis). Dotted grey line denotes no change between rest and task states. d) We next compared the covariance matrices for resting state blocks with (e) task state blocks, and (f) computed the task- versus rest-state covariance matrix difference. g) We found that the average covariance between all pairs of brain regions is significantly reduced during task state. h) We found that the average covariance for each brain region, decreased for each brain region during task state. i) Scatter plot depicting the FC (covariance values) of each pair of parcels during task states (y-axis) and rest states (x-axis). Dotted grey line denotes no change between rest and task states. For panels b-f, and h, plots were tested for multiple comparisons using an FDR-corrected $p < 0.05$ threshold. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, grey line indicates the median, and the distribution is visualized using a swarm plot.

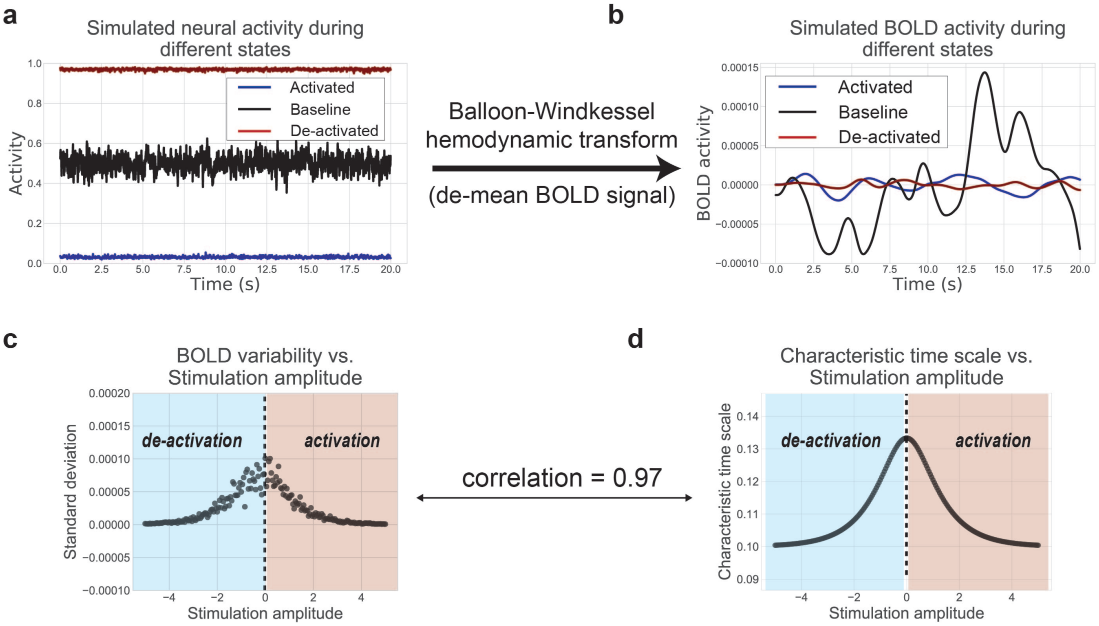


Figure C.6: Supplementary Figure 6. Task-state neural variability reduction is preserved in the BOLD signal in the neural mass model. a) We simulated the neural mass model under the same three stimulus conditions (de-activated, baseline, and activated states) as in Fig 5.7a. b) We subsequently applied the Balloon-Windkessel transformation to the simulated neural activity, a nonlinear transformation from neural activity to the fMRI BOLD signal [Friston et al., 2003]. Notably, the transformation assumes a nonlinear transformation of the normalized deoxyhemoglobin content, normalized blood inflow, resting oxygen extraction fraction, and the normalized blood volume. All BOLD signals were de-meaned such that it is possible to visually compare the time series variance of each stimulus condition. c) We simulated BOLD activity under a range of stimulus conditions and calculated the standard deviation of each time series. d) We calculated the rank correlation of the standard deviation of the BOLD signal across stimulus conditions with the characteristic time scale at each condition.

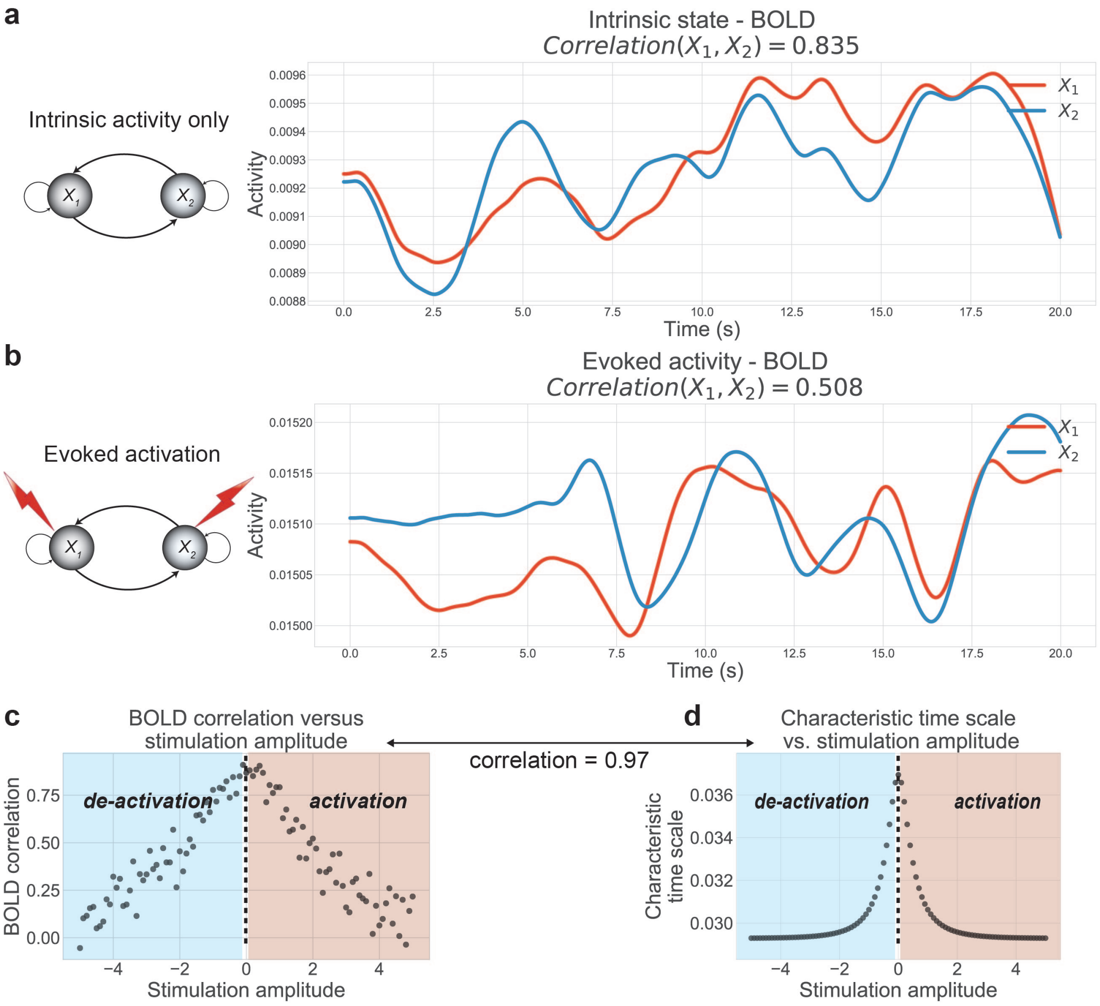


Figure C.7: Supplementary Figure 7. Task-state neural correlation reduction is preserved in the BOLD signal in the two-unit neural mass model. a-b) Using the simulated the neural mass data in Fig 5.8, we applied the Balloon-Windkessel transform to convert our neural data into BOLD data [Friston et al., 2003]. c) We simulated BOLD activity under a range of stimulus conditions and calculated the neural correlation between the two units of each. d) We calculated the rank correlation of the neural correlation of the BOLD signal across stimulus conditions with the characteristic time scale at each condition.

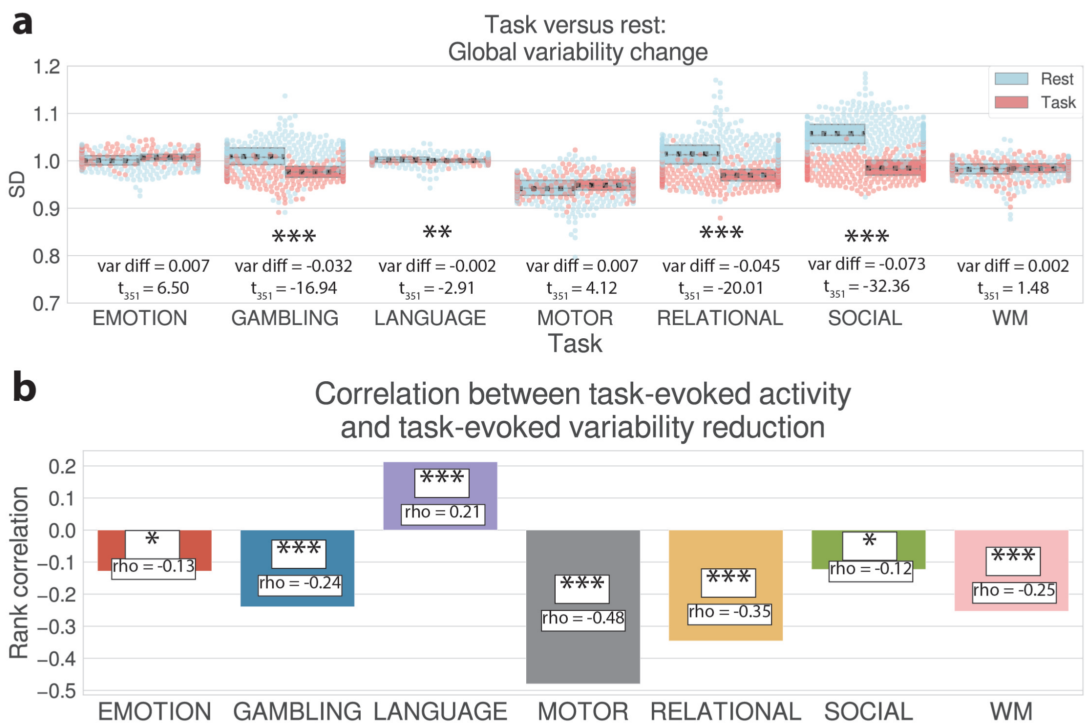


Figure C.8: Supplementary Figure 8. fMRI variability reduction analysis for each of the 7 HCP tasks separately. a) This panel is identical to the analysis performed in Fig 5.3a, except that it was performed on each HCP task separately. Global variability, averaged across all regions, was reduced for 4/7 of the HCP tasks. Global variability was not reduced for the Emotion and Motor tasks, though task-evoked activity was correlated with task-evoked variability reduction across space (see next panel). b) This panel is identical to the analysis performed in Supplementary Figure C.3, except that the spatial correlation was performed on each HCP task separately (and is visualized as a bar plot rather than a scatter plot). Regional task-evoked variability was significantly negatively correlated with the magnitude of task-evoked activation (absolute value) for 6/7 of the HCP tasks. All analyses (in panels a and b) were corrected for multiple comparisons using FDR correction. (*** = FDR-corrected $p < 0.0001$; ** = FDR-corrected $p < 0.01$; * = FDR-corrected $p < 0.05$). Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, and the distribution is visualized using a swarm plot.

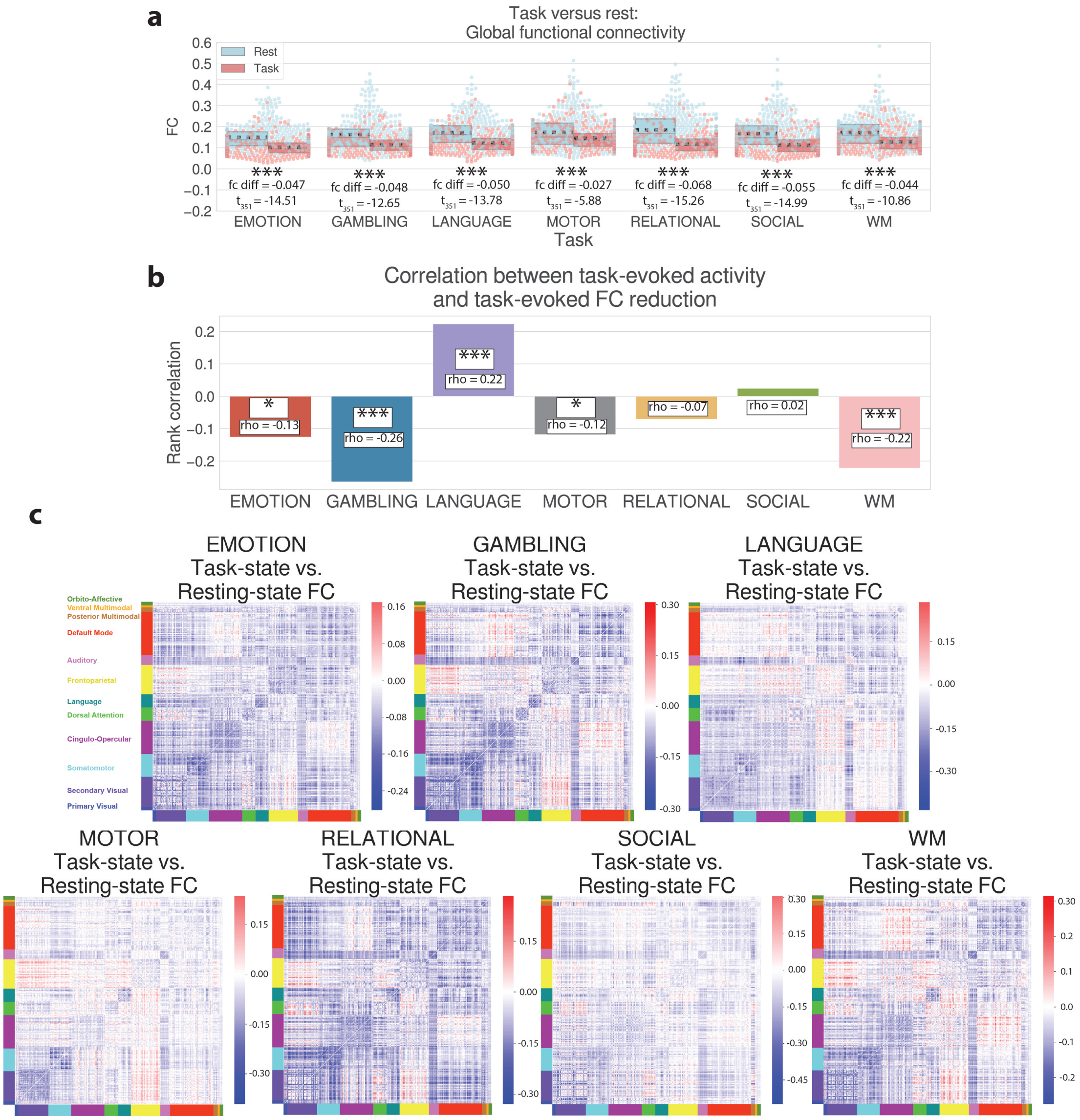


Figure C.9: Supplementary Figure 9. Task versus rest fcMRI analysis for each of the 7 HCP tasks separately. a) This panel is identical to the analysis performed in Fig 5.3G, except that it was performed on each HCP task separately. Whole-brain FC, averaged across all pairs of regions, was reduced for 7/7 of the HCP tasks. b) This panel is identical to the analysis performed in Supplementary Figure C.3, except that the spatial correlation was performed on each HCP task separately (and is visualized as a bar plot). Regional task-evoked FC was significantly negatively correlated with the magnitude of task-evoked activation (absolute value) for 4/7 of the HCP tasks. All analyses (in panels A and B) were corrected for multiple comparisons using FDR correction. (*** = FDR-corrected $p < 0.0001$; ** = FDR-corrected $p < 0.01$; * = FDR-corrected $p < 0.05$). c) Task- versus rest-state FC analysis for each of the 7 HCP tasks separately. (This figure is identical to Fig 5.3f, except that the statistics were performed on each task separately.) Though whole-brain FC differences from task to rest are different for each task, there are mostly FC decreases during task state relative to rest state. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, and the distribution is visualized using a swarm plot.

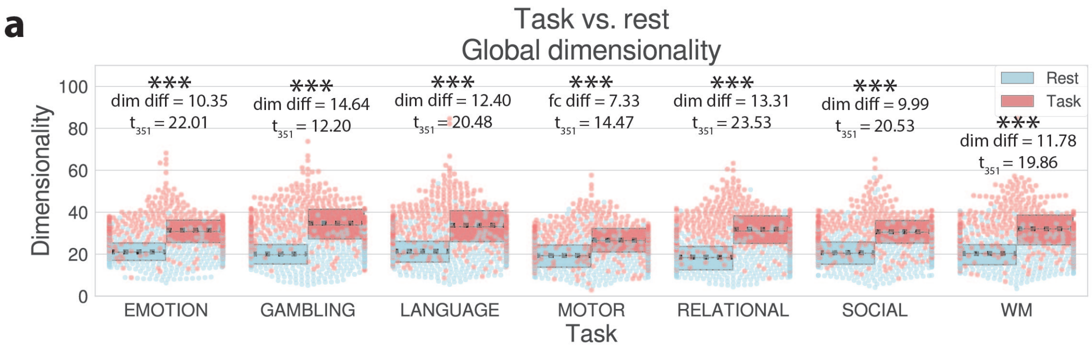


Figure C.10: Supplementary Figure 10. Task versus rest dimensionality comparison for each of the 7 HCP tasks separately. a) This panel is identical to the analysis performed in Fig 5.5a, except that it was performed on each HCP task separately. Whole-brain dimensionality increased from rest to task states for each of the 7 HCP tasks. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, and the distribution is visualized using a swarm plot. (*** = FDR-corrected $p < 0.0001$; ** = FDR-corrected $p < 0.01$; * = FDR-corrected $p < 0.05$)

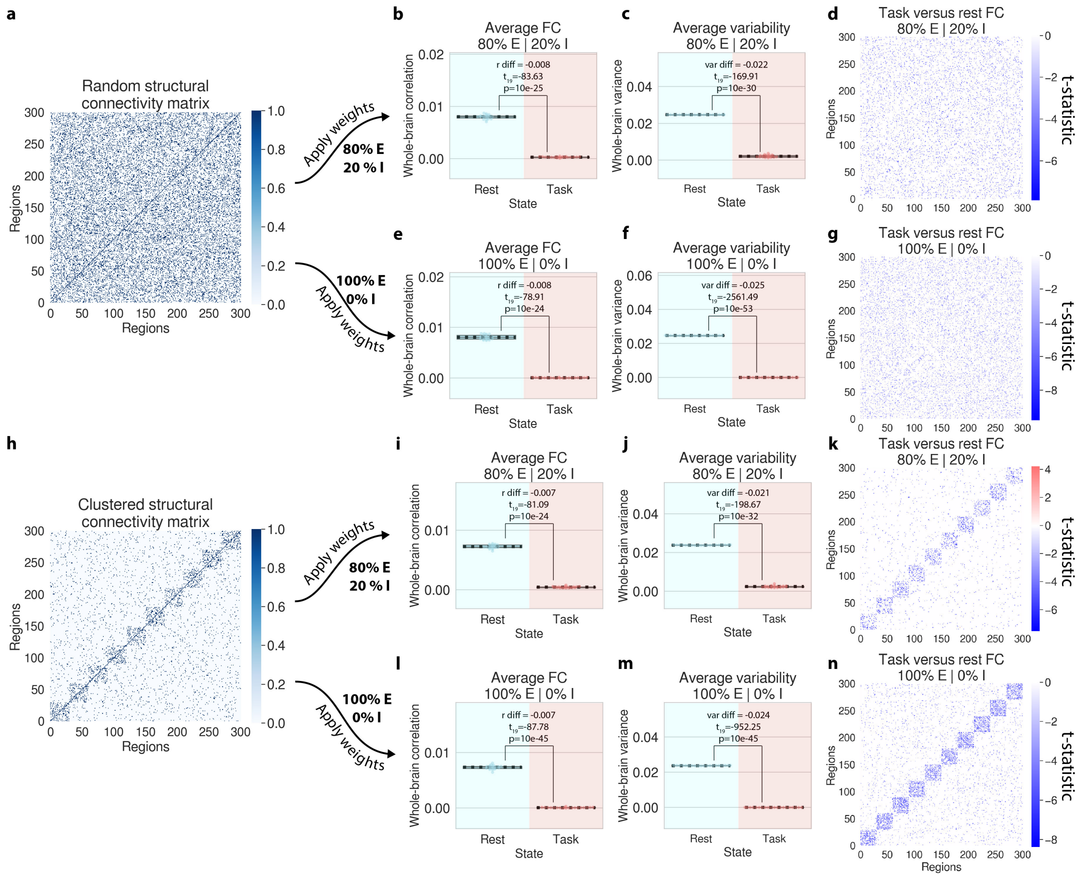


Figure C.11: Supplementary Figure 11. Variability and correlations are quenched in large-scale network models (300 regions) with both random and clustered structural connections. For each structural connectivity matrix, we randomly sampled synaptic weights from a normal distribution with either 100% E connections (given evidence that most long-range connections are excitatory, $\mu = 1.0$, $\sigma = 0.2$ [Joglekar et al., 2018]), or 80% E and 20% I connections ($\mu = 1.0$, $\sigma = 1.2$). For each network model (4 in total), we simulated 20 subjects for 10 seconds each (100ms sampling rate). For simplicity, during the task state, all units were stimulated with a fixed input. a) Random structural connectivity matrix (20% connectivity density) for an example subject. b) The average across all pairwise correlations during the rest and task states for the network model with 80% E and 20% I connections. The rest state exhibits higher correlations than the task state. c) The variability (variance across time) averaged across brain regions during the rest and task states for the network model with 80% E and 20% I connections. The rest state exhibits higher variability than the task state. d) The task minus rest FC matrix (correlation difference) between all 300 regions. Correlations decreased from rest to task states. e-g) The same analyses as b-d, but using only excitatory connections only. h) Clustered structural connectivity matrix (10 communities, 20% within-community density, 3% out-of-community density). i-k) The same analyses as b-d, but using the clustered connectivity matrix with 80% E and 20% I connections. l-n) The same analyses as b-d, but using the clustered connectivity matrix with 100% E connections. Boxplots indicate the interquartile range of the distribution, dotted black line indicates the mean, and the distribution is visualized using a swarm plot. Plots d, g, k, n, were corrected for multiple comparisons and thresholded using an FDR-corrected $p < 0.05$.

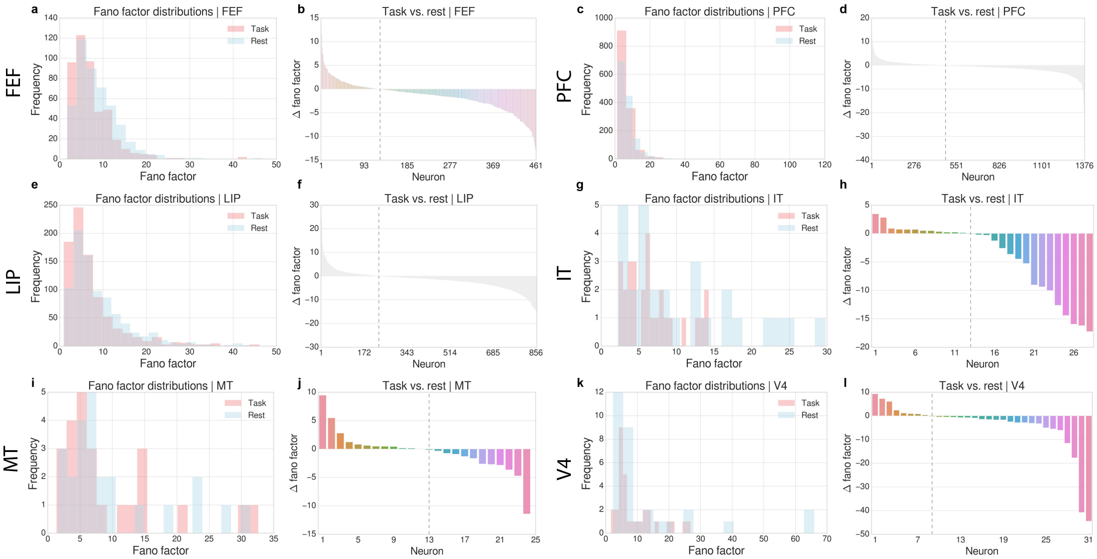


Figure C.12: Supplementary Figure 12. Rest (ITI) to task state (task cue) changes in fano factor analyzed for each neuron individually across the six cortical areas for the exploratory subject. (This is not a mean-field analysis.) a) The distribution of fano factor across all neurons in FEF (from all recording sessions) for the rest (ITI) and task state (cue) periods. b) For each individual neuron in FEF, we calculated the change in fano factor from the rest to task state period. c,d) Same as a, b, but for PFC. e,f) Same as a, b, but for LIP. g,h) Same as a, b, but for IT. i,j) Same as a, b, but for MT. k,l) Same as a, b, but for V4.

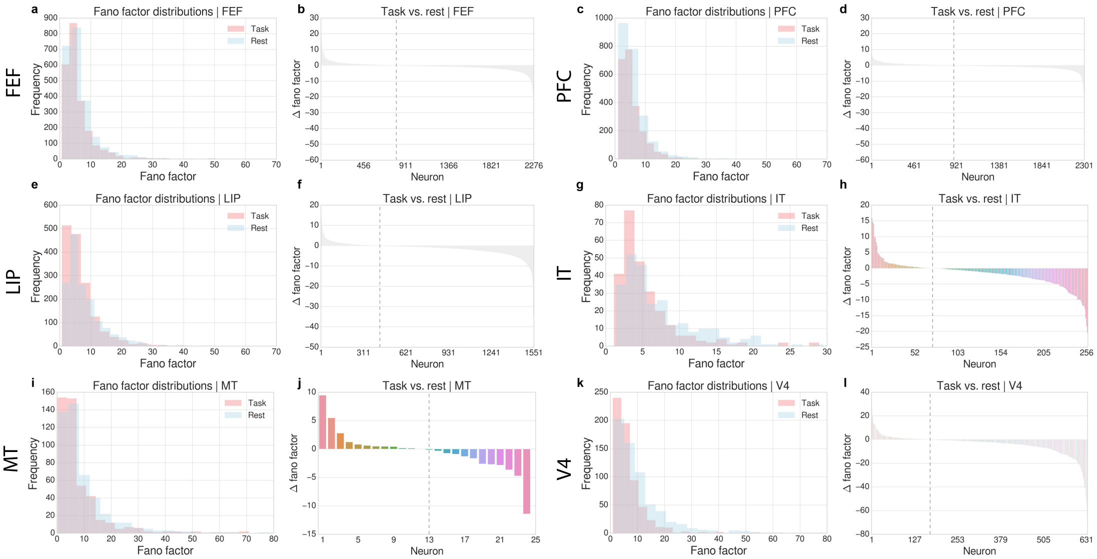


Figure C.13: Supplementary Figure 13. Rest (ITI) to task state (task cue) changes in fano factor analyzed for each neuron individually across the six cortical areas for the replication subject. (This is not a mean-field analysis.) a) The distribution of fano factor across all neurons in FEF (from all recording sessions) for the rest (ITI) and task state (cue) periods. b) For each individual neuron in FEF, we calculated the change in fano factor from the rest to task state period. c,d) Same as a, b, but for PFC. e,f) Same as a, b, but for LIP. g,h) Same as a, b, but for IT. i,j) Same as a, b, but for MT. k,l) Same as a, b, but for V4.

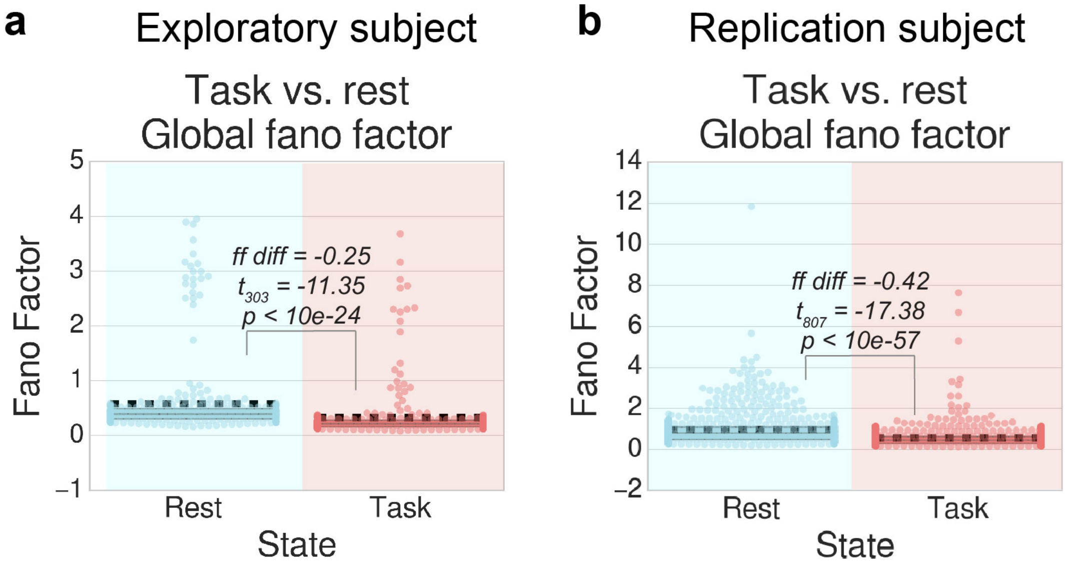


Figure C.14: Supplementary Figure 14. The average fano factor change from rest (ITI) to task (cue) periods for both the exploratory and replication NHP subjects. a) Exploratory subject. b) Replication subject.

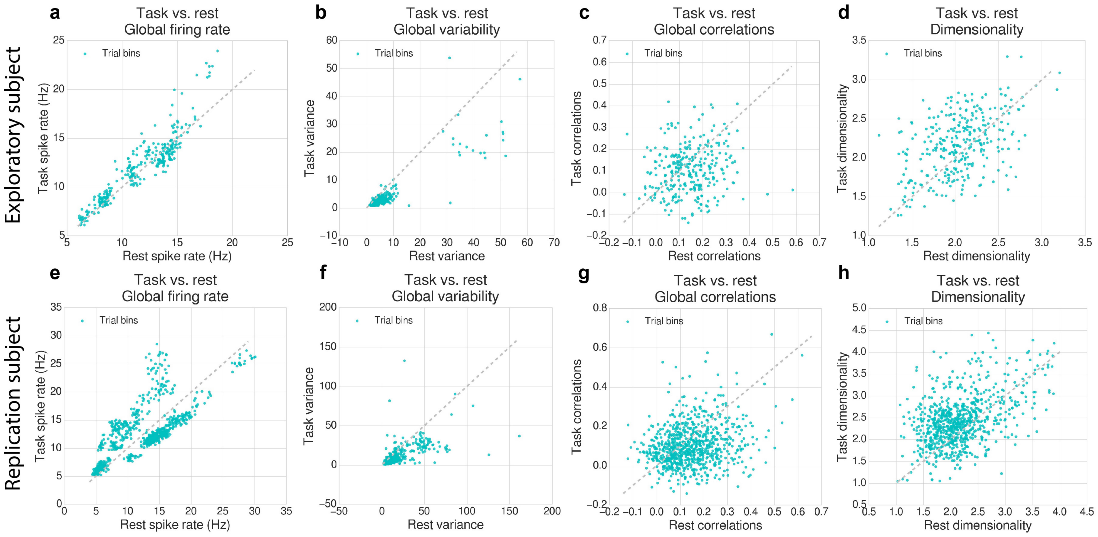


Figure C.15: Supplementary Figure 15. Scatter plot representations of averaged neural statistics (firing rate, variability, correlations, dimensionality) during rest and task state periods. All data are identical to those reported in Figures 5.2b-d, Figure 5.5b, and Supplementary Figures C.1b-d, but are visualized as a scatter plot. In each scatter plot, every point reflects the statistic (mean, variance, correlations, dimensionality) estimated across a bin of 25 contiguous trials. (Rest periods were defined as the ITI preceding the task cue onset.) Statistics were averaged across all recording sites, and included all recording sessions. a) The firing rate (averaged across six cortical areas) for task (y-axis) and rest (x-axis) states. b) The variance (averaged across six cortical areas) for task (y-axis) and rest (x-axis) states. c) Correlations (averaged across all pairwise correlations) for task (y-axis) and rest (x-axis) states. d) Dimensionality (i.e., participation ratio) of all six cortical areas during task (y-axis) and rest (x-axis) states. e-h) The same as a-d, but for the replication subject.