2021

Data-Based And Theory-Based Network Models Of Perturbations To Neural Dynamics

Jennifer Ann Stiso

University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Neuroscience and Neurobiology Commons

Recommended Citation

https://repository.upenn.edu/edissertations/4141

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/4141
For more information, please contact repository@pobox.upenn.edu.
Data-Based And Theory-Based Network Models Of Perturbations To Neural Dynamics

Abstract
Much of neuroscience is centered on uncovering simple principles that constrain the behavior of the brain. When considering the formation of neural architectures, similar structures can be recreated following the principles of minimizing wiring and maximizing topological complexity. However, a similar understanding of neural dynamics on top of these structural connections has not yet been achieved. One promising strategy for identifying underlying principles of neural dynamics is quantifying and modeling the response of neural systems to perturbation. Here, we use a spectrum of data- and theory-based network models to characterize the response of neural systems to different types of perturbations. We report how functional networks change in the context of pathological epileptic activity and brain-computer interface control. We also specifically test one possible principle: that activity is constrained to spread along connections in both the context of brain-computer interfaces and direct electrical stimulation. In the first study, we demonstrate across a wide variety of functional connectivity metrics and frequency bands that epileptic activity increases amplitude-based functional interactions, an observation that can now be incorporated into future theory-based models. In a second study, we determine that modeling activity that is constrained to spread along connections suggests why certain connections are important for brain-computer interface learning; specifically, these connections support sustained activity in attention regions. In our third study, we demonstrate that modeling activity changes from direct electrical stimulation using white matter connectivity explains more variance than models with rewired connections. This model generates testable predictions about which individuals, regions, and time points would lead to successful applications of direct electrical stimulation. Overall, this work demonstrates the potential uses of a range of data- and theory-based models for uncovering simple guiding principles that determine the behavior of a system. It also uses one specific principle - that activity is constrained to spread along connections - to understand the role of specific connections that may support learning, and provide a method to optimize individually tailored stimulation therapies for a specific outcome.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Neuroscience

First Advisor
Danielle S. Bassett

Second Advisor
Timothy H. Lucas

Keywords
Biological modeling, Brain dynamics, Network control theory, Network neuroscience, Neurophysiology

Subject Categories
Neuroscience and Neurobiology

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/4141
To my parents, Cindy and Nick, for instilling in me a love and appreciation for science.
We would like to thank Ann Sizemore for her help with Chapter 2, Box 2: Applied Algebraic Topology, as well as figure design. We would like to thank Ankit N. Kambhati for helpful discussions regarding the application of NMF to BCI learning and network control theory to stimulation (Chapters 4 and 5). We would also like to thank Jason Kim for his help developing the control theoretic model used in Chapter 5. We thank Youssef Ezzyat, Daniel Rizzuto, Michael Kahana and other members of the Kahana lab for guidance and providing classifier output for Chapter 5. We would also like to thank Michael Sperling and others at the Hospital at the University of Pennsylvania and Jefferson University Hospital for subject recruitment and stimulation monitoring (Chapters 3 and 5). We thank Blackrock Microsystems for providing neural recording and stimulation equipment. We acknowledge support from the National Institute of Mental Health National Research Service Award (F31MH120925), the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, the ISI Foundation, the Paul Allen Foundation, the Army Research Laboratory (W911NF-10-2-0022), the Army Research Office (Bassett-W911NF-14-1-0679, Grafton-W911NF-16-1-0474, DCIST- W911NF-17-2-0181), the Office of Naval Research, the National Institute of Mental Health (2-R01-DC-009209-11, R01 – MH112847, R01-MH107235, R21-M MH-106799), the National Institute of Child Health and Human Development (1R01HD086888-01), National Institute of Neurological Disorders and Stroke (R01 NS099348), the National Science Foundation (BCS-1441502, BCS-1430087, NSF PHY-1554488 and BCS-1631550), and French program "Investissements d’avenir" ANR-10-IAIHU-06; "ANRNIH CRCNS" ANR-15-NEUC-0006-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.
ABSTRACT

DATA-BASED AND THEORY-BASED NETWORK MODELS OF PERTURBATIONS TO NEURAL DYNAMICS

Jennifer Stiso
Danielle S. Bassett
Timothy H. Lucas

Much of neuroscience is centered on uncovering simple principles that constrain the behavior of the brain. When considering the formation of neural architectures, similar structures can be recreated following the principles of minimizing wiring and maximizing topological complexity. However, a similar understanding of neural dynamics on top of these structural connections has not yet been achieved. One promising strategy for identifying underlying principles of neural dynamics is quantifying and modeling the response of neural systems to perturbation. Here, we use a spectrum of data- and theory-based network models to characterize the response of neural systems to different types of perturbations. We report how functional networks change in the context of pathological epileptic activity and brain-computer interface control. We also specifically test one possible principle: that activity is constrained to spread along connections in both the context of brain-computer interfaces and direct electrical stimulation. In the first study, we demonstrate across a wide variety of functional connectivity metrics and frequency bands that epileptic activity increases amplitude-based functional interactions, an observation that can now be incorporated into future theory-based models. In a second study, we determine that modeling activity that is constrained to spread along connections suggests why certain connections are important for brain-computer interface learning; specifically, these connections support sustained activity in attention regions. In our third study, we demonstrate that modeling activity changes from direct electrical stimulation using white matter connectivity explains more variance than models with rewired connections. This model generates testable predictions about
which individuals, regions, and time points would lead to successful applications of direct
electrical stimulation. Overall, this work demonstrates the potential uses of a range of
data- and theory-based models for uncovering simple guiding principles that determine the
behavior of a system. It also uses one specific principle - that activity is constrained to
spread along connections - to understand the role of specific connections that may support
learning, and provide a method to optimize individually tailored stimulation therapies for
a specific outcome.
# TABLE OF CONTENTS

ACKNOWLEDGEMENT ........................................................... iv

ABSTRACT ................................................................. v

LIST OF TABLES ............................................................. viii

LIST OF ILLUSTRATIONS ................................................ xi

CHAPTER 1 : Introduction .................................................... 1

CHAPTER 2 : Spatial embedding imposes constraints on neuronal network architectures ........................................... 9

CHAPTER 3 : Changes in functional connectivity associated with interictal epileptiform discharges (IEDs) in intracranial EEG .................................................. 35

CHAPTER 4 : Learning in brain-computer interface control evidenced by joint decomposition of brain and behavior .................................................... 75

CHAPTER 5 : White matter network architecture guides direct electrical stimulation through optimal state transitions ......... 145

CHAPTER 6 : Conclusion ....................................................... 196

APPENDIX ................................................................. 202

BIBLIOGRAPHY ............................................................. 202
LIST OF TABLES

TABLE S3.1:  
Subject Information.  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 190
# LIST OF ILLUSTRATIONS

| FIGURE 1.1 : | Schematic of network models and perturbations. | 8 |
| FIGURE 2.1 : | Effect of wiring minimization and communication efficiency on network topology. | 13 |
| FIGURE 2.2 : | Spatial distribution of intrinsic neural activity. | 16 |
| FIGURE 2.3 : | Community structure obtained with spatially embedded and non-embedded null models. | 19 |
| FIGURE 2.4 : | Schematic of network measures. | 30 |
| FIGURE 2.5 : | Applied algebraic topology. | 32 |
| FIGURE 3.1 : | Schematic of methods. | 41 |
| FIGURE 3.2 : | Flow chart of methods. | 42 |
| FIGURE 3.3 : | Changes in Functional Connectivity Associated with IEDs. | 57 |
| FIGURE 3.4 : | Changes to specific connections. | 60 |
| FIGURE 3.5 : | Similarity across predictors, measures, and bands. | 63 |
| FIGURE 3.6 : | Individual differences. | 66 |
| FIGURE S3.1 : | Filter Responses. | 72 |
| FIGURE S3.2 : | Differences in groups of connections for the number of IEDs. | 73 |
| FIGURE S3.3 : | Strength distributions for a single subject. | 73 |
| FIGURE S3.4 : | Differences in effect sizes between contacts with IEDs present and absent. | 74 |
| FIGURE 4.1 : | Schematic of non-negative matrix factorization. | 82 |
| FIGURE 4.2 : | BCI task and performance. | 90 |
| FIGURE 4.3 : | Performance loading is associated with learning. | 98 |
FIGURE 4.4 : Spatial distribution of subgraph edges that are consistent across participants. .......................... 101
FIGURE 4.5 : Temporal expression of ranked subgraphs. ................. 103
FIGURE 4.6 : Separation of the ability to modulate attention is associated with learning. ............................... 106
FIGURE S4.1 : Effect of Regularization on Triangles. ...................... 126
FIGURE S4.2 : NMF Parameters. ........................................ 127
FIGURE S4.3 : NMF Parameter Consistency. .............................. 128
FIGURE S4.4 : Mean Subgraphs. ......................................... 129
FIGURE S4.5 : Consistency Across Thresholds. ............................ 130
FIGURE S4.6 : Consistency for Null Data. .................................. 131
FIGURE S4.7 : Connectivity to Electro-occulogram (EOG) Channels. 132
FIGURE S4.8 : Alternate target states with the same magnitude. ...... 133
FIGURE S4.9 : Montages. ................................................ 134
FIGURE S4.10 : Optimal Control Target States. .......................... 135
FIGURE S4.11 : Normality of Performance. ................................ 136
FIGURE S4.12 : Normality of Performance Loading for Empirical Data. 137
FIGURE S4.13 : Normality of Performance Loading for Null Data. .... 138
FIGURE S4.14 : Normality of Consistency Data. .......................... 139
FIGURE S4.15 : Normality of Peak Temporal Expression Data. ......... 140
FIGURE S4.16 : Normality of Energy Difference to Attention States. ... 141
FIGURE S4.17 : Normality of Energy Difference to Motor Imagery States. 142
FIGURE S4.18 : Normality of Energy Difference for Alternate Control Set and for Null Data. ............................. 143
FIGURE S4.19 : Normality of Energy Difference for Various State Controls. ........................................ 144
FIGURE 5.1 : Schematic of Methods. ................................... 148
FIGURE 5.2: Post-Stimulation Brain State Depends on White Matter Network Architecture. 154
FIGURE 5.3: Longer-Distance Trajectories Require More Stimulation Energy. 155
FIGURE 5.4: Topological and Spatial Constraints on the Energy Required for Stimulation-Based Control. 162
FIGURE 5.5: Role of Local Topology Around the Region Being Stimulated. 163
FIGURE 5.6: Network Topology and Brain State Predict Energy Requirements. 166
FIGURE S5.1: Effects of Different Parameter Choices in Open Loop Control. 188
FIGURE S5.2: Alternative Ways to Operationalize Brain States and Dynamic Properties. 189
FIGURE S5.3: Robustness and Selection of Parameters for Optimal Control. 191
FIGURE S5.4: Relationships Between Energy and Distance in Topological and Spatial Null Models. 192
FIGURE S5.5: Additional Spatially Embedded Null Models. 193
FIGURE S5.6: Additional Analyses Related to Regional Topology. 194
FIGURE S5.7: Electrode Coverage. 195
FIGURE 6.1: Cooperation between models. 199
Modern neuroscience began with the identification of the nervous systems as a system of discrete, interacting neurons (Ramón y Cajal (1996)). Descriptions of these systems have since grown to include their tremendous complexity; in humans, they contain 86 billion neurons of at least 60 possible types, with potentially thousands of connections per neuron, and 85 billion non-neural support cells (Herculano-Houzel (2009); Masland (2004)). Much of neuroscience seeks to understand the macroscale activity of this complex system parsimoniously, in terms of the fewest possible underlying principles. Progress towards this type of understanding has been achieved in other complex systems with many interacting parts, such as birds in a flock (King and Sumpter (2012)), or grains in a sand pile (Perrot and Rémila (2011)). Here, scientists can understand macroscale properties such as the shapes of flocks or the occurrences of avalanches in sand by modeling a few rules governing how different parts of the system interact. In neuroscience, one example of this level of understanding is found in literature investigating how neural architectures form within the confines of the skull (Stiso and Bassett (2018)). Specifically, many of the patterns of observed connections can be recreated by simulated systems that seek to (1) minimize the physical cost of wiring and (2) maximize topological complexity between functionally distinct regions (Kaiser and Hilgetag (2006)) (Fig 1A). The strength of this and similar models is that it provides an intuition regarding the core spatial and topological constraints guiding the system, and will accurately produce quantifiable features of the empirical system. However, a similar understanding of the dynamic functional changes that evolve on top of this structural backbone has not yet been achieved. One promising step towards this understanding is characterizing the response of activity in the brain to diverse kinds of perturbations, which are often difficult to create and observe in living systems. In this work, we explore three different types of perturbations to neural systems in humans using different types of network models.

Below, we will outline the major features of models and experiments included in this body of work. We will discuss the defining features of network models, and common applications
in neuroscience. We will then describe one important dimension on which these network models can vary: the extent to which they are data-based or theory-based. Next, we will discuss different methods of perturbation that can be used in neural systems. Lastly, we will summarize the models and perturbations used in each chapter.

Network Models

Network models are a large, diverse class of models that characterize a system as a group of discrete units and the interactions between them (Butts (2009)). In these systems, the discrete units are referred to as nodes, and the interactions between units are referred to as edges (Fig. 1B). In neural systems, the nodes and edges can have different biological instantiations depending on the level of detail at which researchers seek to model the system, and the types of interactions will provide insights into different research questions (Bassett and Sporns (2017)). For example, nodes could be single neurons, small groups of neurons in a circuit, or brain regions consisting of millions of neurons recorded from a single source. Edges are commonly quantified as either structural or functional. Structural edges represent estimates of physical connections between neurons or groups of neurons. These connections are typically visualized with non-invasive diffusion imaging or invasive histological tract tracing. Functional edges do not necessarily imply the presence of physical connections between nodes, but rather quantify the statistical similarity in the behavior of nodes. For example, functional connections might be estimated as the correlation between the activity timeseries of two different regions. Functional connections with a higher correlation would indicate more similar activity. Given the focus of this work on neural systems in humans where cellular imaging is difficult, we will adopt a definition of a node as a brain region. Edges are defined as either structural connections estimated from diffusion imaging or functional connections estimated by some similarity measure.
Models range from data-based to theory-based

Once a definition for nodes and edges has been chosen, researchers can select from a variety of models to address questions at different levels. These models vary along multiple dimensions, including their biological realism, spatial or temporal scale of analysis, and weight towards structure or function (Bassett et al. (2018b)). Here, we focus on one particular dimension that is independent from all those previously listed: the extent to which a model is data-based or theory-based. Models that are more data-based may provide more accurate descriptions of observed data, but this approach makes it difficult to infer the theoretical mechanisms that led to that data. Therefore, it is difficult to make claims about underlying principles from data-based models alone, but these models can capture much of the complexity and realism of the true system. By contrast, more theory-based models will typically specify a few simple rules that the system follows, and quantify the extent to which these rules recreate behaviors of interest (Gerstner et al. (2012)). These models better allow for claims about underlying mechanisms, but often times must sacrifice detail to posit simple, interpretable principles. While these two types of models are on opposite ends of a spectrum, they do not exist in isolation. Often times, work done using more data-based models is highly informative for future work on theory-based models, and both are useful for understanding the macroscale behavior of a system (Gao and Ganguli (2015)).

Common uses of network models in neuroscience span this continuum. For example after creating a data-based description of a structural or functional network, researchers often use graph theory to quantify summaries of the patterns of observed connections and report changes associated with learning, or disease, or differences compared to to null models (De Vico Fallani et al. (2014); Chiang and Haneef (2014); Bassett and Mattar (2017); Avena-Koenigsberger et al. (2018). Researchers also stipulate theory-based models of network development or activity that recreate distilled observations from the data-based models (Betzel et al. (2016a); Yan et al. (2017b)). One such theory-based model of activity with recent application to neuroscience is network control theory - which stipulates a simple, linear model of activity spread along connections in a complex system (Bryson (1996)).
This model posits a theory of how the behavior of a system changes - by spreading along connections. In this work, we assess models at three points on the data-theory continuum: one that is heavily data-based, one model that mixes data-based models of functional connections with theory-based model of activity, and a third that is more heavily theory-based. Each approach can add to the understanding of underlying principles of neural dynamics by characterizing the system response to different types of perturbations.

Different experimental methods of perturbation in neural systems

A useful strategy for identifying underlying principles guiding the macroscale behavior of the system is to quantify, characterize, and predict how that system will respond to perturbations, or small changes in the state of the system. Part of the usefulness of this approach comes from the causal inferences that can be made from well-controlled perturbations. Full observations of the effects of perturbations to every part of the system can be a major advancement towards a complete understanding of the underlying rules of the system (Barker et al. (2000)). A barrier to widespread use of perturbations in biological systems is that it can be difficult or impossible to design methodologically sound and ethical experiments that allow for such causal inferences (Marinescu et al. (2018)). For example, pathological brain activity observed in epilepsy has been thought of as a perturbation to ongoing brain dynamics; however these types of perturbations cannot ethically or practically be performed in a randomized manner. Whether or not perturbations can be interpreted as causal, there are several key insights that can be gained from observing consistencies in neural activity that are associated with or caused by some perturbations. The simplest of these insights involves quantifying the changes in the systems’ behavior pre- and post-stimulation (Khambhati et al. (2018a); Tomasino et al. (2014). Alternatively, one could use the effects of perturbations to test theories regarding guiding principles of the system (Stiso et al. (2018); Deng et al. (2019). Additionally, one can simply try to accurately predict the output of the system for a given perturbation (Becker et al. (2018)). In order to carry out these strategies using network models in neural systems, we need ways of systematically
perturbing the ongoing dynamics in the brain.

In neural systems, there are two broad categories of perturbations that can be studied: exogenous and endogenous. Exogenous perturbations are those that come from an external source, such as magnetic or electric stimulation (Folloni et al. (2019); Rose et al. (2016); Sironi (2011)). Within electrical stimulation, perturbations can further be broken down into indirect perturbations, such as those achieved with transcranial direct current stimulation (tDCS) or transcranial alternating current stimulation (tACS), and direct electrical stimulation (Sironi (2011)). The latter involves invasively implanting electrodes into the brain to directly apply current to neural tissue, and is currently an FDA-approved treatment for epilepsy, Parkinson’s disease, and essential tremor (Sironi (2011)). Endogenous perturbations, in contrast, are large changes in activity not resulting from an external electromagnetic pulse. Rather, these perturbations can arise from pathological or healthy ongoing brain activity (Stringer et al. (2016); Musall et al. (2019)). One example of pathological neural perturbations are interictal epileptiform discharges (IEDs) (Conrad et al. (2020)). Many individuals with epilepsy will display transient bursts of neural firing near their seizure onset zone that show a prominent spike in activity in a local region, but do not spread to become a seizure (Prince and Connors (1986). Non-pathological endogenous perturbations can come from volitional modulation of spatially localized neural activity above its baseline levels (Jeunet et al. (2016)). This type of modulation is commonly used to train brain-computer interfaces (BCIs) to respond to local changes in neural activity. In this work, we use both exogenous and endogenous perturbations to quantify, characterize, and predict the neural response to stimulation.

Specific models and methods of perturbation in this work

Here, we first review work that identifies potential unifying principles of the development of neural architectures, and then detail three projects that assess both data-based and theory-based models of neural dynamics, using models of both endogenous and exogenous perturbations. The first project quantifies how summaries of global functional connectivity
change in association with IEDs. Here, intracranial EEG (iEEG) data is used to estimate functional networks in 143 individuals with medically refractory epilepsy. The functional networks here are a data-based network model of the system, since they accurately quantify statistical similarity between brain regions but make no claims about the underlying principle that gave rise to those interactions. We then assess how characterizations of these networks differ during IEDs, a type of endogenous perturbation seen in epilepsy. We find that functional interactions tend to increase during IEDs, and that these increases are not driven only by pathological tissue. These findings can help provide evidence for validation of future theory-based models and aid understanding of the underlying principles of the epileptic brain’s propensity towards synchrony.

The second project asks how functional networks support learning to volitionally modulate one’s own brain activity in order to control a BCI. Here, sensor-level functional networks are created from magnetoencephalography data in 20 individuals who are learning to control a BCI. We then used non-negative matrix factorization to identify sparse subgraphs that make up each person’s functional network, and tested whether any of several graph properties were associated with BCI learning rate. We tested graph theoretic summary statistics from the original data-based model of activity, as well as statistics taken from a theory-based model of activity change in the brain. We found some theory-based model properties were associated with learning. Specifically, the theory-based model’s ability to activate sensors near regions important for sustaining attention was positively associated with learning, indicating an important role for attention processes in successful control.

Lastly, we report the results of a study investigating the efficacy of a theory-based model for predicting the outcomes of direct electrical stimulation. Here, we use iEEG recordings and diffusion weighted imaging in participants undergoing a direct electrical stimulation paradigm to reconstruct a structural network, and empirical functional activity before and after stimulation. We then define a theory-based model from network control theory to quantify activity spread along white matter tracts during stimulation to affect changes in
activity across the whole brain. Here, we show that this simple model of activity spread explains more variance in observed stimulation-induced changes in brain activity than similar null models using rewired structural connectivity networks. Additionally, the model gives us predictions on which features of brain activity, regional connectivity, and global connectivity would be most effective in facilitating stimulation targeted to improve memory. Ultimately, this model demonstrates some validity of the underlying principle of activity spread along white matter tracts, and also provides testable hypotheses regarding when and where to stimulate to reach certain patterns of brain activity in future experiments.

Overall, this body of work provides samples from three points on the data-theory spectrum of network models, and demonstrates how each model can add insight into how the brain functions as a system. We add new observations about properties of the epileptic brain and its propensity towards synchrony, posit a specific system whose supported dynamics are key for BCI learning, and suggest features for brain activity and connectivity that could facilitate desired stimulation outcomes. Collectively, these studies provide new insights into how dynamics are constrained by structure and pathology. Leveraging invasive neural stimulation and recordings, we provide unique evidence that more fully characterizes the macroscale behavior of the human brain in response to perturbation.
Figure 1.1: **Schematic of network models and perturbations.** (A) A schematic of principles underlying the formation of neural architectures. There exists a trade-off between wiring cost and topological complexity. Models that recreate the most features of neural systems exist at a balance between these two constraints. (B) An example of a structural (top) and functional (bottom) network model of a brain. While both models consist of nodes connected by edges, the edges in structural models correspond to physical connections, here visualized as a synapse, and the edges in functional models correspond to a measure of similarity between activity at nodes, here visualized as a correlation. (C) Where different models fall on the continuum between data-based (top-left) and theory-based (bottom-right). Adapted with permission from (Bassett et al. (2018b)). (D) Examples of endogenous (top) and exogenous (bottom) perturbations. Examples of exogenous perturbations include DBS (deep brain stimulation) and TMS (transcranial magnetic stimulation). Examples of endogenous perturbations include IEDs (interictal epileptiform discharges) and BCIs (brain-computer interfaces).
CHAPTER 2 : Spatial embedding imposes constraints on neuronal network architectures

Abstract

Recent progress towards understanding circuit function has capitalized on tools from network science to parsimoniously describe the spatiotemporal architecture of neural systems. Such tools often address the system’s topology divorced from its physical instantiation. Yet, for embedded systems like the brain, physical laws directly constrain processes of network growth, development, and function. Here, we review the rules imposed by the space and volume of the brain on the development of neuronal networks and show that these rules give rise to a specific set of complex topologies. These rules also affect the repertoire of neural dynamics that can emerge from the system, and thereby inform our understanding of network dysfunction in disease. We close by discussing new tools and models to delineate the effects of spatial embedding.
Network topology versus geometry in neural systems

In contemporary neuroscience, increasing volumes of data are being used to answer the question of how heterogeneous and distributed interactions between neural units might give rise to complex behaviors. Such interactions form characteristic patterns across multiple spatial scales, spanning from molecules and cells, to brain regions and lobes (Bassett and Sporns (2017)). An intuitive language in which to describe such interactions is network science, which elegantly represents interconnected systems as sets of nodes linked by edges. Nodes often represent proteins, neurons, subcortical nuclei, or large cortical areas, and edges often represent either (i) structural links in the form of chemical bonds, synapses, or white-matter tracts, or (ii) functional links in the form of statistical relations between nodal activity time series. Generally, the resultant network architecture can be fruitfully studied using tools from graph theory to obtain mechanistic insights pertinent to cognition (Hutchison et al. (2013)), above and beyond those provided by studies of regional activation (Tononi et al. (1994)) (Box 1).

In particular, several fundamental questions in neuroscience are quintessentially network questions concerning the physical relationships between functional units. How does the physical structure of a circuit affect its function? How does coordinated activity at small spatial scales give rise to emergent phenomena at large spatial scales? How might alterations in neurodevelopmental processes lead to circuit malfunction in psychiatric disorders? How might pathology spread through cortical and subcortical tissue giving rise to the well-known clinical presentations of neurological disease? These questions collectively highlight the fact that the brain — and its multiple networks of interacting units — is physically embedded into a fixed three-dimensional enclosure. Natural consequences of this embedding include diverse physical drivers of early connection formation and physical constraints on the resultant adult network architecture. An understanding of the system’s constitution and basal dynamics therefore require not only approaches to quantify and predict network topology, but also tools, theories, and methods to quantify and predict network geometry and its role in both enabling and constraining system function.
In this review, we provide evidence to support the notion that a consideration of the brain’s physical embedding will prove critical for a holistic understanding of neural circuit function. We focus our comments on the utility of informing this consideration with emerging computational tools developed for the characterization of spatial networks. Perhaps as a historical artifact of its origins in mathematics or its initial applications to abstract informational systems, network science often addresses the topology of systems in a way that is devoid of clear spatial characteristics (Ducruet and Beauguitté (2014)). The field has steadily developed tools and intuitions for spatially embedded network systems (Barthélemy (2011)). In the light of these recent technical developments, we begin by recounting observations from empirical studies addressing the question of how brain networks are embedded into physical space. Next, we discuss the relevance of this spatial embedding for an understanding of network function and dysfunction. We complement these empirical discussions with a more technical exposition on the relevant tools, methods, and statistical approaches to be considered when analyzing brain networks. Finally, we outline open questions regarding network architecture and circuit function, the answers to which will require a thorough appraisal of the role of physical space in brain network physiology.

Physical constraints on network topology and geometry

There are diverse processes that guide the formation of structural connections in neural systems (Bullmore and Sporns (2012); Chen et al. (2017)). Evidence from genetics suggests that neurons with similar functions as operationalized by similar gene expression tend to have more similar connection profiles than neurons with less similar functions (Bullmore and Sporns (2012); French and Pavlidis (2011); Rubinov et al. (2015)), with the greatest similarity appearing at highly interconnected, metabolically demanding hubs (Arnatkevičiūtė et al. (2017)). Of course, it is important to note that some spatial similarity of expression profiles is expected due to the influence of spatial gradients of growth factors during development (Bullmore and Sporns (2012)). However, evidence suggests that interareal connectivity profiles in rodent brains are even more correlated with gene co-expression than expected
Figure 2.1: **Effect of wiring minimization and communication efficiency on network topology.** Networks were generated by modulating the balance between a constraint on wiring (in the figure referred to as a *spatial cost*) and a constraint on information routing efficiency (in the figure referred to as a *temporal cost*). The parameter $\beta$, which ranges between 0 and 1, tunes this balance by weighting spatial cost against temporal cost. When $\beta = 0$ only the spatial cost is considered, while when $\beta = 1$ only the temporal cost is considered. (A) Examples of networks at different values of $\beta$ when only the spatial constraint exists (left), when only the temporal constraint exists (right), and when the two constraints are balanced (middle). Root nodes are shown in green and all other nodes are shown in yellow. (B) Spatial costs (blue) and temporal costs (red) vary as a function of $\beta$. This figure was adapted with permission from (Budd and Kisvárday (2012)).
simply based on such spatial relationships (French and Pavlidis (2011)). This heightened correlation could be partially explained by observations in mathematical modeling studies that neurons with similar inputs (and therefore potentially performing similar functions) tend to have more similar connection profiles than neurons with dissimilar inputs (Vertes et al. (2012)).

Yet, while genetic coding and functional utility each play important roles, a key challenge lies in summarizing the various constraints on connection formation in a simple and intuitive theory that can guide future predictions. One particularly acclaimed theory first outlined by Ramon y Cajal in 1899 (Ramón y Cajal (1996)) is that physical constraints on space, time, and material over development underlie connection formation. Metabolism related to neural architecture and function is costly, utilizing 20% of the body’s energy, despite comprising only 2% of its volume (Laughlin and Sejnowski (2002)). Even the development of axons alone extorts a large material cost (Bullmore and Sporns (2012)). The existence of these pervasive costs motivated early work to postulate that wiring minimization is a fundamental driver of connection formation. Consistent with this hypothesis, axons in the brain seem to occupy a nearly optimal volume to minimize metabolic costs (Chklovskii et al. (2002)). Additionally, the neural architecture of multiple species (Cherniak (1994); Young (1992) and at multiple scales (Song et al. (2005)) across different methods of data collection (Cherniak et al. (2010); Song et al. (2014); Rubinov (2016)) are predominantly comprised of wires extending over markedly short distances (Budd and Kisvárday (2012); Bullmore and Sporns (2012)).

However, mounting evidence suggests that pressures for wiring minimization may compete against pressures for additional topological complexity (Young (1992)) that could facilitate efficient communication (Bassett et al. (2010a); Kaiser and Hilgetag (2006); Zalesky et al. (2012)). Early evidence supporting the role of efficient communication came from the observation that one can fix the network architecture of inter-areal projections in the macaque cortex (and later human (Bassett et al. (2010a)), mouse(Rubinov et al. (2015)),

14
and dendritic arbors (Budd and Kisvárday (2012); Bullmore and Sporns (2012)) and then rearrange the location of areas in space to obtain a configuration with significantly lower wiring cost (Kaiser and Hilgetag (2006)). Interestingly, the connections whose length is decreased most also tend to be those that shorten the characteristic path length – one of many ways to quantify how efficiently a network can communicate (Kaiser and Hilgetag (2006)). Notably, computational models that instantiate both constraints on wiring and efficient communication produce topologies more similar to the true topologies than models that instantiate a constraint on wiring minimization alone (Chen et al. (2013); Vertes et al. (2012). Moreover, models that allow for changes in this tradeoff over developmental time periods better fit observed connectome growth patterns than prior models, positing a mechanism of early connection to nearby sources, coupled with later expansion of older densely connected clusters to create topological diversity (Nicosia et al. (2013)). It is worth noting that other properties have been proposed as drivers in addition to communication efficiency, such as fine scale chemical mechanisms of chemotaxis (Tessier-Lavigne and Goodman (1996) and large-scale mechanisms driving functional diversity via long distance connections (Betzel and Bassett (2018)). Alternatives to the hypothesis that communication efficiency is a key driver include the preservation of hubs specifically (Rubinov (2016)).

It is precisely this balance between wiring minimization and communication efficiency that is thought to produce the complex network topologies observed in neural systems, along with markedly precise spatial embedding (Henderson and Robinson (2014); Kaiser (2017)). A simple illustration of this precise embedding lies in the allometric scaling of white versus grey matter across species (Jerison (1975); Changizi (2001)). To better understand how this scaling relates to the topology of a single organism, it is useful to consider methods that can simultaneously (rather than independently) assess topology and geometry. One such method that has proven particularly useful in the study of neural systems from mice to humans is Rentian scaling, which assesses the efficiency of a network’s spatial embedding (Bassett et al. (2010a); Sperry et al. (2017); Pineda-Pardo et al. (2015)). Originally developed in the context of computer circuits, Rentian scaling describes a power-law scaling
Figure 2.2: **Spatial distribution of intrinsic neural activity.** Principal gradients of functional connectivity calculated in the structural connections of both humans and macaques. The first two principal gradients explained approximately 40% of the observed variance. (A) (Left) A scatter plot of the first two principal gradients, with transmodal regions shown in red, visual regions shown in blue, and sensorimotor regions shown in green. (Right) The same colors are used to show the distribution of points visualized on a cortical surface. The pattern suggests the existence of a macroscale gradient of connectivity that reflects the systematic integration of information across different sensory modalities. (B, *Left*) The minimum geodesic distance (mm) between each point on the cortical surface and the positive peaks of the first principal gradient. The peaks are shown as white circles. (B, *Right*) A scatter plot depicting the relationship between distance and location on the trans-modal gradient. Put differently, transmodal regions with high values in the principal gradient are maximally distant from unimodal regions with low values in the principal gradient. This figure was adapted with permission from (Betzel et al. (2016c)).

The relationship between the number of nodes in a volume and the number of connections crossing the boundary of the volume (Bassett et al. (2010a); Bullmore and Sporns (2012)). The existence of such a power law relationship with an exponent known as Rent’s exponent is consistent with an efficient spatial embedding of a complex topology (Alcalde Cuesta et al. (2017)). Additionally, the Rent’s exponent of connections in the human brain is proportional to the allometric scaling of grey and white matter volume across species, creating a putative link between the efficient embedding of a single system and the scaling of connectomes across evolution.

Reflections of physical constraints in local to global network topology

Across species, the brain consistently exhibits a set of topological features at local (single regions), meso- (neural circuits), and global (entire connectome) scales that can be simply explained by a few spatial wiring rules (Avena-Koenigsberger et al. (2014); Sporns et al. (2004); Henderson and Robinson (2014)). At the local scale, multiple modalities have been
used to demonstrate that a key conserved topological feature is the existence of hubs, or nodes of unexpectedly high degree (van den Heuvel et al. (2012); Seidlitz et al. (2018)). Such hubs emerge naturally in computational models in which the location of nodes are fixed in space, and edges between nodes are rewired to minimize average wiring length and to maximize topological efficiency by minimizing the average shortest path length (Box 1), though the number and degree of hubs varies systematically with the relative importance of the two constraints (Budd and Kisvárday (2012); Chen et al. (2013)) (Fig.1). Importantly, when both constraints are balanced, networks contain several hubs of varying degrees, consistent with the topology observed in brain networks (Chen et al. (2013)). In brain networks, hubs tend to be linked by connections that are longer than expected (Roberts et al. (2016); van den Heuvel and Sporns (2013)), although their exact physical placement enables low wiring cost given the presence of hubs (Gollo et al. (2018)). It is notable that such constraints can be implemented within the natural processes of development; for example, in adult C. elegans, hub neurons have been tracked back to the earliest born neurons in the embryo, which accumulate a large number of connections along the normative growth trajectory (Kaiser (2017); Varier and Kaiser (2011)).

At the mesoscale, a key conserved topological feature is modularity, or the existence of internally dense and externally sparse communities of nodes (Avena-Koenigsberger et al. (2014); Hilgetag and Goulas (2016); Henderson and Robinson (2014)). The strength of modularity in a network is commonly quantified using a modularity quality index (Box 1). In computational models, this index obtained under pressures of wiring minimization and communication efficiency (quantified with path length) was more similar to that empirically measured in the connectomes of the macaque and C. elegans than to that obtained under either constraint separately (Chen et al. (2013, 2017)). Again it is notable that such constraints can be implemented within the natural processes of development; for example, in Drosophila, communities form when many neurons are born in a similar temporal window, and therefore typically share a common progenitor type, and therefore a similar spatial location and genetic profile (Kaiser (2017); Chiang et al. (2011)). Genetically similar neurons
being born in close proximity are likely to connect to one another, forming densely connected functional groups. Spaces between modules can form cavities or cycles, or intuitively holes in the network, that can be identified with emerging tools from applied algebraic topology (Box 2) (Sizemore et al. (2017)). The locations, prevalence, and weight structure of these cycles differs markedly between geometric and random networks (Kahle (2018, 2009)), with patterns of functional connectivity among neurons exhibiting characteristics similar to those observed in spatially constrained geometric networks (Giusti et al. (2015)). It will be interesting in future to gain a deeper understanding of the relations between cycles and modules, and their emergence through the spatially constrained processes of development.

At the global scale, a key conserved topological feature is small-worldness, or the confluence of unexpectedly high clustering and short path length (Box 1) (Watts and Strogatz (1998)). Such an architecture is thought to be particularly conducive to a balance between local information processing within the clusters, and global information transmission across the topologically long distance connections (Shih et al. (2015)). Similar to the existence of hubs, modules, and cavities, small-world architecture in a network can naturally arise from spatial constraints on wiring (Kaiser and Hilgetag (2004)). Intuitively, clusters tend to form in spatially nearby regions in order to minimize wiring cost, while long distance connections facilitating efficient communication tend to form only occasionally due to their elevated wiring cost (Bassett and Bullmore (2017)). In concert with these empirical observations, computational models that account for wiring economy produce networks with small-world architecture reminiscent of that observed in real neural systems (Avena-Koenigsberger et al. (2014)). Collectively, these studies demonstrate the influence of parsimonious wiring rules on complex network topology. Future work could be directed to better understand the aspects of connectome topology that remain unexplained and thus may arise from more subtle rules (Chen et al. (2017)).
Figure 2.3: Community structure obtained with spatially embedded and non-embedded null models. 

(A) A schematic of a spatially-informed null model. The model expects fewer long distance connections than short distance connections. 

(B) A schematic of the anticipated difference between the spatial null model and the Newman-Girvan (NG) null model; spatial communities will have longer distance connections and not capture clustering of spatially nearby regions. 

(C) Differences in the association matrices between the two models. Positive (negative) numbers indicate when two nodes were more likely to be co-assigned to the same module under the spatial (NG) model. 

(D) The difference in the participation coefficient between the spatial and NG models. The participation coefficient quantifies how diverse a node’s connections are across modules. 

(E) The difference in spatial spread of modules in both models; the spatially embedded model tends to produce modules that cover larger distances. This figure was adapted with permission from (Margulies et al. (2016)).
Relevance of network geometry for dynamics and cognition

Pressures for wiring minimization and communication efficiency can exist alongside developmental processes that produce non-isotropically structured organs that result in patterning across multiple overlapping signaling gradients (Tessier-Lavigne and Goodman (1996)). It is intuitively possible that such processes could also explain the observed differences in the network topologies of different sectors of the brain (Scholtens et al. (2014); Batalle et al. (2017)), which can impinge on the functions that those sectors are optimized to perform (Box 3). Indeed, prior work has noted the co-existence of complex structural topologies and spatial gradients of specific function (Jbabdi et al. (2013)), although it has been difficult to achieve a mechanistic understanding of exactly how the two relate to one another. One particularly promising recent line of investigation attempting to link the two mechanisms has proposed the existence of a set of primary spatial gradients that explains variance in large-scale connectivity (Huntenburg et al. (2018); Margulies et al. (2016)) (Fig. 2A). In both humans and macaques, the primary axis of variance is bounded on one end by the transmodal default mode system, and on the other end by the unimodal sensory systems (Margulies et al. (2016)) (Fig. 2B). Notably, this gradient is tightly linked to the geometry of the network, with the regions located at one end having maximal spatial distances from the regions located at the other end (Margulies et al. (2016)). Additionally, the regions located at the peaks of the transmodal gradient have substantial overlap with structural hubs (whose putative role in the wiring economy has been discussed) in human connectomes (van den Heuvel and Sporns (2013); Petersen and Sporns (2015); Rubinov (2016)). Put simply, such evidence supports the notion that the cortex is fundamentally organized along a dimension of function from concrete to abstract, and that dimension manifests clearly in the network’s spatial embedding.

The specific topology and spatial geometry of brain networks has important implications for the patterns of neural dynamics that one would expect to observe. Consider, for example, the patterns of intrinsic activity noted consistently across species, individuals, and imaging modalities in the default mode system (Raichle (2015)). The consistent architecture of
correlations between regional time series in this system suggests a role for time-invariant structural features in organizing these dynamics. Consistent with this suggestion, functional connectivity in the default mode is more similar to its structural connectivity than other systems (Horn et al. (2014)). Recent work addressing the mechanisms of the stable intrinsic activity patterns in the default mode has posited the existence of so-called lag threads, or spatial progressions of whole-brain activity patterns at non-zero time lags (Mitra et al. (2015)). Notably, regions of the default mode participate in consistent lag thread motifs, where changes in the activity in one region reliably lead to changes in the activity of another region (Mitra et al. (2015); Raichle (2015)). It has been postulated that these lag threads, both within and outside of the default mode, arise from infra-slow oscillations in membrane potential that travel between cortical layers (Mitra et al. (2018)), although further work parsing the relative role of passive propagation along structural pathways versus active neuromodulation in these patterns is needed.

In addition to characteristic dynamics of activity across the default mode, the brain also shows reliable wave-like cortical dynamics in both task and rest that are important for neural computation (Richardson et al. (2005); Rubino et al. (2006); Schoch et al. (2018)). Different aspects of these dynamics can be replicated with biophysical models of neuronal activity that account for the delay of activity propagation across axons, indicating that connection topology and distance of connections might be important for their characteristic spread (Jirsa and Haken (1996)). Using whole brain human connectomes, these models can recreate metastable patterns of waves, sources, and sinks, where such patterns tend to emanate from hubs in the network more than non-hubs (Roberts et al. (2018)). Together, these results support the notion that features of the network topology created by the spatial embedding of the brain influence the reliable patterns of dynamics observed in the cortex.

Relevance of network geometry for disease
The spatial architecture of brain networks not only impacts our understanding of dynamics and cognition, but also our understanding of neurological disease and psychiatric disor-
ders. Mounting evidence suggests that many diseases and disorders of mental health can be thought of fruitfully as network disorders, where the anatomy and physiology of cross-regional communication can go awry (Braun et al. (2018a)). Intuitively, spatial anisotropies of developmental processes, or spatial specificity of pathology could also explain alterations in the spatial characteristics of brain networks (Bassett et al. (2018a)). Though there are many different neurological diseases with pathologies related to the spatial embedding of the brain (for a review, see (Fornito et al. (2015))), we will limit our discussion in this correspondence to epilepsy, a particularly common neurological disease, and in schizophrenia, a particularly devastating psychiatric disorder.

Despite a diverse pathophysiology but a renitent unifying biological manifestation, epilepsy is characterized by altered network dynamics in the form of seizures that display spatially consistent patterns. For example, an ictal period often begins with a marked spatial decorrelation followed by a period in which abnormally synchronized activity propagates in consistent spatial patterns (Jirsa et al. (2014); Wendling et al. (2003)). In addition to broad patterns of spatial decorrelation, individual siezures also show stereotyped patterns of both spiral waves and travelling waves of activity (González-Ramírez et al. (2015); Richardson et al. (2005); Martinet et al. (2017)). In silico studies have demonstrated that a simple adaptive model of synaptically coupled and spatially embedded excitatory neurons can reproduce many basic features of these waveforms, including their speed and the size of the wavefront (González-Ramírez et al. (2015). We have noted that travelling waves are not unique to epilepsy, however, marked differences in wave propagation in healthy and epileptic cortical tissue suggests that the precise spatial progression is important, potentially supported by distinct underlying microstructures (Benucci et al. (2007)). Finally, even interictal dynamics are altered in epilepsy, as manifest by marked decreases in average functional connectivity across the brain combined with local increases in functional connectivity and efficiency in default mode areas (Bonilha et al. (2012); DeSalvo et al. (2014)). These connectivity patterns have some utility in predicting seizure spread, but the guiding principles leading to these changes and how they relate to fine scale patterns of activity
remains unclear (Jirsa et al. (2017)).

While its pathophysiology is quite distinct from that implicated in epilepsy, schizophrenia is also a condition marked by severe network disturbances that have broad ramifications for cognitive function (Bassett et al. (2008); Zalesky et al. (2012)). Some of these network alterations appear to selectively affect connections of certain physical lengths, reflecting an alteration in the network’s spatial embedding (Alexander-Bloch et al. (2013)). Specifically, evidence suggests a reduced hierarchical structure and increased connection distance in the anatomical connectivity of multimodal cortex in patients with schizophrenia compared to healthy controls, indicative of less efficient spatial wiring (Bassett et al. (2008)). Moreover, in functional brain networks, patients display longer high-weight connections, decreased clustering, and increased topological efficiency in comparison to healthy controls (Alexander-Bloch et al. (2013)). The lack of strong, short-distance functional connections is in line with evidence from animal studies suggesting an over-pruning of synapses in childhood onset schizophrenia (Alexander-Bloch et al. (2013)). Additionally, the location of hubs (with high metabolic cost) coincides tightly with grey matter loss in schizophrenia (Gollo et al. (2018)). Here, the intuitions gained from a consideration of the network’s spatial embedding offer important directions for future work in linking non-invasive imaging phenotypes with invasive biomarkers of neural dysfunction in disease.

Statistics, Null Models, and Generative Models

In the previous sections, we outlined developmental rules for efficient wiring and we discussed the reflections of these rules in spatial patterns of healthy and diseased brain dynamics. Collectively, the studies that we have reviewed motivate the broader use and further development of sophisticated and easily-implementable tools for the analysis of a network’s spatial embedding (Kaiser (2011)). Here we outline the current state of the field in developing effective network statistics, network null models, and generative network models that account for spatial embedding.
Network Statistics. A simple way to examine network architecture in the context of spatial embedding is to incorporate the Euclidean distance of connections into local, meso-scale, and global statistics (Alexander-Bloch et al. (2013); Duarte-Carvajalino et al. (2012)). Arguably the simplest local statistics that remain spatially sensitive are moments of the distribution of edge lengths in the network. One can also compute graph metrics that have been extended to consider space, such as the physical network efficiency and the physical edge betweenness (Buhl et al. (2004)). The physical network efficiency uses the physically (rather than topologically) shortest path, then takes the inverse of the harmonic mean of this length, while the physical edge betweenness provides the fraction of shortest physical paths between all node pairs that traverse a given edge (Papadopoulos et al. (2018)). One could also define a physical clustering coefficient in a similar manner. Finally, one can assess the system for Rentian scaling as described earlier, providing information on how efficiently the complex network topology has been embedded into the physical space (Bassett et al. (2010a); Sperry et al. (2017); Pineda-Pardo et al. (2015)). In the context of neural systems, these spatially informed graph statistics can be used to account for the physical nature of information processing, propagation, and transmission.

Complementing local and global graph statistics is an assessment of a network’s community structure, a mesoscale property frequently assessed by considering the existence and strength of network modules (Garcia et al. (2018)). From that community structure, one can determine the spatial embedding of communities, for example by assessing their laterality in bilaterally symmetric systems such as the brain (Doron et al. (2012); He et al. (2018)). One of the most common ways to assess community structure is to maximize a modularity quality function, which identifies assortative modules with dense within-module connectivity and sparse between-module connectivity (Newman and Girvan (2003)) (see Betzel et al. (2018) for methods to identify non-assortative communities). Statistically, this algorithm compares the strength of observed connections between two nodes in a community to that expected under a given a null model. The most commonly used null model in this context is the Newman-Girvan or configuration model, which preserves the strength distribution
of the network (Newman and Girvan (2003)). However, this null model operationalizes a purely topological constraint – the strength distribution – and does not acknowledge any spatial constraints that may exist in the system. For this reason, many investigators across scientific domains have begun developing alternative null models that account for physical constraints (Betzel et al. (2016c); Expert et al. (2010); Sarzynska et al. (2015)) on their system of interest.

In the context of brain networks, it is worth considering three distinct null models for modularity maximization that incorporate information about the physical space of the network’s embedding. First, one can directly incorporate the wiring minimization constraint observed in brain networks by defining a null model with a probability of connection between two nodes that decays exponentially as a function of distance (Betzel et al. (2016c)) Fig.3A-B. Using this model, one can detect different and more spatially distributed modules than those obtained when one uses the configuration model (Betzel et al. (2016c)) (Fig.3C-E). Second, one can employ gravity models (Expert et al. (2010)), which account for the number of connections expected given a certain distance (typically a power law or inverse of distance), weighted by the relative importance of each location (typically a quantification of the population or size of a given location) (Expert et al. (2010); Sarzynska et al. (2015)). Third, one can employ radiation models designed to capture flow of information between regions, by weighting distance functions by the flux or flow of each location (Sarzynska et al. (2015)). Of course, there exists no single correct null model for community detection that will suit every question in neuroscience. However, we propose that many studies could test tighter, more targeted hypotheses about community structure in brain networks by using a null model that accounts for the brain’s spatial nature.

Network Null Models. When considering a network representation of a neural system, one often computes a statistical quantity of interest and then compares that quantity to that expected in a random network null model. If the observed quantity is significantly greater
than or less than that expected, one concludes that the network under study shows meaningful architecture of potential relevance to the biology. Perhaps the most common random network null model is that which randomly permutes the locations of edges in the network while preserving the number of nodes, number of edges, and edge weight distribution. However, one may also be interested to determine whether observed statistics are different from what one would expect simply from the spatial embedding or wiring rules of the network (Roberts et al. (2016); Samu et al. (2014); Wiedermann et al. (2016). To address these questions, one can rewire the observed network by conditionally swapping two links if the swap preserves the mean wiring length of the network (Samu et al. (2014)). By pairing this model with a reduced null model in which connections are only swapped if they reduce connection length, one can assess the role of long distance connections in the network, which will be preserved in the spatial null but not preserved in the reduced null (Samu et al. (2014)). In addition to preserving the mean wiring length, one might also wish to preserve the full edge length distribution by, for example, (1) fitting a function to the relationship between the mean and variance of edge weights and their distances, (2) removing the effect of that relationship from the data, (3) randomly rewiring the network, and (4) adding the effect back into the rewired network (Roberts et al. (2016)).

To complement insights obtained from edge swapping algorithms, one can also construct null model networks by stipulating a wiring rule a priori while fixing the locations of nodes within the embedded system. In this vein, studies have fruitfully used null models based on minimum spanning tree and greedy triangulation methods (Cui et al. (2018); Smit et al. (2016). A minimum spanning tree is a graph that connects all of the nodes in a network such that the sum of the total edge weights is minimal. To extend this notion to spatial networks, one can preserve the true geographic locations of all nodes in the empirical network and compute the minimum spanning tree on the matrix of Euclidean distances between all node pairs (Papadopoulos et al. (2018)). Representing the opposite extreme is the greedy triangulation model, which is particularly relevant for the study of empirical networks that are planar (lying along a surface) as opposed to non-planar (lying within a volume). In the
context of neural systems, planar or planar-like networks are observed in vasculature, and in thinned models of cortex that either consider a single lamina (Schmid et al. (2017)). To construct a greedy triangulation null model, one can preserve the true geographic locations of all nodes in the empirical network and iteratively connect pairs of nodes in ascending order of their distance while ensuring that no edges cross. After constructing such minimally and maximally wired null models, one can calculate relative measures of wiring length, physical efficiency, physical betweenness centrality, and community structure by normalizing the empirical values by those expected in the two extremes (Papadopoulos et al. (2018)).

**Generative Network Models.** Generative network models can be used to test hypotheses about the rules guiding network growth, development, and evolution (Betzel and Bassett (2017)). Often, an ensemble of generative models are constructed, and summary graph statistics from the empirical network are compared to the statistics of each of the generative models with the goal of inferring which wiring rule was most likely to have produced the observed architecture (Klimm et al. (2014); Betzel et al. (2016a); Vertes et al. (2012)). Evidence from such studies suggests that spatially embedded models tend to more accurately reproduce network measures of large-scale neural systems than models that do not account for space (Klimm et al. (2014)). One particularly influential study considered 13 generative models that all incorporated a wiring probability that increased with distance (Betzel et al. (2016a)). Consistent with other work, the authors found that the model that only included the wiring minimization constraint was unable to recreate long distance connections of individual connectomes in humans (Betzel et al. (2016a); Chen et al. (2017); Vertes et al. (2012)). Successive generative models were then added that attempted to recreate certain aspects of topology in addition to these geometric constraints (Betzel et al. (2016a)). The models that performed the best were those that preserved homophilic attraction such that connections preferentially formed between nodes that had similar connection profiles (Betzel et al. (2016a)). Generative models can also be used to determine the implicit geometric structure that would give rise to graphs with specific topological properties(Allard et al.
(2017)), and directly assess how Euclidean space the network is embedded in relates to this geometry. Continued advancement of generative network models, and inclusion of additional biological features such as bilateral symmetry, serves as an exciting approach to test mechanistic predictions about how network topology forms in spatially embedded neural systems.

Concluding Remarks and Future Directions

The spatial embedding of the brain is an important driver of its connectivity, which in turn directly constrains neural function and by extension behavior. Emerging tools from network science can be used to assess this spatial architecture, thereby allowing investigators to test more specific hypotheses about brain network structure and dynamics. While we envision that the use of these tools will significantly expand our understanding, it is also important to acknowledge their limitations. In particular, the majority of currently available network tools make the simplifying assumption that all of the relations of interests are strictly dyadic in nature, and exist between inherently separable components (Butts (2009)). In truth, however, features that arise from spatial embedding can also manifest as continuous or overlapping maps and gradients (Jbabdi et al. (2013)), motivating the use of tools from applied algebraic topology that can account for non-dyadic interactions (Box 2). As the field moves forward, we envision existing and yet-to-be-developed tools for characterizing the spatial embedding of brain networks will prove critical for our understanding of network processes underlying cognition, and alterations to those processes accompanying disease.
Box 1: Simple Network Statistics

In a network representation of the brain, units ranging from neurons or neuronal ensembles to nuclei and areas are represented as network nodes and unit-to-unit interactions ranging from physical connections to statistical similarities in activity time series are represented as network edges. These connections can have independent topological, and spatial distances (Fig. 1A). The architecture of the network can be quantitatively characterized using statistics from graph theory (Rubinov and Sporns (2010)). Here, we mathematically define some of the topological statistics mentioned elsewhere in this paper (Fig. 1B).

- **Degree and Strength.** The degree of a node is the number of connections it has. In a binary graph encoded in the adjacency matrix $A$, where two regions $i$ and $j$ are connected if $A_{ij} = 1$, and not connected if $A_{ij} = 0$, then the degree $k_i$ is defined as $k_i = \sum_{i,j \in N} A_{ij}$, where $N$ is the set of all nodes. In a weighted graph, where $A_{ij}$ is the strength of the connection between nodes $i$ and $j$, then the strength $s_i$ is defined as $s_i = \sum_{i,j \in N} A_{ij}$.

- **Path Length and Network Efficiency.** The term path length frequently refers to the average length of the shortest path in a network. The shortest path between any two nodes is given by the path requiring the fewest hops. The network efficiency is given by the inverse of the harmonic mean of the shortest path length. To be precise, we can write the path length of node $i$ as $L_i = \frac{1}{n} \sum_{i \in N} \frac{1}{\sum_{j \neq i} d_{i,j}}$, where $d_{i,j}$ is the shortest path length between two nodes and $n$ is the number of nodes.

- **Clustering Coefficient.** The clustering coefficient can be used to quantify the fraction of a node’s neighbors that are also neighbors with each other. Specifically, the clustering coefficient of node $i$ is given by $C_i = \frac{1}{n} \sum_{i \in N} \frac{2t_i}{k_i(k_i-1)}$, where $t_i$ is the number of triangles around node $i$ (Watts and Strogatz (1998)). The clustering coefficient of the network is the average clustering coefficient of all of its nodes.

- **Modularity.** While several modularity quality functions exist, the most common is
Figure 2.4: **Schematic of network measures.** (A) An illustration of network space (topology) and physical space (geometry). The network is embedded into a physical space, indicated by the $x$- and $y$-axes. The topological and physical distances between the nodes are not necessarily related. (B) The network representation enables the calculation of local, mesoscale, and global features to describe the pattern of connections in topological space (as shown here) as well as the pattern of connections in physical space (as we describe in the main text). This figure was adapted with permission from (Bassett and Sporns (2017)) and from (Garcia et al. (2018)).

$$Q = \sum_{ij}[A_{ij} - \gamma P_{ij}]\delta(c_ic_j),$$

where $Q$ is the modularity quality index, $P_{ij}$ is the expected number of connections between node $i$ and node $j$ under a specified null model, $\delta()$ is the Kroenecker delta, and $c_i$ indicates the community assignment of node $i$. The tuning parameter $\gamma$ ranges from $(0, \infty)$ and can be used to tune the average community size.

**Box 2: Applied Algebraic Topology**

While graph theory is a powerful and accessible framework for analyzing complex networks, complementary information can be gained by using different mathematical formalisms. Here, we describe an alternative approach to studying structure in networks that relies on tools developed in the field of applied algebraic topology, specifically persistent homology.
Persistent homology can be used to study intrinsically mesoscale structures called cycles and cliques (Sizemore et al. (2018)). Cycles are looped patterns of cliques which may enclose a cavity, or topological void, within the network. Cycles are all-to-all connected subsets of nodes in a network. The presence of many, large cliques indicates many highly connected units are present in the network (Reimann et al. (2017)). Cliques are all-to-all connected subsets of nodes in a network. The presence of many, large cliques indicates many highly connected units are present in the network (Reimann et al. (2017)). Cliques are all-to-all connected subsets of nodes in a network. The presence of many, large cliques indicates many highly connected units are present in the network (Reimann et al. (2017)). Cliques are all-to-all connected subsets of nodes in a network. The presence of many, large cliques indicates many highly connected units are present in the network (Reimann et al. (2017)). Cliques and cavities by definition reside within a binary graph, however one can expand a weighted network into a sequence of binary graphs via iterative thresholding (Giusti et al. (2015); Petri et al. (2013)). Then using persistent homology one can track the birth, persistence, and death of cavities along this sequence which gives a wholistic insight into the global network (Fig. I Box 2, panel A).

In random graphs, the number of births and deaths across thresholds follows a characteristic pattern (Kahle (2009)). At high thresholds and low edge density, a few low dimensional cavities exist, while at low thresholds and high edge density, more high-dimensional cavities exist (Fig. IB) (Kahle (2009); Horak et al. (2009)). Interestingly, geometric graphs – which can be used to instantiate spatial constraints on the topology – show a markedly different distribution. There are many low dimensional cavities, and fewer cavities with increasing dimension (Kahle (2018, 2009)) (Fig. IC). This general pattern has been recapitulated in functional networks constructed from firing of hippocampal neurons, indicating a geometric rather than random nature to neuronal co-firing (Giusti et al. (2015)). Furthermore, the persistent homology of human connectomes (Sizemore et al. (2017)) and rat microcircuits (Dotko et al. (2016)) is distinct from that expected in a minimally wired null model. In humans, the presence of widespread subcortical connections leads to more cavities being born at high densities (Sizemore et al. (2017)), while rat microcircuits display more high dimensional cavities in general (Dotko et al. (2016)). Further investigation into how wiring rules shape the topology of neural systems may shed light on how the brain’s spatial embedding shapes connectivity across scales and species.
Figure 2.5: **Applied algebraic topology.** (A) An illustration of thresholding a weighted network across different densities ($\rho$). At $\rho_1$, a cavity of dimension 1 is born (shown in yellow), which then dies at $\rho_3$. (B) The characteristic pattern of births and deaths (called Betti curves) for cycles of dimension 1 (yellow), dimension 2 (red), and dimension 3 (blue) from a random network. The cliques of each dimension are shown near each corresponding Betti curve for reference. (C) The same pattern, but for geometric networks. Different lines of the same color indicate different dimensions of embedding. This figure was adapted with permission from (Giusti et al. (2015)).
Box 3: Control Theory

Network control theory provides a potentially powerful approach for modeling neural dynamics (Schiff (2011)). Hailing from physics and engineering, network control theory characterizes a complex system as composed of nodes interconnected by edges, and then specifies a model of network dynamics to determine how external input affects the nodes’ time-varying activity (Liu et al. (2011)). Most studies of network control in neural systems stipulate a linear, time-invariant model of dynamics: $\dot{x}(t) = Ax(t) + Bu(t)$, where $x$ is some measure of brain state, $A$ is a structural connectivity matrix, $u$ is the input into the system (exogenous stimulation, or endogenous input from other brain regions), and $B$ selects the control set, or regions to provide input to (Yan et al. (2017b)). Assuming this model of dynamics, one can calculate the control energy required to reach specific brain states, which can be used as a state-dependent measure of the efficiency of control (Betzel et al. (2016b)). Control theory can also posit control metrics that quantify how efficiently a node would drive the brain to various states. Two commonly used metrics are average controllability and modal controllability (Pasqualetti et al. (2014b)). When every node is included in the control set, average controllability is proportional to the average energy required to drive the node to any state (Jeganathan et al. (2018)). Conversely, modal controllability is high in nodes where a small input will result in large perturbations to all eigenmodes of the system, and is interpreted to be high in nodes that can easily drive the brain to hard-to-reach states (Gu et al. (2015a)).

If these properties are important for helping the brain transition between states, one would expect them not to be randomly distributed across the cortex, but to be clustered into spatially constrained, functionally relevant systems. More specifically, one might expect functional systems that drive the brain to many accessible states, such as the default mode system, to have high average controllability, while regions that drive the brain to hard-to-reach, cognitively demanding states (executive control areas) to have high modal controllability. Data from healthy human adults supports these two hypotheses (Gu et al. (2015a)). Moreover, both average and modal controllability increase across development and are cor-
related with cognitive performance generally (Tang and Bassett (2017)). The manner in
which network control tracks individual differences reflects the fact that the capacity for a
network to enact control is dependent upon its topology (Menara et al. (2018a)). Further
efforts are needed to distill exactly how spatial embedding and wiring constraints impinge
on that control capacity, and how it is altered in psychiatric disorders (Jeganathan et al.
(2018)).
CHAPTER 3: Changes in functional connectivity associated with interictal epileptiform discharges (IEDs) in intracranial EEG

This chapter is based on an ongoing project with Stiso, J., Caciagli, L., Hadar, P., Davis, K.A., Lucas, T.H. and Bassett, D. S. Changes in human intracranial EEG derived functional connectivity associated with interictal epileptiform discharges (IEDs)
Abstract

All forms of epilepsy are defined by a propensity for periods of hypersynchronous activity during seizures. Understanding this unifying property would be advanced by a thorough quantification of how the sub-seizure synchrony of the epileptic brain responds to small perturbations. Here, we use transient focal bursts of epileptic activity called interictal epileptiform discharges (IEDs) to characterize this response. Specifically, we use a group of 143 participants with intracranial EEG (iEEG) coverage to report changes in 5 functional connectivity measures associated with three properties of IEDs: their presence, spread, and number. We perform this analysis in 5 frequency bands in order to contextualize our findings with ongoing neural processes at different spatial and temporal scales. We find that across frequency bands, both the presence and spread of IEDs tend to lead to independent increases of functional connectivity, but only in functional connectivity measures influenced by the amplitude, rather than phase, of a signal. We find that these increases are not explained by simple subgroups of connections such as the weakest connections in the brain, or only connections containing the seizure onset zone. We then report patterns of similarity across different band and measure combinations, and find the presence of IEDs impacts high frequencies (gamma and high gamma) and low frequencies (theta, alpha, and beta) differently, though responses within each group are similar. We then use grouped LASSO regression to identify which individual-level features explain differences in changes in functional connectivity associated with IEDs. While no feature robustly explains observed differences, the most consistently included predictor across bands and measures was the anatomical locus of IEDs. Overall, this work provides compelling evidence for increases in global synchrony associated with IEDs, and adds a thorough exploration of different functional connectivity measures, frequency bands, and IED properties to the literature on the impacts of IEDs. These observations show a disruption of several types of ongoing neural dynamics associated with IEDs. Additionally, they provide a starting point for future models of how small perturbations affect neural systems and how those systems support the hypersynchrony seen in epilepsy.
Epilepsy is a heterogeneous neurological disease that is characterized by a predisposition towards seizures (Falco-Walter et al. (2018)). While the different subtypes of epilepsy manifest with different types of seizures, levels of severity, co-morbidity of symptoms, the macroscale behavior of the brain exhibits a marked tendency towards hypersynchronous activity of the neural system that characterizes a seizure (Falco-Walter et al. (2018)). In order to better understand this property of hypersynchrony that leads to seizures, it is useful to understand how small perturbations impact endogenous patterns of synchrony in the brain. An interictal epileptiform discharge (IED) is a sub-seizure epileptic waveform that does not trigger broad hypersynchrony in a seizure. Nevertheless, IEDs often precede seizure activity and occur in spatially proximal regions (Prince and Connors (1986)). IEDs have been considered the simplest phenomenon from which to study epilepsy (Prince and Connors (1986)) and are known to perturb ongoing cognitive processes (Ung H et al. (2017)). Therefore represent a promising candidate to better understand the synchrony of the epileptic brain through it’s response to small perturbations. Quantifying the consistent changes in synchrony resulting from IEDs would represent an important step towards understanding the underlying properties of the epileptic brain that lead to seizures.

In EEG recordings, IEDs are single sharp, spike-like waveforms that result from a burst of firing from a small group of neurons (Prince and Connors (1986)). IEDs have quantifiable impacts on neural dynamics and behavior. For example these focal perturbations impact a battery of cognitive tasks, even when those IEDs occur outside the tissue supporting seizure formation: the seizure onset zone (Aldenkamp and Arends (2004); Ung H et al. (2017)). Additionally, IEDs are associated with spindle (Dahal et al. (2019)) and neural spiking (Keller et al. (2020)) activity in regions distant from the source of the IED. These findings both suggest that IEDs impact dynamics outside the population of cells producing the spike; however a direct quantification of those effects in a large sample of source-level recordings has been difficult to obtain, partially due to the rarity of the data. Work using scalp EEG, or fMRI has indirectly addressed this question by trying to separate out the contribution of IEDs to group differences in synchrony observed between individuals with
epilepsy and controls. Some work suggests that differences in intrinsic functional connectivity seen between individuals with and without epilepsy are due largely to IEDs, rather than to endogenous dynamics (Shamshiri et al. (2017)). However, depending on the cohort and method for quantifying synchrony, these conclusions can vary (van Houdt et al. (2015)). Many studies have also detailed local or system-level changes in synchrony associated with IEDs, especially near the onset of IEDs (Laufs et al. (2007); Tong et al. (2019); Lopes et al. (2014)). These findings are corroborated by nonlinear biophysical modeling studies showing a lack of global differences in functional connectivity associated with IEDs, but some local changes (Courtiol et al. (2020)). However, quantifying only local changes makes it difficult to connect with the global synchrony seen in diverse seizure states.

A thorough investigation of the effect of IEDs on global synchrony is warranted to provide key observations of the response of an epileptic brain to perturbations. These observations could help future researchers identify principles underlying the universal hypersynchrony seen in different epilepsies. Here, we consider a thorough investigation to have two important properties: (1) it will comprehensively test different quantifications of global synchrony and (2) it will use a large sample to identify changes that are consistent across individuals. In EEG recordings, synchrony is quantified by an array of functional connectivity measures that identify statistical similarities between two signals. Additionally, EEG recordings can quantify synchrony between different neural processes that are best reflected in either low or high frequencies and broadband activity. Low-frequency activity (theta, alpha, beta) in a single region is thought to reflect aligned fluctuations in the membrane potential of a large, local population of neurons. These fluctuations have the potential to modulate spiking activity to occur at the peaks of these oscillations (Fries (2005)). Distant regions with similar low-frequency activity, and therefore high functional connectivity, could be structurally connected regions with high-amplitude oscillatory activity where local spiking activity may or may not be modified (Schneider et al. (2020)). Some theories of cortical communication suggest a role for these spatially broad oscillations in top-down processing and attention (Canolty and Knight (2010); Riddle et al. (2019)). Higher frequencies (especially high-
gamma) in a single region are thought to correlate, though imperfectly (Leszczynski et al. (2020)), with local spiking activity (Crone et al. (1998a); Manning et al. (2009)). Multiple regions showing similar patterns in high frequency activity and high functional connectivity are thought to have similar local patterns of spiking activity. Both gamma and high-gamma connectivity are theorized to support bottom-up processing (Canolty and Knight (2010); Riddle et al. (2019)). Identifying which, if any of these processes are changing during IEDs can help connect basic principles of hypersynchrony to underlying biological processes generating the activity.

Here, we use 143 individuals from the RAM dataset (Ezzyat et al. (2017)), a publicly available dataset of individuals with epilepsy and intracranial EEG (iEEG) recordings to quantify changes in functional connectivity associated with IEDs that are not explained by changes in regional activity (Fig. 3.1A). We quantify connectivity based on phase (imaginary phase-locking value), amplitude (orthogonal amplitude envelope correlation) and amplitude-weighted phase (imaginary coherence) in 5 bands (3 low and 2 high): theta, alpha, beta, gamma, and high gamma (Fig. 3.1B). We also use the full broadband signal to investigate two measures that are not phase/amplitude separated: maximal cross-correlation and autoregressive fit. We then use linear regression to quantify the amount that each of these metrics changes based on the presence of an IED, the number of IED sequences in a 1-second time window, and the average number of contacts containing an IED in each 1-second time window (Fig. 3.1C and D). This information will allow us to determine whether there are consistent changes to functional connectivity associated with IEDs, whether these effects impact local spiking connectivity or spatially distributed lower-frequency connectivity, and whether these effects selectively impact phase- or amplitude-based connectivity. After identifying increases in functional connectivity associated with IEDs, we ask if these changes are driven by specific edges. We first hypothesize that increases might be driven by changes in the extreme ends of the distribution. Second, given that interictal connectivity is characterized by a globally disconnected (Burns et al. (2014)), but strongly intraconnected (Juárez-Martinez et al. (2018)) seizure onset zone, we hypothesized increases might
be driven selectively by changes within the seizure onset zone. After answering these basic questions, we next explore patterns of similarity across measures and bands. We conclude by assessing possible sources of individual differences in the magnitude of the effects of IEDs. This work provides a thorough quantification of changes in global functional connectivity beyond changes in activity associated with different properties of IEDs, and identifies several candidate neural processes that are disrupted during these simple perturbations. These observations contextualize IEDs within the complex ongoing dynamics of the brain and can now be used to validate models that test principles underlying the hypersynchrony in epilepsy.

Methods

Participants (RAM Dataset)

A publicly available dataset of 251 individuals undergoing intraoperative monitoring of their seizures was obtained from http://memory.psych.upenn.edu/RAM. In each subject, monitoring was conducted with either grid electrocorticography electrodes, stereo electroencephalography electrodes, or both. Recordings were completed while participants completed one of 5 different tasks testing either memory or free recall. Some participants underwent stimulation during these tasks. This data was collected as part of a multisite, collaborative effort at the University of Pennsylvania, Columbia University, Dartmouth College, Emory University, Thomas Jefferson University, the Mayo Clinic, the National Institutes of Health, the University of Texas Southwestern, Medtronic Inc, and the Lawrence Livermore National Labs.

Preprocessing Electrocorticography Data

First, raw data from the RAM dataset was segmented into 5 second or longer task-free epochs from either before or after task completion. If no information was available regarding the timing of task events, or if this information was inconsistent, the recording session was not processed. Data were then downsampled to the lowest sampling rate used across
Figure 3.1: Schematic of methods. (A) Data undergo automatic IED and artifact detection before being separated into 1-second windows. (B) The different functional connectivity measures being tested. (C) The different summaries of global functional connectivity being tested. Red region indicates the seizure onset zone. (D) Schematic of regression equation used to estimate effect sizes for 3 different properties of IEDs: their presence, number, and spread.
Figure 3.2: Flow Chart of Methods. (Preprocessing) Epochs of at least 5 seconds undergo preprocessing steps 1-5. Data are downsampled to 500 Hz, and rereferenced to a common average (CAR). (IED Detection) Clean data then undergo detection and selection of IEDs. (Artifact Detection) The same clean data have transient artifacts, specifically sharp electrode drift and periods of flatlining detected. These epochs are removed from further analyses. (Quality Control) Entire datasets are rejected based on any of 3 criteria numbered in the box. (Functional Connectivity) Data are then segmented into 1-second windows, and prewhitened within each window. All windows then undergo three manipulations: a multitaper fast Fourier transform (FFT), low-pass filter, and Hilbert transform. FFTs and Hilbert transforms are performed in 5 frequency bands. The multitaper FFT is then used to calculate coherence and power in each band. Low-pass filtered data is used to calculate a cross-correlation and autoregressive fit. Hilbert-transformed data is used to calculate an amplitude envelope correlation, and the phase-locking value. Connectivity measures are then summarized as the strength (mean connectivity) in 4 groups of channels: those in the seizure onset zone (SOZ), those outside the SOZ, each contact individually, and all channels. (Regression) Lastly, we fit a permutation-based linear model for each subject to obtain the standardized coefficients for a linear model associating a categorical indicator of an IED, the spread of IEDs, and the number of IEDs per window with each measure of strength. Power is included as a nuisance covariate in order to assess changes in functional connectivity that cannot be explained by changes in activity.
recording sites - 500 Hz - using the resample() function in MATLAB. Electric line noise and its harmonics at 60, 120, and 180 Hz (all sites were in the United States) were filtered out using a zero phase distortion 4th order stop-band Butterworth filter with a 1 Hz width. This filtering was implemented using the butter() and filtfilt() functions in MATLAB. For impulse and step responses of this filter see Supplemental Figure S3.1.

We then sought to remove individual channels that were noisy or had poor recording quality. The size of the dataset prevented us from visually inspecting each recording thoroughly, so we rejected channels using both the notes provided within the RAM dataset and using strict automated methods. After removing channels marked as low quality in the notes, we further removed electrodes that had either (1) a line length greater than three times the mean (Ung H et al. (2017)), (2) a z-scored kurtosis greater than 1.5 (Owen et al. (2017)), or (3) a z-scored power-spectral density dissimilarity measure greater than 1.5 (Betzel et al. (2017)). The dissimilarity measure was the average of one minus the Spearman’s rank correlation of that signal with the signals of all other channels. These automated methods should remove channels with excessive high-frequency noise, electrode drift, and line noise, respectively.

Data were then demeaned and detrended. Detrending was used instead of a high-pass filter to avoid inducing filter artifacts (de Cheveigné and Nelken (2019)). Channels were then grouped by grid or depth electrode, and common average referenced within each group. Following the common average reference, plots of raw data and power spectral densities were visually inspected by an expert researcher with 6 years of experience working with electrocorticography data (J.S.) to ensure that data were relatively clean.

**Automatic Interictal Epileptic Discharge (IED) Detection**

Automatic IED detection is still an open area of research and there is no consensus on the best practice (Brown III et al. (2007)). In clinical settings, epileptologists still manually mark EEG recordings for spikes and seizures when monitoring patients, though there are
not high levels of agreement across different experts (Gaspard et al. (2014)). We chose to
use an IED detector from Janca et al. (Janca et al. (2014)) because it is sensitive, fast, and
requires relatively little data per subject.

This Hilbert-based method dynamically models background activity and looks for outliers
from that background. Specifically, the algorithm first downsamples the data to 200 Hz,
and applies a 10-60 Hz bandpass filter. The envelope of the signal is then obtained by
taking the square of this Hilbert-transformed signal. In 5 second windows with overlap of
4 seconds, a threshold $k$ is calculated as the mode plus the median and used to identify
IEDs. The initial $k$ value is set to 3.65, which was determined through cross-validation in
the original paper (Janca et al. (2014)).

In order to remove false positives potentially caused by artifacts, we apply a spatial filter
to identified IEDs. Specifically, we remove IEDs that are not present in a 50 ms window of
IEDs in at least 3 other channels. The 50 ms window was taken from papers investigating
the biophysical properties of chains of IEDs, which tended to last less than 50 ms (Lai et al.
(2007)). Spikes detected within 80% of contacts within 2 ms were discarded. A subset of
spikes were then randomly selected and validated by a board certified epileptologist (K.D.).

In order to quantify the spread and number of IEDs, we also needed to determine where
sequences of temporally proximal IEDs began and ended. We employed an algorithm used
previously in refs (Conrad et al. (2020); Tomlinson et al. (2016)). IEDs occurring within 50
ms of the first IED in the sequence, or within 15 ms of the previous IED were considered
part of the same sequence. Sequences were discarded if 50% of the spikes in the sequence
occurred within 2 ms of each other.

Alternate IED Detector

To ensure that our results were not due to the specific IED detector used, we reproduced
key results using an alternate detector. The alternate detector was the Delphos (Detector
of ElectroPhysiological Oscillations and Spikes) detector (Roehri et al. (2017)). This de-
The detector was designed to detect and distinguish between both oscillations and IEDs in the time-frequency representation of the signal (Roehri et al. (2017)). IEDs are detected by analyzing the time width and frequency spread of peaks above a given threshold. Here, we chose a threshold value of 50 because it discovered a similar number of IEDs as our primary detector. This way, we know that changes between detectors are not driven by a different number of IEDs, but by differences in the properties of the individual IEDs discovered.

**Transient Temporal Artifact Rejection**

Visual inspection of the data revealed two types of temporally transient artifacts that we sought to remove from further analyses. The first type was sharp channel drift. Here, most channels simultaneously jump to a higher voltage before slowly drifting back to their original level. These sharp transients can cause artifacts that mimic oscillations when filtered (de Cheveigné and Nelken (2019)). To automatically detect these sharp artifacts we calculated the rate of change of the time series and looked for large outliers. An absolute threshold of 30,000 microvolts per sample was used, as it was determined to be well outside the normal range for several randomly selected datasets, while still capturing artifacts. All time points where at least half of all channels contained values greater than this threshold were removed from further analyses.

The type of second artifact we observed was the presence of large periods of flatlining across all channels. This flatlining could induce artificially large connectivity estimates as well as increased false positives in the IED detector. Since the detector dynamically calculates threshold values, periods of flat lining will lead to especially low thresholds and therefore many detected IEDs. Therefore, in addition to removing the periods of flatlining themselves, we also remove 5 seconds of data following the artifact. To identify periods of flatlining, we look for time points in the data with extremely small variance across all channels. We use a value of 300 microvolts because it was shown to be outside of the normal range but able to capture flat lining for several randomly selected subjects. All identified artifacts were visually inspected by an expert research with 6 years of experience working with electro-
corticography data (J.S.) to ensure that identification was working as expected.

**Dataset Rejection**

Our last quality control measure was to remove entire datasets from analysis that were especially noisy. Visual inspection of the data revealed three features indicative of low-quality data: (1) oddly shaped power spectral densities, (2) power spectral densities with many sources of line noise, and (3) many temporal artifacts. To identify datasets with uncharacteristic power spectra, we first calculated the average power spectral density across all channels using Welch’s method in 500 ms windows (pwelch() in MATLAB). We then calculated the Spearman’s rank correlation between the power spectral densities of all datasets. We removed the 10% of datasets with the lowest correlations.

To remove datasets with high levels of line noise, we started with the same power spectral densities calculated above and sought to identify those that were not fit well with a smooth curve. We then fit a smooth spline function to each power spectral density using fit() in MATLAB with a smoothing parameter of 0.01 Hz. Frequencies with notch filters, and frequencies below 10 Hz were excluded from the fit, since they often deviated from the smooth curve and increased error rates. For each dataset, we then calculated the sum of squared errors between the smooth fit and the power spectral density. Power spectral densities with line noise at many frequencies would not be fit well by the smooth curve, and have higher errors. We then removed the 10% of datasets with the highest error.

Lastly, we rejected any dataset with greater than 1,000 time points containing temporal artifacts. Most of these datasets were also removed by one of the other methods. Lastly, we visually inspected the remaining data to confirm that the remaining datasets looked clean. After performing these steps, we were left with data from 181 subjects.
Dynamic Functional Connectivity

In this work, we sought to quantify how functional connectivity between brain regions changes in association with IEDs. To test for these changes, we first split all task-free data into 1 second non-overlapping windows. If a window contained an IED, that window was realigned to start 1 sample before the first IED in that window. Successive windows would be shifted later so that they did not overlap. Once windows were defined, data were prewhitened within each window using the Fieldtrip Toolbox for MATLAB (Oostenveld et al. (2011)) (ft_preprocessing() function with 'derivative' parameter). To systematically explore the changes in functional connectivity following IEDs, we calculated 5 different commonly used metrics. For band-limited measures in the theta (θ, 4-8 Hz), alpha (α, 9-15 Hz), beta (β, 16-25 Hz), and gamma (γ, 36-70 Hz) bands, we wished to investigate amplitude-based, phase-based, and combined metrics. Therefore, we calculated orthogonal amplitude envelope correlations (amplitude), imaginary phase-locking value (phase), and imaginary multitaper coherence (combined). We also calculated the power in each band for each window to include as a covariate in later analyses. For broadband measures, we sought to characterize both undirected functional and directed effective connectivity. Therefore, we calculated the maximal cross-correlation, as well as a vector autoregressive model.

Orthogonal Amplitude Envelope Correlation: The amplitude envelope correlation (AEC) quantifies the extent to which signals from two channels change amplitude synchronously. For each bandpass filtered time-series \( y_n(t) \) at channel \( n \), the instantaneous amplitude is obtained from the analytic signal \( z_n(t) \) where \( z_n(t) = y_n(t) + iy_{Hn}(t) = \rho_n(t)e^{i\phi_n(t)} \) and \( \phi_n(t) \) is the instantaneous phase. Correlations between these amplitude envelopes are highly susceptible to artificial correlations due to volume conduction (Hipp et al. (2012); Nolte et al. (2019)). In electrophysiological signals specifically, similarities between channels at location A and B could be caused by the signal from a third source C spreading through the tissue and cerebrospinal fluid (CSF) that is being picked up by both channels, despite region C having no functional relationship to regions A and B. This process is called volume conduction, and while it is a much bigger problem for sensor-level recordings (EEG and
magnetoencephalography (MEG)), it can still have effects on iEEG data. Therefore, we take only the orthogonal components of each signal before calculating their correlation.

Here, we account for volume conduction using the method from Nolte et al. (Nolte et al. (2019)). We chose this method as opposed to the method from Hipp et al. (Hipp et al. (2012)) because this method uses a global normalization constant rather than one fit to each time point and is therefore much faster. Assuming Gaussian distributed data, which is a reasonable but not perfect fit to short segments of electrophysiological data, we can obtain the portion of the analytic signal $z_m$ that is orthogonal to $z_n$ by subtracting $z_n$ multiplied by the real part of the coherency spectrum between the two channels. Specifically, we first normalize the analytic signal from channels $n$ and $m$ such that $\langle |z_n|^2 \rangle = \langle |z_m|^2 \rangle = 1$. Here, the expected value is taken over time points. We then calculate coherency $c$ as $c = \frac{\langle z_n z_m^* \rangle}{\left( \langle |z_n|^2 |z_m|^2 \rangle \right)^{1/2}}$ where $*$ refers to the complex conjugate. Lastly, we calculate the correlation between $z_n$ and $z_m - \text{real}(c)z_n$. The resulting value ranges from -1 to 1, with 1 indicating perfectly correlated signals, and -1 indicating perfectly anti-correlated signals. The absolute value of the orthogonal amplitude envelope correlation was taken before averaging across edges.

**Bandpass Filter - Imaginary Phase-Locking Value (PLV)**

The phase-locking value quantifies how consistent the phase offset between two channels is over time regardless of the signal amplitude. The imaginary phase locking value removes 0-phase lag contributions to this value, which could arise from volume conduction. Using the bandpassed signal, $y_n(t)$, obtained with the parameters listed above, we calculate the imaginary phase locking value as $iPLV_{i,j} = \frac{1}{T} \text{imag}(e^{-i\phi_i(t)} - \phi_j(t))$. Here, $\phi_i$ is the instantaneous phase in channel $i$ for a given frequency band. This process was implemented using custom code in MATLAB, taken from (Bruña et al. (2018)). Instantaneous phase requires a narrow frequency range in order to be biologically interpretable; therefore we do not compute this measure on the high gamma band (Bastos and Schoffelen (2016)). This measure ranges between 0 and 1, where 1 indicated consistent and small phase offsets, and
0 indicates inconsistent or large phase offsets.

**Multitaper Fourier Transform - Imaginary Coherence and Power**

*Multitaper Fourier Transform:* Multitaper fast Fourier transforms (FFTs) use multiple tapers in order to better control spectral leakage at high frequencies (Subhash Chandran et al. (2016)). They can be used to obtain the cross-spectral density, $S_{i,j}$, between any two channels $i$ and $j$ as well as the power spectral density $S_{ii}$ for channel $i$. The multitaper FFT was computed on 30 logarithmically spaced frequencies between 4 and 150 Hz using DPSS tapers with 4 Hz smoothing. Trials were zero-padded to the maximum period length for a given frequency. This calculation was implemented in MATLAB with the Fieldtrip toolbox (Oostenveld et al. (2011)).

*Imaginary Coherence:* Coherence quantifies the consistency of phase offsets between two channels, weighted by their signal amplitude. Imaginary coherence ignores the contribution of 0-phase lag signals to this value, which could arise from volume conduction. Mathematically, the complex valued coherence $c$ between two signals is $c = \frac{S_{i,j}}{\sqrt{S_{ii}S_{jj}}}$. The imaginary coherence $C$ is then taken as the absolute value of the imaginary component of coherence, $|\text{imag}(c)|$. Imaginary coherence ranges from 0 to 1. Imaginary coherence values were averaged across all frequencies within a band.

*Power:* We wished to quantify changes to connectivity above what would also be explained by changes in activity. Here, we quantify activity as spectral power. Spectral power for each band and each channel was calculated by taking the logarithm (base 10) of all frequencies that fell within a given band.

**Low-Pass Filter - Cross-Correlation**

*Low Pass Filter:* For broadband measures, we first removed sources of high-frequency noise using a low pass filter. Data were low pass filtered at 200 Hz using a 0-phase lag 4th order Butterworth filter. This filter was implemented using the MATLAB package Fieldtrip.

*Cross Correlation:* The cross-correlation is the maximal correlation between two signals
that can be achieved across shifts of one signal relative to the other. The correlation coefficient was normalized such that the autocorrelation at zero lag was equal to one. The cross-correlation was implemented with the MATLAB function xcorr(). The resulting values range between 0 and 1.

**Low-Pass Filter - Effective Connectivity from Autoregressive (AR) Models**

*Autoregressive (AR) Models:* All the measures described above are undirected functional connectivity measures that will always have the same relationship between node \( i \) and \( j \) as between node \( j \) and node \( i \). We also wished to capture a directed measure of connectivity in which the influence of node \( i \) on node \( j \) could be different from the influence of node \( j \) on node \( i \). To accomplish this goal, we calculated a measure of directed connectivity from the weights of a first-order autoregressive model fit to the data. The weights were obtained by solving for \((A)\) in the equation \( \mathbf{x}(t) = A\mathbf{x}(t-1) + \epsilon(t) \), where \( \mathbf{x} \) is the timeseries data, and \( \epsilon \) is an error term. Fits were calculated using the arfit package in MATLAB (Schneider and Neumaier (2001)). Connectivity values from AR models range from -1 to 1, so the absolute value was taken before summarizing. AR models assume a linear relationship between signals and assume the data are stationary. For short time windows (here, 1 second) this is a reasonable assumption.

*Estimating Effects of IEDs on Functional Connectivity*

The overarching goal of this study is to quantify the effect of IEDs on each of these functional connectivity metrics. We also wished to be able to answer the three specific questions. Are changes in connectivity driven by shifts in the weakest or strongest connections? Are changes larger within the seizure onset zone than outside it? Do these effects vary reliably across regions or individuals? To answer these questions, we calculated the effects of IEDs on 5 different summary measures of the distributions of connections. These measures included the average connectivity across all contacts (full coverage), the skew of connectivity across all contacts, the strength of containing the seizure onset zone, the strength of connectivity contained contacts outside the seizure onset zone, and the strength of connectivity in each
individual contact (Fig 3.1C). In supplemental analyses, we also report the changes in connectivity in two additional summaries: (1) containing any contacts where IEDs are present and (2) containing any contacts where IEDs are absent.

We next quantify the specific features of each IED sequence that we hypothesized would impact functional connectivity uniformly. The first feature is the presence of an IED in a window, regardless of the properties of that IED. The second is the number of sequences within a window. A sequence is defined as all IEDs that occur within 50 ms of the first IED, or within 15 ms of the last IED in that timeframe. If the leader channel had multiple spikes in this sequence, the initial sequence was split into multiple sequences. The third is the average spread across each sequence within the window (Fig 3.1D).

For each dependent variable, we calculated the coefficient of each IED predictor in a permutation-based linear model that contained power and recording session as nuisance covariates. If two or more of the three covariates of interest were perfectly colinear, one of them was removed from the model. The results presented in the main text are only from the subset of 143 subjects that had coefficient values for all three predictors. Outliers outside of 3 standard deviations from the mean distribution for each band and measure combination were removed from further analyses. Permutation-based models were used because both the spike spread and the number of spikes are highly skewed, non-normal variables that could lead to artificially large estimations in parametric models. Additionally, permutation-based models more effectively down-weight outliers in the distribution.

*Individual-Level Variables*

We investigated the relative efficacy of 12 variables in explaining individual differences in the changes in functional connectivity associated with IEDs. These 12 variables were sex, race, handedness, age, locus of IEDs, hemisphere of IEDs, etiology, the presence of a lesion, the age of seizure onset, the subject’s average task performance, the institution of treatment, and the type of contacts (grid or depth electrodes). Of these variables, sex, race, institution,
age, type of coverage, and handedness were included as they were recorded in the public data release. If subjects had missing values in any of these demographic variables, they were filled in with the most common demographic for categorical variables, or the mean demographic for continuous variables. Other variables required processing before including them in the model or were not included in the public data release.

Clinical Variables

We received additional clinical data that was not available in the public release from the creators of the original study. This included the age that seizures began for each patient, whether or not the patient had a lesion, and the underlying etiology. The underlying etiology was put into one of 10 categories: (1) traumatic brain injury; (2) infection; (3) neurocutaneous syndrome; (4) neoplasia; (5) stroke; (6) malformation of cortical development; (7) medial temporal sclerosis; (8) hypoxemic ischemic encephalopathy; (9) other identified etiologies in a different category; and (10) unknown etiology. The variable with missing values for the most people was etiology, which was missing in 80% of our 160 participants. Individuals with missing values were not included in our final analysis.

Contact Variables

The locus and hemisphere of IEDs were obtained from information provided in the public data release. MNI coordinates provided were registered to the Schaefer parcellation (Schaefer et al. (2018)) of 7 cognitive systems (Yeo et al. (2011)). It is standard practice to register atlases to a subject-specific space, rather than registering subject coordinates to a common space. To assure that we were confident in our system assignments, once each contact had been assigned a system, we manually checked that the physician-assigned regions of each contact and their given system matched. Participants with missing MNI coordinates were excluded from these analyses (76.3%).

Task Performance

The last included variable was each participant’s average task performance on a subset of 5 tasks. Three tasks were free recall tasks and 2 were spatial navigation tasks. For all
tasks, performance ranges from 0 to 1; however, a value of 1 in free recall tasks indicated
good performance, while a value of 1 in navigation tasks indicated poor performance. We
averaged each participant’s reverse score on all free recall tasks with \((1 – \text{navigation})\). Any
trials where the subjects received stimulation were removed. Most subjects (97.5%) com-
pleted at least 1 task, but those who did not were removed from these analyses.

**Group LASSO**

We used group LASSO regression to assess the relative importance of each variable. Group
LASSO applies a penalty to groups of variables and will regularize coefficients of uninforma-
tive groups to 0. Here, all levels of a given categorical variable were grouped together. This
way, all levels of the etiology, for example, variable would be regularized together, rather
than regularizing individual levels. Group LASSO was implemented using the gglasso pack-

**Statistical Analyses**

Here, we set out to complete an exploratory analysis of the effect sizes associated with
changes to functional connectivity and IEDs. We used a strict family-wise Bonferroni
correction in which each band/measure combination was treated as a different hypothesis.
This procedure allowed us to have more confidence in the results we report but its stringency
may contribute to false negatives.

Before any significance testing, subjects with effect sizes greater than 3 standard deviations
greater than the mean were removed. Quantile-quantile plots were then used to check if
distributions were normal. Since distributions appeared to deviate from a normal distri-
bution, one sample permutation tests were used to test whether distribution means were
significantly different from 0 (Figs. 3.3-3.4). Statistics are reported with the test statistic,
the number of observations, and the \(p\)-value. Similarity was assessed with Spearman’s cor-
relations (Figs. 3.5-3.6) to avoid excessive influence of outliers. Differences within versus
between groups in similarity matrices were assessed with Pearson’s correlations between the
lower diagonal of entries, and a binary mask specific to each hypothesis. Pearson’s correlations assume independent observations, which is not the case for the entries in the similarity matrices used for Figs. 3.5 and 3.6. Using tests intended for independent observation on clustered data can result in p-values that are artificially small due to the artificially high sample size (Sainani (2010)). However, in these analyses, it is not the assessment of the significance of correlations that interests us, but rather the relative similarity values that are found with different masks. Because of this, we include Pearson’s $r$ as a simple and interpretable estimate of similarity.

**Data and Code**

Code is available at github.com/jastiso/interictal_fc.

Data is publicly available at http://memory.psych.upenn.edu/RAM.

**Results**

We sought to characterize the changes in global functional connectivity associated with a simple form of epileptic activity - an IED - in individuals with medically refractory epilepsy. Accordingly, we use a large sample of 143 participants in the RAM dataset, and we obtain at least 5 seconds of task-free, clean iEEG data. We automatically detect artifacts, and exclude those time points from further analysis. We next automatically detect IEDs, and segment the data into consecutive 1-second windows. If a window contained an IED, the start of the window was shifted to align with the start of the IED.

We select a comprehensive sample of functional connectivity measures common to iEEG analysis. Specifically, we calculate 3 measures of band-limited functional connectivity and 2 measures of broadband functional connectivity in each window. We used the following band-limited measures: orthogonal amplitude envelope correlation, imaginary coherence, and imaginary phase-locking value. These measures were all calculated in the theta ($\theta$, 4-8 Hz), alpha ($\alpha$, 9-15 Hz), beta ($\beta$, 16-25 Hz), and gamma ($\gamma$, 36-70 Hz) frequency bands. Amplitude influenced measures (orthogonal amplitude envelope correlation and imaginary
coherence) were also calculated in the high gamma range ($\gamma^+$, 71-150 Hz). The phase locking value is not calculated in the high gamma range because the instantaneous phase is only biologically interpretable for narrow bands (Bastos and Schoffelen (2016)). We used the following broadband measures: cross-correlation and an autoregressive fit (Fig. 3.1B).

Lastly, we quantified the effect of 3 different properties of IEDs on each band-measure combination using a permutation-based linear model including the presence of an IED, the number of IEDs in a window, and the average spread of every IED in the window. In order to identify drivers of spatial distributions of observed effects, we calculate the effects of IEDs on the total strength across all contacts, the skew of edges across all contacts, the strength of only those contacts within (or outside) the seizure onset zone, and in each contact individually (Fig. 3.1C).

**Quantifying the impact of IEDs on functional connectivity**

We first asked which band-measure combinations would show consistent changes in the strength of functional connections for each of the 3 IED predictors. We find different patterns of responses across predictors. The presence of IEDs increases functional connectivity, but largely in the orthogonal amplitude envelope correlations (Fig. 3.3A, one-sample permutation test, Bonferroni-corrected $p$-values ($n=16$) $d_{\theta,\text{AEC}}(142) = 0.28$, $p = 3.2 \times 10^{-3}$, $d_{\alpha,\text{AEC}}(141) = 0.25$, $p = 3.2 \times 10^{-2}$, $d_{\beta,\text{AEC}}(143) = 0.46$, $p < 1.0 \times 10^{-4}$, $d_{\beta,i\text{Coh}}(140) = 0.19$, $p < 1.0 \times 10^{-4}$, $d_{\gamma,\text{AEC}}(142) = 0.45$, $p < 1.0 \times 10^{-4}$, $d_{\gamma,i\text{PLV}}(142) = 0.10$, $p = 6.4 \times 10^{-3}$, $d_{\gamma^+,\text{AEC}}(142) = 0.50$, $p < 1.0 \times 10^{-4}$). The number of IEDs in a window typically does not consistently affect functional connectivity. The only exceptions were the orthogonal amplitude envelope correlation in the beta band and high gamma band (Fig. 3.3B, one-sample permutation test, Bonferroni-corrected $p$-values ($n=16$) $d_{\beta,\text{AEC}}(143) = 0.40$, $p < 1.0 \times 10^{-4}$, $d_{\gamma^+,i\text{Coh}}(141) = 0.01$, $p = 3.2 \times 10^{-3}$). The spread of IEDs within a window largely increases both amplitude (orthogonal amplitude envelope correlation) and amplitude weighted phase (imaginary coherence) based measures (Fig. 3.3C, one-sample permutation test, Bonferroni-corrected $p$-values ($n = 16$) $d_{\theta,\text{AEC}}(142) = 0.06$, $p = 6.4 \times 10^{-3}$,
\( d_{\theta,i\text{Coh}}(142) = 0.06, p < 1.0 \times 10^{-4}, d_{\theta,i\text{PLV}}(143) = -0.07, p < 1.0 \times 10^{-4}, d_{\alpha,AEC}(142) = 0.08, p < 1.0 \times 10^{-4}, d_{\alpha,i\text{Coh}}(142) = 0.07, p < 1.0 \times 10^{-4}, d_{\beta,AEC}(141) = 0.13, p < 1.0 \times 10^{-4}, d_{\beta,i\text{Coh}}(141) = 0.08, p < 1.0 \times 10^{-4}, d_{\gamma,AEC}(141) = 0.22, p < 1.0 \times 10^{-4}, d_{\gamma,i\text{Coh}}(141) = 0.06, p < 1.0 \times 10^{-4}, d_{\gamma+,AEC}(139) = 0.12, p < 1.0 \times 10^{-4}, d_{\gamma+,i\text{Coh}}(142) = 0.03, p < 1.0 \times 10^{-4}, d_{\text{broadband,xcorr}}(140) = -0.09, p = 6.2 \times 10^{-3} \).

The predictor for the spread of IEDs also has effect sizes an order of magnitude smaller than the other two predictors. While individuals can have both positive and negative changes to functional connectivity associated with IEDs, here we use a large sample (\( n = 143 \)) of source-level data to demonstrate consistent increases in global functional connectivity as a result of the presence and spread of IEDs.

To ensure that results are driven by IEDs themselves and not by artifacts picked up by our specific IED detection algorithm, we repeated the above analysis with a second IED detector. This second detector found significant effects for all the measures listed above, except the cross-correlation (spread), theta imaginary phase-locking value (spread), and beta as well as high gamma orthogonal amplitude envelope correlation (number) (Fig. 3.2, indicated with dark red asterisks). This algorithm also identified the only additional measure that had a mean significantly different from 0 was gamma imaginary phase-locking value for the spread of IEDs. This observation indicated that while many of our findings generalize to multiple detectors, the selection of IED detector can influence the quantification of how neurophysiology changes in association with IEDs.

**Quantification in subsets of connections**

We next asked if these global effects were driven largely by smaller parts of the network. Because we found no reproducible large effect sizes for the number-of-IEDs predictor, we limit this analysis to the presence and spread of IEDs (for the number of IEDs, see Fig. S3.2). We hypothesized that increases in global connectivity were driven by a strengthening of the weakest connections, possibly connecting the seizure onset zone to the rest of the brain, or a strengthening of selectively already strong, isolated connections within the seizure onset zone. We broke these hypotheses first into two questions: (1) could increases be
Figure 3.3: **Changes in Functional Connectivity Associated with IEDs.** (A) Distributions of coefficients for the presence of an IED obtained from permutation based regression including all IED predictors, power in a given band, and the recording session. Columns indicate different frequency bands, and colors indicate different measures. Bright red asterisks indicate significant distributions after multiple comparisons correction that could also be reproduced with a different spike detector. Dark red asterisks indicate those were not reproducible with another IED detector. * = \( p < 0.05 \), ** = \( p < 0.01 \), *** = \( p < 0.001 \), **** = \( p < 0.0001 \). (B,C) Same as in panel A, but for the coefficients associated with the number of IEDs and the spread of IEDs. Here, AEC stands for amplitude envelope correlation, imag. for imaginary, PLV for phase locking value, and AR for autoregressive model.
driven by the weakest edges; and (2) could increases be driven by edges in the seizure onset zone? To address our first question, we repeated the above analysis with the skew, rather than the strength of edges. Since edge distributions are heavy tailed (see Fig. S3.3), increases in the skew are consistent with a strengthening of the weakest edges in the distribution. We observe far fewer significant changes to the skew of connections than the strength, and we note that these changes tend to be negative. Significant changes were seen in theta, alpha, and beta imaginary coherence (spread), beta orthogonal amplitude envelope correlation (spread), gamma orthogonal amplitude envelope correlation (presence) and cross-correlation (presence) (Fig. 3.4A, one-sample permutation test, Bonferroni-corrected p-values \( n = 16 \) \( d_{\gamma,AEC}(141) = -2.13, p < 1.0 \times 10^{-4}, d_{\theta,iCoh}(143) = -0.39, p = 0.026, d_{\alpha,iCoh}(143) = -0.45, p = 6.4 \times 10^{-2}, d_{\beta,AEC}(142) = 0.35, p = 6.4 \times 10^{-2} \)). This finding indicates that most changes observed above are not driven by a strengthening of weak connections.

We address our second question by subtracting coefficients for the effect of IEDs on connections inside the clinically defined seizure onset zone (SOZ) from connections outside the seizure onset zone, for subjects where this data was available \( n = 103 \). Large positive numbers would indicate greater effects inside the SOZ. We find that after multiple comparison’s correction, only the effect sizes for the spread of IEDs were larger between the seizure onset zone and the rest of the brain (Fig. 3.4B). While these differences are seen in all but the theta band, they are most pronounced in the connectivity of the high gamma band (Fig. 3.4B, one-sample permutation test, Bonferroni-corrected p-values \( n = 16 \) \( d_{\alpha,iCoh}(102) = 0.04, p = 0.03, d_{\beta,AEC}(101) = 0.06, p = 0.02, d_{\gamma,iCoh}(101) = 0.03, p = 9.6 \times 10^{-3}, d_{\gamma+,AEC}(100) = 0.06, p < 1.0 \times 10^{-4}, d_{\gamma+,iCoh}(99) = 0.02, p = 0.02 \)). Some measures, indicated with a grey asterisk in Fig. 3.4, showed significant differences before multiple comparisons correction, and also suggested larger effects inside the seizure onset zone. Overall, we find that most changes are not driven by increases within the seizure onset zone, with one notable exception being changes associated with the spread of IEDs in the high gamma band.
We then tested a third hypothesis, that changes were driven by channels containing IEDs. These changes could be due to biological disruptions in ongoing interactions or spurious changes in statistical similarity associated with filtering artifacts arising from the spikes themselves. We repeated our analysis in channels containing IEDs for subjects who only displayed IEDs in a portion of contacts (n = 99). We then calculated differences in coefficients between contacts that contained IEDs and contacts that did not. We find 5 significant differences after multiple comparisons correction for the spread of IEDs in the imaginary coherence in the theta band, the imaginary coherence in the alpha band, the orthogonal amplitude envelope correlation and imaginary coherence in the gamma band, and the imaginary coherence in the high gamma range (Fig. S3.4, one sample permutation test, Bonferroni-corrected p-values (n = 16) $d_{\theta,iCoh}(99) = 0.04, p = 0.04$, $d_{\alpha,iCoh}(98) = 0.03, p = 0.04$, $d_{\gamma,AEC}(96) = 0.08, p < 1.0 \times 10^{-4}$, $d_{\gamma,iCoh}(97) = 0.03, p = 6.4 \times 10^{-3}$, $d_{\gamma+iCoh} = 0.03, p = 1.0 \times 10^{-4}$). We further tested if only connections between channels without IEDs showed significant changes associated with IEDs in bands and measures that showed both significant increases in global connectivity associated with IEDs (Fig. 3.3) and significant differences between contacts with and without IEDs (one sample permutation test, Bonferroni-corrected p-values (n = 16) $d_{\theta,iCoh}(98) = 0.04, p < 1.0 \times 10^{-4}$, $d_{\alpha,iCoh}(100) = 0.04, p = 6.4 \times 10^{3}$, $d_{\beta,AEC}(97) = 0.06, p < 1.0 \times 10^{-4}$, $d_{\gamma,AEC}(98) = 0.10, p < 1.0 \times 10^{-4}$, $d_{\gamma+iCoh}(98) = 0.03, p < 1.0 \times 10^{-4}$, $d_{\gamma+iCoh} = 0.01, p = 0.33$). We found that all measures except for imaginary coherence in the high gamma band remain significant after only considering connections between contacts not containing an IED. Overall, this finding suggests that while increases in functional connectivity are sometimes larger in affected regions, they are not limited to those regions.

*Relationships between IED predictors and changes to connectivity within frequency bands and connectivity measures*

Above, we explored a large space of IED predictors and their effects on different frequency bands and connectivity measures. In order to distill these many findings into broader,
Figure 3.4: Changes to specific connections. (A) The coefficient associated with each predictor for explaining changes to the skew of the distribution of functional connectivity. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$. Grey asterisks indicate significant differences before multiple comparisons correction. (B) The difference between coefficients found for changes to connections within the seizure onset zone versus outside it.
more unifying principles about the response of neural systems to simple epileptic events, it would be useful to determine whether there are consistent associations between different predictors, frequency bands, or measures. To identify these associations, we calculate similarity with Spearman’s correlations of participants’ coefficients between IED predictors and band-measure combinations.

We first calculated pairwise correlations between the distribution of participants’ coefficients for each of the IED predictors, in each band-measure combination. We find that relationships are largely consistent across bands and therefore show results from each measure averaged across bands (**Fig. 3.5A**). We find that the strongest effect is a significant positive correlation between the coefficients for the number of IEDs and the presence of an IED (**Fig. 3.5A**, Spearman’s correlation, Bonferroni-corrected \( p \)-values \( n = 16 \) mean \( r_{AEC} = 0.66 + / - 0.035 \), mean \( r_{iCoh} = 0.58 + / - 0.058 \), mean \( r_{iPLV} = 0.60 + / - 0.037 \), \( r_{xcorr} = 0.52 \), \( r_{AR} = 0.67 \), all \( p < 2.55 \times 10^{-5} \)). The correlations between the coefficients for the presence of an IED and the spread of an IED tend to be significant and positive but weaker (**Fig. 3.5A**, Spearman’s correlation, Bonferroni-corrected \( p \)-values \( n = 16 \) mean \( r_{AEC} = 0.456 + / - 0.034 \), mean \( r_{iCoh} = 0.45 + / - 0.025 \), mean \( r_{iPLV} = 0.50 + / - 0.052 \), \( r_{xcorr} = 0.44 \), \( r_{AR} = 0.39 \), all \( p < 4.01 \times 10^{-4} \)), and the correlation between the coefficients number of IEDs and the spread of IEDs tends to be nonsignificant and much weaker (**Fig. 3.5A**, Spearman’s correlation, Bonferroni-corrected \( p \)-values \( n = 16 \) mean \( r_{AEC} = 0.054 + / - 0.029 \), mean \( r_{iCoh} = 0.097 + / - 0.051 \), mean \( r_{iPLV} = 0.086 + / - 0.067 \), \( r_{xcorr} = 0.051 \), \( r_{AR} = 0.096 \), all \( p < 0.013 \)). For the number and spread of IEDs, only the correlation in the theta band imaginary phase-locking value and coherence was statistically significant for these predictors. Due to the lack of consistent changes in functional connectivity resulting from the number of IEDs and the predictor’s high similarity with the presence of an IED, we exclude this predictor from further analysis.

Next, we sought to characterize similarities between each band-measure combination. Here we test three hypotheses. First, we hypothesize that coefficients will be similar across
measures within a given frequency band. Our second hypothesis is that coefficients will be similar within a measure, across frequency bands. Our third hypothesis is that there will be a patterned similarity across bands. Specifically, the low frequencies (theta, alpha, and beta) will be similar to each other, but dissimilar to high frequencies (gamma and high gamma) and vice versa. High and low frequencies are often described in the literature as showing opposing patterns, and are theoretically proposed to have opposite roles (Canolty and Knight (2010)). To test these three hypotheses, we calculate similarity matrices for each predictor. Each element of the similarity matrix is the Spearman’s correlation between the distribution of coefficients across subjects for two band-measure combinations. We show these matrices sorted by frequency band and by measure (Fig. 3.5B).

To test our first hypothesis, we correlate the upper triangles of these pairs of similarity matrices with masks that operationalize each hypothesis. For example, in the band similarity hypothesis, the mask has ones for entries from the same band and zeros for entries from different bands. We find similar correlations to the frequency band mask for both the spread and presence of IEDs (Fig. 3.5C, Pearson’s correlation with Bonferroni-corrected p-values \( n = 3 \) \( r_{\text{presence}}(112) = 0.37 \) \( r_{\text{spread}}(112) = 0.34 \), though the correlation is slightly stronger for the presence of IEDs. To test our second hypothesis, we repeated the same analysis, but with a measure mask. Here, we find that correlations are strongest for the spread-of-IEDs predictor (Fig. 3.5C, Pearson’s correlation with Bonferroni-corrected p-values \( n = 3 \) \( r_{\text{presence}}(112) = 0.27 \), \( r_{\text{spread}}(112) = 0.41 \)). To test our third hypothesis, we repeated the same analysis with a grouped frequency (high or low) mask. Here, we find much larger correlations in the presence of IEDs (Fig. 3.5C, Pearson’s correlation with Bonferroni-corrected p-values \( n = 2 \) \( r_{\text{presence}}(112) = 0.62 \), \( r_{\text{spread}}(112) = 0.40 \)). Additionally, this mask explained much more variance than each band individually for the presence, but not spread, of IEDs. Ultimately, we find that both the bands and measures tend to have similar effects. Additionally, the presence of IEDs seems to show more similar effect sizes within bands than within measures, though those effects differ between high and low frequencies.
Figure 3.5: **Similarity across predictors, measures, and bands.** (A) Spearman’s correlations between predictors. A bar plot showing the average correlation and standard error for each measure across bands. Error bars are the standard error across bands, when there are multiple bands to test. Colors indicate different measures. (B) Similarity matrices between all band-measure combinations. (Left) Matrices ordered by band. (Right) Matrices ordered by measure. (C) Correlations with a frequency band mask (left), measure mask (middle), and frequency band group, i.e. high or low (right) for each predictor. Error bars show the 95% confidence interval for each correlation.

**Explanatory Sources of Individual Variability**

We next sought to understand sources of individual variability in effect sizes. However, participants in this study have coverage over a unique set of regions, making it difficult to determine whether differences across individuals are due to differences in coverage or in characteristics of the participants. To test whether it is likely that changes across subjects are due to differences in electrode placement, we wished to quantify whether effect sizes were more similar within individuals or within regions. To quantify similarity, we calculated the Spearman’s correlation between the vector of effect sizes across all bands, measures, and predictors between each individual contact (**Fig. 3.6A and B**). We then calculated the correlation between these similarity matrices and masks selecting for entries from either...
the same individual or the same region. Here, regions are clinician-provided labels for each contact. We find a much larger Pearson’s $r$ value for the subject mask, compared to the region’s mask (Pearson’s correlation, Bonferroni-corrected for multiple comparisons ($n = 2$) $r_{\text{subj}}(2, 521, 117) = 0.14$, $r_{\text{region}}(2, 521, 117) = 1.34 \times 10^{-3}$). This observation motivated us to continue investigating individual, rather than regional differences in effect sizes.

We next tested which participant-specific features would best predict effect sizes for each band and measure using group LASSO regression. Group LASSO applies a penalty to groups of coefficients in a regression equation, allowing coefficients that do not greatly increase the explanatory power of the model to be regularized to zero. A parameter $\lambda$ scales the regularization, such that a larger $\lambda$ will result in a stricter penalty. Here, we chose the value of $\lambda$ that minimized mean squared error across 5-fold cross validation.

We hypothesized that behavioral, clinical, contact, and demographic factors might impact the magnitude of effect sizes of IEDs on functional connectivity. We include one behavioral variable, which is the average task performance across a battery of 5 possible free recall and spatial navigation tasks completed by each participant. We include 3 clinical variables: (1) whether the participant had a lesion, (2) the age that seizures began, and (3) the underlying etiology of their epilepsy. We include 5 demographic variables: (1) age, (2) race, (3) sex, (4) handedness, and (5) institution. Lastly, we include 3 contact variables: (1) the cognitive system (Yeo et al. (2011)) that contains the contact with the most IEDs, (2) the hemisphere that contains the contact with the most IEDs, and (3) whether the participant had grid contacts, depth contacts, or both. Of 143 participants, 99 participants had all these fields of information.

We restrict our analyses to only measures that had effect sizes significantly different from 0 in both IED detectors, the orthogonal amplitude envelope correlation, and imaginary coherence. The variables included in each grouped LASSO regression are shown in the table in Figure 3.6C, ranked by the size of their regularized beta value. For categorical variables, the largest beta value assigned to a contrast at a given level was used. For both the
presence and spread of IEDs, we find that some bands and measures are not well explained by any of the included measures (high gamma imaginary coherence for presence, alpha orthogonal amplitude envelope correlation, beta orthogonal amplitude envelope correlation and imaginary coherence, and high gamma orthogonal amplitude envelope correlation and imaginary coherence for spread). The most commonly included variable was the cognitive system containing most IEDs (IED locus). This variable was included in 14 out of 16 models. Most often, this effect was driven by large effect sizes in the frontoparietal control system.

Discussion

Here, we complete a thorough quantification of the changes in functional connectivity associated with IEDs in a large sample of 143 individuals. We test 5 different frequency bands as well as broadband signal for changes associated with 5 different measures of functional connectivity and 3 properties of IEDs. One key insight from our studies is that despite tremendous heterogeneity in IED properties both within and between individuals (Conrad et al. (2020)), we observe consistent increases in functional connectivity associated with IEDs. Our study also gives us the capability to specify which properties of IEDs are responsible for the observed changes. We find that once a single IED has occurred, more IEDs within a time window do not further disrupt ongoing functional connectivity, whereas IEDs that spread to more contacts further increase connectivity. Additionally, individuals whose functional connectivity changes significantly in response to the presence of an IED tend to show large changes associated with the spread of IEDs. Our work reveals that only amplitude-influenced measures (orthogonal amplitude envelope correlations and imaginary coherence) show changes, while phase-based and broadband measures remain more stable on average. This result is consistent with work showing that these two dimensions are not redundant biologically (Siems and Siegel (2018)). Across bands, we see that effects tend to be similarly sized within low and high frequency groups individually, but not across them. We also observe a high degree of individual variability in effect sizes. No single feature robustly explained the variation observed across all measures and bands, but the cognitive
Figure 3.6: Individual differences. (A) Similarity matrices for coefficient distributions across all predictors, bands, and measures, but between contacts. (Left) Matrices ordered by region. (Right) Matrices ordered by participant. (B) Correlations for masks selecting for participants, or regions. Error bars are the 95% confidence interval for each correlation. (C) Predictors included for each model. Predictors listed at the top of lists of a given color had higher regularized beta values than those at the bottom. Hem. is an abbreviation for hemisphere.
system that IEDs originate from explains the most variance. The changes reported here contribute substantial evidence towards an important unanswered question in epilepsy research regarding the impact of IEDs on functional connectivity. Therefore, these results have important implications for the preprocessing, analyses, and interpretation of research conducted with human iEEG data.

Evidence for the disruption of multiple physiological processes by IEDs

Our results show that the responses to the presence of IEDs are more similar within groups of high and low frequencies than with each band individually. While we cannot infer the mechanisms that generate oscillations from only their frequency band, previous research has reported differing roles for oscillations in low- and high-frequency bands, respectively. For example, early intracranial recordings show that lower frequency oscillations (<30 Hz) tend to be much more spatially distributed than higher frequency oscillations (Crone et al. (1998a,b)). Further work has shown that power in the highest frequency band we investigate here, high gamma, can be correlated with localized spiking activity (Manning et al. (2009); Leszczyński et al. (2020)). Additionally, while low frequency signals are thought to impact top-down cognitive processes, high frequency signals are thought to reflect bottom-up sensory driven processing (Singer et al. (2001); Richter et al. (2017)). Here we provide evidence that each of these processes is disrupted during IEDs but that each might be impacted differently.

The results shown here point to two important future directions. First, it would be of interest to investigate changes in cross-frequency interactions and the slope of the power spectral density associated with IEDs. High and low frequencies have also been shown to have a preferential interaction direction, where the amplitude of higher frequencies is modulated by the phase of lower frequencies (Canolty and Knight (2010)). While these interactions can be spurious and require strict testing for the presence of oscillations above the 1/f background and against null models that preserve autocorrelation (Gerber et al. (2016); Tort et al. (2010)), a detailed investigation of which of these interactions are changed dur-
ing IEDs would help further elucidate the neurophysiological processes that they disrupt. Second, opposing changes in high- and low-frequency bands could be explained by changes in the slope of the 1/f background, rather than changes in each band individually (Cole and Voytek (2017)). Here, since we do not see strongly anti-correlated patterns between low and high frequencies, we do not think this is driving the observed patterns. However, investigating changes in the slope of the 1/f background, as well as other non-sinusoidal features of the signal would also add significantly to our understanding of how IEDs impact ongoing physiology.

**Implications for iEEG researchers**

Our work has important implications for researchers who use iEEG data to understand basic features of human brains that are independent of epilepsy and cannot be answered with other methods (Parvizi and Kastner (2017)). There exist multiple methods for discounting the impact of IEDs on this type of research, including removing time points with IEDs, removing channels with IEDs, or including IED data. Here, we observe statistically significant increases in functional connectivity in a sample of individuals an order of magnitude larger than a typical iEEG study. The population-level effects extended beyond the contacts that showed IEDs. Additionally, some individuals showed large changes in functional connectivity associated with IEDs. Given that iEEG research is well-suited to within-individual experimental designs, we recommend based on these results that researchers remove entire time points that contain IEDs, rather than simply the channels that contain frequent IEDs. This approach will likely make results obtained from this population more consistent with the noninvasive EEG and fMRI literature in non-epileptic populations, and will aid inferences made about generalized processes not linked to epilepsy.

**Unexplained variance in connections and individuals**

Here, we test a few simple hypotheses about which connections might be contributing to increases in functional connectivity. Specifically, we test if increases in functional connectiv-
ity are preferentially explained by (1) the weakest edges in the distribution, (2) connections in the seizure onset zone, and (3) between regions that contain IEDs. While all of these measures show some significant effects, especially for the spread of IEDs, none of them recreate the patterns observed in our initial analysis for changes in functional connectivity. Interictal connectivity changes from healthy controls show complex patterns of increases and decreases across systems (Widjaja et al. (2013); Liao et al. (2010)). Additionally, these changes are temporally dynamic, but culminate in broad increases in connectivity during seizure initiation (Burns et al. (2014)). Given that IEDs are often considered to be a simple instantiation of epileptic activity that leads to seizures (Prince and Connors (1986)), the specific connections contributing to increases likely vary depending on the specific seizure phenotype of the individual.

Given the large variability in effect sizes across individuals, we also sought to characterize which features of individual participants explained some of the observed variance. The most consistently included variable was the cognitive system that most IEDs originated from. The system from which IEDs originate might be capturing information about the specific type of epilepsy, i.e. frontal or temporal lobe, that each participant has. While we did not have access to this information directly, we expect these different phenotypes to show different patterns, due to their diverse spatial distributions of pathology and connectivity changes from controls (Widjaja et al. (2013); Liao et al. (2010). Additionally, theoretical work on the spread of perturbations in brain networks suggests different directions of travel and patterns for perturbations originating in different regions (Mišić et al. (2015); Vázquez-Rodríguez et al. (2020)). It is likely that both clinical and generalizable neural features are contributing to the explanatory power of the cognitive system where the IED originated.

While it is unsurprising that no single variable captured individual variability well, it is worth considering that some band-measure combinations were not well-explained by any combination of variables. However, the cohort used in this study was not selected to be homogeneous, and contains many differences that are not captured by our variables. For
example, we include no information about the shape or frequency of IEDs, treatment plans while in the epilepsy monitoring unit, or the type of epilepsy. A meta-analysis of smaller studies with this information available might better be able to address individual variability in observed effects.

**Limitations**

This study presented a thorough characterization of how global functional connectivity changes during IEDs and simplifies the space of possible biophysical interactions that are affected during an IED. However, the findings from this study should be interpreted in light of limitations in our approach and methods. First, we sought to characterize changes to the macroscale behavior of the system, quantified as global connectivity. However, this approach prevents us from identifying even robust local changes in connections that counteract the shift in mean connection weight. Second, we chose to investigate a window size of 1 second, but it is possible that other time windows would better capture changes across different frequency bands. Third, we used a permutation-based linear model to estimate effect sizes, which precludes us from identifying nonlinear effects associated with any of our predictors. Fourth, we tested for changes in the magnitude of functional connectivity metrics, but this approach creates inconsistency for metrics that use negative values (AR model, amplitude envelope correlation) by preventing the identification of changes that shift from large positive to large negative values.

One important final consideration is the potential for filtering artifacts to influence our results (de Cheveigné and Nelken (2019)). Filters, like those used when bandpass filtering data, can induce spurious oscillations and connectivity when applied to sharp transient activity or steps. For transparency, we show the results of applying the filters used in our preprocessing steps in Fig. S3.1. We are confident that our results are not due purely to this type of artifact for the following reasons: (1) the lack of strong changes in phase-locking value associated with IEDs, (2) the lack of difference between changes to functional connectivity between contacts with IED present or absent (Fig S3.4), and (3) the significant
differences seen in functional connectivity from only contacts with no IEDs. However, it is still possible that these artifacts impact our findings in small ways.

**Future Directions**

In addition to future studies investigating other aspects of neural activity, this exploratory work facilitates several avenues of hypothesis-driven research. While we were explicitly looking for large effects that could be visible after stringent multiple comparisons correction, specific hypotheses about more homogeneous groups that are subsets of this population could elucidate more subtle changes. For example, frontal versus temporal lobe epilepsy cases have been show to have changes to functional connectivity in different systems at rest (Widjaja et al. (2013); Liao et al. (2010)). Future work could test whether the presence or spread of IEDs impacts interactions between those systems selectively, or whether changes are only seen in resting-state activity. Additionally, while work has already shown a relationship between the presence of IEDs and task performance, an extension of this work to task data could assess whether that relationship extends to other features of IEDs such as their number and spread. Lastly, specific etiologies can have very homogeneous neural abnormalities. It would be interesting to investigate whether deviations from the average profile presented here are localized to regions of pathology.
Supplemental Figures

Figure S3.1: **Filter responses.** (A) The result of notch filtering a single spike with the same parameters used in the main text. (B,C) The same as panel A, but for the bandpass and low-pass filters used in the main text.
Figure S3.2: **Differences in groups of connections for the number of IEDs.** *(Top)*
Distributions of coefficients for changes in the skew of connections associated with the number of IEDs. Obtained from permutation-based regression including all IED predictors, power in a given band, and the recording session. Columns indicate different frequency bands, and colors indicate different measures. Bright red asterisks indicate significant distributions after multiple comparisons correction that were reproduced with a different spike detector. * = \( p < 0.05 \), ** = \( p < 0.01 \), *** = \( p < 0.001 \), **** = \( p < 0.0001 \). *(Bottom)*
The same as the top, but for the coefficients for connections outside the seizure onset zone subtracted from within the seizure onset zone.

Figure S3.3: **Strength distributions for a single subject.** The distribution of strength values across connections and time points for the first participant in the dataset.
Figure S3.4: **Differences in effect sizes between contacts with IEDs present and absent.** Distributions of coefficients for the presence (top), number (middle), and spread (bottom) of an IED. Columns indicate different frequency bands, and colors indicate different measures. Grey asterisks indicate significant distributions before multiple comparisons correction. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$. 
CHAPTER 4: Learning in brain-computer interface control evidenced by joint decomposition of brain and behavior

Abstract

Objective: Motor imagery-based brain-computer interfaces (BCIs) use an individual’s ability to volitionally modulate localized brain activity, often as a therapy for motor dysfunction or to probe causal relations between brain activity and behavior. However, many individuals cannot learn to successfully modulate their brain activity, greatly limiting the efficacy of BCI for therapy and for basic scientific inquiry. Formal experiments designed to probe the nature of BCI learning have offered initial evidence that coherent activity across spatially distributed and functionally diverse cognitive systems is a hallmark of individuals who can successfully learn to control the BCI. However, little is known about how these distributed networks interact through time to support learning.

Approach: Here, we address this gap in knowledge by constructing and applying a multimodal network approach to decipher brain-behavior relations in motor imagery-based brain-computer interface learning using magnetoencephalography. Specifically, we employ a minimally constrained matrix decomposition method – non-negative matrix factorization – to simultaneously identify regularized, covarying subgraphs of functional connectivity, to assess their similarity to task performance, and to detect their time-varying expression.

Main Results: We find that learning is marked by diffuse brain-behavior relations: good learners displayed many subgraphs whose temporal expression tracked performance. Individuals also displayed marked variation in the spatial properties of subgraphs such as the connectivity between the frontal lobe and the rest of the brain, and in the temporal properties of subgraphs such as the stage of learning at which they reached maximum expression. From these observations, we posit a conceptual model in which certain subgraphs support learning by modulating brain activity in sensors near regions important for sustaining attention. To test this model, we use tools that stipulate regional dynamics on a networked system (network control theory), and find that good learners display a single subgraph whose temporal expression tracked performance and whose architecture supports easy modulation of sensors located near brain regions important for attention.

Significance: The nature of our contribution to the neuroscience of BCI learning is there-
fore both computational and theoretical; we first use a minimally-constrained, individual specific method of identifying mesoscale structure in dynamic brain activity to show how global connectivity and interactions between distributed networks supports BCI learning, and then we use a formal network model of control to lend theoretical support to the hypothesis that these identified subgraphs are well suited to modulate attention.
Introduction

Both human and non-human animals can learn to volitionally modulate diverse aspects of their neural activity from the spiking of single neurons to the coherent activity of brain regions (Sitaram et al. (2017); Fetz (1969); Steiner et al. (2014)). Such neural modulation is made possible by routing empirical measurements of the user’s neural activity to a screen or other external display device that they can directly observe (Sitaram et al. (2017); Graimann et al. (2010); Moxon and Foffani (2015)). Referred to as a brain-computer interface (BCI), this technology can be used not only to control these external devices, but also to causally probe the nature of specific cognitive processes (Bassett and Sporns (2017); Reiner et al. (2014); Okazaki et al. (2015)), and offers great promise in the treatment of neural dysfunction (Young et al. (2014); Renton et al. (2015); Lofthouse et al. (2012)). However, translating that promise into a reality has proven difficult (Thibault et al. (2016); Hamedi et al. (2016); Ahn and Jun (2015)) due to the extensive training that is required and due to the fact that some individuals who undergo extensive training will only achieve moderate control (Curran and Stokes (2003); Jeunet et al. (2016); Moxon and Foffani (2015)). A better understanding of the neural processes supporting BCI learning is an important first step towards the development of BCI therapies and the identification of specific individuals who are good candidates for treatment (Curran and Stokes (2003); Jeunet et al. (2016)).

While BCIs vary widely in their nature, we focus on the common motor imagery based BCIs where subjects are instructed to imagine a particular movement to modulate activity in motor cortex. Performance on motor imagery based BCIs has been associated with a diverse array of neural features, demographic factors, and behavioral measures (Hammer et al. (2014); Kleih and Kübler (2015); Bamdadian et al. (2014); Jeunet et al. (2016); Guillot et al. (2008)). Neural features predicting performance are frequently identified in areas associated with either performing or imagining action; for example, better performance is associated with higher pre-task activity in supplementary motor areas (Halder et al. (2011)) and larger grey matter volume in somatomotor regions (Halder et al. (2011)). Interestingly, performance has also been predicted by activity in a diverse range of other cognitive sys-
tems relevant for sustained attention, perhaps due to the high cognitive demands associated with BCI learning (Jeunet et al. (2016)). Specifically, better performance is associated with greater parietal power suppression in the $\alpha$ band, midline power suppression in the $\beta$ band, and frontal and occipital activation with motor power suppression in the $\gamma$ band (Bamdadian et al. (2014); Grosse-Wentrup et al. (2011); Frey et al. (2013)). The role of sustained attention in BCI control is corroborated by the fact that personality and self-report measures of attention predict successful learning (Hammer et al. (2012)). The heterogeneity of predictors suggests the possibility that individual differences in the interactions between cognitive systems necessary for action, action planning, and attention might explain the idiosyncratic nature of BCI control, although these interactions are challenging to quantify (De Vico Fallani and Bassett (2018); Bassett and Sporns (2017)).

Assessing the interactions between cognitive systems has historically been rather daunting, in part due to the lack of a common mathematical language in which to frame relevant hypotheses and formalize appropriate computational approaches. With the recent emergence and development of network science (Newman (2010)), and its application to neural systems (Bullmore and Sporns (2009)), many efforts have begun to link features of brain networks to BCI learning specifically and to other types of learning more generally. In this formal modeling approach (Bassett et al. (2018b)), network nodes represent brain regions or sensors and network edges represent statistical relations or so-called functional connections between regional time series (De Vico Fallani et al. (2014)). Recent studies have demonstrated that patterns of functional connections can provide clearer explanations of the learning process than activation alone (Bassett et al. (2015)), and changes in those functional connections can track changes in behavior (Bassett and Mattar (2017)). During BCI tasks, functional connectivity reportedly increases within supplementary and primary motor areas (Hamedi et al. (2016)) and decreases between motor and higher-order association areas as performance becomes more automatic (Corsi et al. (2018)). Data-driven methods to detect putative cognitive systems as modules in functional brain networks have been used to demonstrate that a particularly clear neural marker of learning is reconfiguration.
of the network’s functional modules (Kambhati et al. (2018c); Li et al. (2019)). Better performance is accompanied by flexible switching of brain regions between distinct modules as task demands change (Bassett et al. (2010b); Ramos-Nuñez et al. (2017); Gerraty et al. (2018)).

While powerful, such methods for cognitive system detection are built upon an assumption that limits their conceptual relevance for the study of BCI learning. Specifically, they enforce the constraint that a brain region may only affiliate with one module at a time (Kambhati et al. (2018b)), in spite of the fact that many regions, comprised of heterogeneous neural populations, might participate in multiple neural processes. To address this limitation, recent efforts have begun to employ so-called soft-partitioning methods that detect coherent patterns in mesoscale neural activity and connectivity (Kambhati et al. (2018b); Chai et al.; Dipasquale et al. (2015); Leonardi et al. (2013)). Common examples of such methods are independent component analysis and principal component analysis, which impose pragmatic but not biological constraints on the orthogonality or independence of partitions. An appealing alternative is non-negative matrix factorization (NMF), which achieves a soft partition by decomposing the data into the small set of sparse, overlapping, time-varying subgraphs that can best reconstruct the original data with no requirement of orthogonality or independence (Lee and Seung (1999)). Previous applications of this method to neuroimaging data have demonstrated that the detected subgraphs can provide a description of time varying mesoscale activity that complements descriptions provided by more traditional approaches (Kambhati et al. (2018b)). For example, some subgraphs identified with NMF during the resting state have similar spatial distributions to those found with typical module detection methods, while others span between modules (Kambhati et al. (2018b)). As a minimally constrained method for obtaining a soft partition of neural activity, NMF is a promising candidate for revealing the time-varying neural networks that support BCI learning.

Here, we investigate the properties of dynamic functional connectivity supporting BCI learn-
ing. In individuals trained to control a BCI, we use a wavelet decomposition to calculate single trial phase-based connectivity in magnetoencephalography (MEG) data in three frequency bands with stereotyped behavior during motor imagery: $\alpha$ (7-14 Hz), $\beta$ (15-30 Hz), and $\gamma$ (31-45 Hz) (Fig. 4.1, step 1). We construct multimodal brain-behavior time series of dynamic functional connectivity and performance, or configuration matrix (Fig. 4.1, step 2 and 3), and apply NMF to those time series to obtain a soft partition into additive subgraphs (Lee and Seung (1999)) (Fig. 4.1, step 4). We determine the degree to which a subgraph tracks performance by defining the performance loading as the similarity between each subgraph’s temporal expression and the time course of task accuracy (Fig. 4.1, step 5). We first identify subgraphs whose performance loading predicted the rate of learning and then we explore the spatial and temporal properties of subgraphs to identify common features across participants. We hypothesize that subgraphs predicting learning do so by being structured and situated in such a way as to easily modulate patterns of activity that support sustained attention, an important component of successful BCI control (Jeunet et al. (2016)). After demonstrating the suitability of this approach for our data (Fig. S4.1A-B), we test this hypothesis by capitalizing on recently developed tools in network control theory, which allowed us to operationalize the network’s ability to activate sensors located near regions involved in sustained attention as the energy required for network control (Gu et al. (2017)). Collectively, our efforts provide a network-level description of neural correlates of BCI performance and learning rate, and a formal network control model that explains those descriptions.

Methods

Participants

Written informed consent was obtained from twenty healthy, right-handed subjects (aged 27.45 ± 4.01 years; 12 male), who participated in the study conducted in Paris, France. Subjects were enrolled in a longitudinal electroencephalography (EEG) based BCI training with simultaneous MEG recording over four sessions, spanning 2 weeks. All subjects were BCI-naive and none presented with medical or psychological disorders. The study was
Figure 4.1: **Schematic of non-negative matrix factorization.**

1. MEG data recorded from 102 gradiometers is segmented into windows \((t_1, t_2, t_3, t_4, \ldots t_n)\) that each correspond to the feedback portion of a single BCI trial.
2. A Morlet *wavelet decomposition* is used to separate the signal into \(\alpha (7-14 \text{ Hz}), \beta (15-30 \text{ Hz}), \) and \(\gamma (31-45 \text{ Hz})\) components.
3. In each window, and for each band, functional connectivity is estimated as the *weighted phase-locking index* between sensor time series. Only one band is shown for simplicity. The subject’s performance on each trial is also recorded.
4. The lower diagonal of each trial (highlighted in grey in panel (3)) is reshaped into a vector, and vectors from all trials are concatenated to form a single *configuration matrix*. The subject’s time-varying performance forms an additional row in this configuration matrix. This matrix corresponds to \(A\) in the NMF cost function.
5. The NMF algorithm decomposes the configuration matrix (composed of neural and behavioral data) into \(m\) subgraphs with a *performance loading* (where \(m\) is a free parameter), with three types of information: (i) the weight of each edge in each subgraph, also referred to as the *connection loading* (viridis color scale), (ii) the performance loading (purple color scale) and (iii) the time varying expression of each subgraph (black line graphs). The performance loading indicates how similar the time-varying performance is to each subgraph’s expression. The connections and performance loadings together comprise \(W\) in the NMF cost function, and the temporal expression comprises \(H\).
6. Across bands and subjects, we then group subgraphs by their ranked performance loading for further analysis.
approved by the ethical committee CPP-IDF-VI of Paris.

**BCI task**

Subjects were seated in a magnetically shielded room, at a distance of 90 cm from the display screen. Subjects’ arms were placed on arm rests to facilitate stability. BCI control features including EEG electrode and frequency were selected in a calibration phase at the beginning of each session, by instructing the subjects to perform motor imagery without any visual feedback.

The BCI task consisted of a standard 1 dimensional, two-target box task (Wolpaw et al. (2003)) in which the subjects modulated their EEG measured $\alpha$ [8-12 Hz] and/or $\beta$ [14-29 Hz] activity over the left motor cortex to control the vertical position of a cursor moving with constant velocity from the left side of the screen to the right side of the screen. The specific sensor and frequency selected to control the BCI were based on brain activity recorded during a calibration phase before each day of recording. Here, subjects were instructed to perform the BCI task, but received no visual feedback; specifically, the target was present on the screen, but there was no ball moving towards the target. Each subject completed 5 consecutive runs of 32 trials each for the calibration phase. The EEG features (sensor and frequency) with the largest R-squared values for discriminating motor imagery conditions from rest conditions were used in the subsequent task.

Both cursor and target were presented using the software BCI 2000 (Schalk et al. (2004)). To hit the target-up, the subjects performed a sustained motor imagery of their right-hand grasping and to hit the target-down they remained at rest. Some subjects reported that they imagined grasping objects while others reported that they simply imagined clenching their hand to make a fist. Each trial lasted 7 s and consisted of a 1 s inter-stimulus interval, followed by 2 s of target presentation, 3 s of feedback, and 1 s of result presentation (Fig. 4.2A). If the subject successfully reached the target, the target would change from grey to yellow during the 1 s result section. Otherwise it would remain grey. The feedback portion was the only part of the trial where subjects could observe the effects of their volitional
modulation of motor region activity. Specifically, the subjects saw the vertical position of the cursor change based on their neural activity, as it moved towards the screen at a fixed velocity. Brain activity was updated every 28 ms. In the present study, we therefore restricted our analysis to the feedback portion of the motor imagery task because we were interested in the neural dynamics associated with learning to volitionally regulate brain activity rather than in the neural dynamics occurring at rest.

Subjects completed 4 sessions of this BCI task, where each session took place on a different day within two weeks. Each session consisted of 6 runs of 32 trials each. Each trial had either a target in the upper quadrant of the screen, indicating increased motor imagery was needed to reach it, or a target in lower quadrant of the screen, indicating no change in activity was needed to reach it. Only signals from the motor imagery trials were analyzed. This left us with, before trial rejection due to artifacts, 16 motor imagery trials × 6 runs × 4 sessions, or 384 trials per subject. Each trial was 7 seconds in duration, leading to 3 minute long runs. Combined with the training phase, each session was 1-1.5 hours total.

Neurophysiological Recordings Data

Recording

MEG and EEG data were simultaneously recorded with an Elekta Neuromag TRIUX machine (MEG) and a 74 EEG-channel system (EEG). While EEG and MEG data were recorded simultaneously, only MEG were analyzed because they are less spatially smeared than EEG signals, and therefore more appropriate for network analyses (Cuffin and Cohen (1979)). Signals were originally sampled at 1000 Hz. We also recorded electromyogram (EMG) signals from the left and right arm of subjects, electrooculograms, and electrocardiograms. EMG activity was manually inspected to ensure that subjects were not moving their forearms during the recording sessions. If subjects did move their arms, those trials were rejected from further analyses.
Preprocessing

As a preliminary step, temporal Signal Space Separation (tSSS) was performed using MaxFilter (Elekta Neuromag) to remove environmental noise from MEG activity. All signals were downsampled to 250 Hz and segmented into trials. ICA was used to remove blink and heartbeat artifacts. An FFT of the data from each subject was inspected for line noise, although none was found in the frequency bands studied here. We note that the frequency of the line noise (50 Hz) was outside of our frequency bands of interest. In the present study, we restricted our analyses to gradiometer sensors. Gradiometers sample from a smaller area than magnetometers, which is important for ensuring a separability of nodes by network models (Butts (2009)). Furthermore, gradiometers are typically less susceptible to noise than magnetometers (Garcés et al. (2017)). We combined data from 204 planar gradiometers in the voltage domain using the ‘sum’ method from Fieldtrip’s ft_combine_planar() function, resulting in 102 gradiometers (http://www.fieldtriptoolbox.org/).

Connectivity Analysis

To estimate phase-based connectivity, we calculated the weighted phase-locking index (wPLI) (Vinck et al. (2011)). The wPLI is an estimate of the extent to which one signal consistently leads or lags another, weighted by the imaginary component of the cross-spectrum of the two signals. Using phase leads or lags allows us to take zero phase lag signals induced by volume conduction and to reduce their contribution to the connectivity estimate, thereby ensuring that estimates of coupling are not artificially inflated (Vinck et al. (2011)). By weighting the metric by the imaginary component of the cross spectrum, we enhance robustness to noise (Vinck et al. (2011)). Formally, the wPLI between two time series $x$ and $y$ is given by

$$
\phi(x, y) = \frac{|E\{\text{imag}(\Gamma_{xy})\}|}{E\{|\text{imag}(\Gamma_{xy})|\}}, \tag{4.1}
$$

where $E\{}$ denotes the expected value across estimates (here, centered at different samples), $\Gamma_{xy}$ denotes the cross spectrum between signals $x$ and $y$, and $\text{imag}()$ selects the imaginary
We first segment MEG data from gradiometers into 3 second trials, sampled at 250 Hz. The cross spectrum is then estimated using wavelet coherence (Lachaux et al. (2002)) in each of three frequency bands of interest (α 7-14 Hz, β 15-30 Hz, and γ 31-45 Hz), with wavelets centered on each timepoint. We chose to compute the wavelet coherence – rather than Welch’s method – it does not assume stationarity of the signal (Lachaux et al. (2002)). We implemented the procedure in the Fieldtrip package in MATLAB, with a packet width of 6 cycles and zero-padding up to the next power of two (‘nextpow2’). We then calculate the wPLI as the mean of the imaginary component of the cross spectrum, divided by the imaginary component of the mean of the cross spectrum.

We then construct a network model of these statistical relationships where sensors (N = 102) are nodes, and the weight of the edge between node i and node j is given by the weighted phase-locking value. The graph, G, composed of these nodes and edges is a weighted, undirected graph that is encoded in an adjacency matrix A. By constructing this network model, we can use statistics from graph theory and computational approaches from control theory to quantify the structure of inter-sensor functional relations (Bassett et al. (2018b); Bassett and Sporns (2017)).

Uniformly Phase Randomized Null Model

In order to ensure that our results are not due to choices in preprocessing, the time invariant cross-correlation of neural signals, or the autocorrelation of neural signals, we repeated all of the preprocessing and analysis steps with a uniformly phase randomized null model (Heitmann and Breakspear (2018)). To enhance the simplicity and brevity of the exposition, we will also sometimes refer to this construct simply as the null model. Surrogate data time series from the null model were calculated using a custom function in MATLAB. Essentially, the FFT of the raw data is taken, the same random phase offset is added to every channel, and then the inverse FFT is taken to return the signal to the time domain (Theiler). Mathematically, this process is achieved by taking the discrete Fourier transform of a time
series $y_v$:

$$Y(u) = \sum_{v=0}^{V-1} y_v e^{2\pi uv/V},$$

(4.2)

where $V$ is the length of the time series, $v$ indexes time, and $u$ indexes frequencies. We then multiply the Fourier transform by phases chosen uniformly at random before transforming back to the time domain:

$$y_v = \frac{1}{\sqrt{V}} \sum_{v=0}^{V-1} e^{ia_u} |Y(u)| e^{-i2\pi kv/V},$$

(4.3)

where the phase $a_t \in [0, 2\pi)$.

**Construction of a Multimodal Configuration Matrix**

In this work, we wished to use a data-driven matrix decomposition technique to identify time-varying subgraphs of functional connectivity that support learning. Specifically, for each subject and each frequency band, we created a multimodal configuration matrix of edge weights and BCI performance over time, prior to submitting this matrix to a decomposition algorithm that we describe in more detail below (Fig. 4.1, step 4). We made separate matrices for each frequency band rather than concatenating them into a single matrix because it is easier for the NMF algorithm to converge if there are more time points relative to the number of edges. To construct the matrix, we first vectorize the upper triangle (not including the diagonal) of each trial’s connectivity matrix, and then we concatenate all of the vectors and our one performance measure into an $E \times \tau$ matrix, where $\tau$ is the number of trials (384, if no trials were removed), and $E$ is the number of edges (5151) plus the number of behavioral measures (1). This concatenation process results in a $5152 \times 384$ multimodal (brain-behavior) matrix. In this task, each subject’s performance is recorded as their percentage of successful trials (out of 32) on each run. This measure includes both motor imagery trials, where the target was located in the upper quadrant of the screen, and
rest trials where the target was located in the lower quadrant of the screen. Because this measure was averaged over trials but the connectivity was calculated on individual trials, we interpolate the performance time series to obtain a graded estimate of the percentage of correct trials that is $\tau$ time points long. The performance vector is then normalized to have the same mean as the other rows of the configuration matrix.

**Non-negative Matrix Factorization**

We used a data-driven matrix decomposition method – non-negative matrix factorization (NMF) – to identify time-varying groups of neural interactions and behavior during BCI learning (Lee and Seung (1999)). Intuitively, NMF decomposes a matrix into a set of additive subgraphs with time-varying expression such that a linear combination of these subgraphs weighted by temporal expression will recreate the original matrix with minimal reconstruction error (Lee and Seung (1999); Khambhati et al. (2018b)). The NMF algorithm can also be thought of as a basis decomposition of the original matrix, where the subgraphs are a basis set and the temporal coefficients are basis weights. Unlike other graph clustering methods (Newman and Girvan (2003)), NMF creates a soft partition of the original network, allowing single edges to be a part of multiple subgraphs. Additionally, unlike other basis decomposition methods (Bartholomew (2010); Comon et al. (2015)), NMF does not impose harsh constraints of orthogonality, or independence of the subgraphs; it simply finds the most accurate partition, given that the original matrix is non-negative. In many systems (including those whose edges reflect phase-locking), the non-negativity constraint is not difficult to satisfy; moreover, this constraint is particularly relevant to the study of physical systems, where the presence of a negative edge weight can be difficult to interpret.

Formally, the NMF algorithm will approximate an $E \times T$ configuration matrix $\hat{A}$ by the multiplication of two matrices: $W$, the subgraph matrix with dimensions $E \times m$, and $H$, with dimensions $m \times T$. The matrices $A$, $W$, and $H$ are shown in Fig. 4.1, steps 4 and 5. Here, $E$ is the number of time varying processes (behavior and functional connections derived from MEG data), $T$ is the number of time points, and $m$ is the number of subgraphs. Details of how we solve for $W$ and $H$, as well as parameter selection can be found in the
Supplemental Materials.

Subgraph Inclusion

Most subgraphs are sparse, with distributions of temporal coefficients skewed towards zero (see Fig. S4.4). However, for every subject and every frequency band, one subgraph showed very little regularization (no edges were equal to 0) and had a uniform, rather than skewed distribution of temporal coefficients. These subgraphs are clear outliers from the others, and appear to be capturing global phase-locking across the entire brain, rather than any unique subsystem. To answer this question about the time varying interactions between neural systems, we were particularly interested in differences between the subgraphs that were spatially localized, having edges regularized to zero. Because including these outlier subgraphs would obscure those differences, we removed these subgraphs from all further analyses.

Group Average Subgraphs

After applying NMF to the multimodal brain-behavior matrix, we next turned to a study of the nature of the detected subgraphs after ranking them by performance loading. Specifically, we were initially interested in determining which edges contributed to each ranked subgraph most consistently across the population. For this purpose, we used a consistency based approach to create a group representative subgraph for each ranked subgraph (Roberts et al. (2017)). In this procedure, each subject’s subgraph was first thresholded to retain only the 25% strongest connections (see Fig. S4.5 for evidence that results are robust to variations in this choice). We then constructed an average $N \times N$ subgraph $G$, where $N$ is the number of channels and where each element $G_{ij}$ quantifies how many subjects (out of 20) displayed an edge between region $i$ and region $j$ in their thresholded subgraph. In addition to visually depicting these group representative subgraphs, we also wished to summarize their content in spatial bins. It is important to note that without source reconstruction, meaningful inference about which anatomical regions correspond to which sensors is extremely difficult (Palva et al. (2018)). We therefore binned edges into 10 anatomically defined areas using montages obtained from BrainStorm (Tadel et al. (2011)).
Figure 4.2: **BCI task and performance.** (A) Schematic of the BCI task. First the target, a grey bar in the upper or lower portion of the screen, was displayed for 1 s. Next, the subjects have a 3 s feedback period, where the vertical position of the cursor is determined by their neural activity while it moves horizontally at a fixed velocity. This portion corresponds to the analysis window, indicated with a grey bar in the figure. The result is then displayed for 1 s. If the subject reached the target, it will turn yellow; otherwise it will remain grey. There is a 1 s intertrial interval (ITI) between trials where nothing is displayed on the screen. This sequence is repeated 32 times per run, with 6 runs per session. (B) Each subject’s average performance across four days within two weeks. BCI Score is the percentage of correct trials during that session.
software (neuroimage.usc.edu/brainstorm/Tutorials/MontageEditor). For parsimony, and acknowledging the limits of anatomical inference from sensor data, we refer to each of these bins as a different lobe (frontal, motor, parietal, occipital, and temporal) in a given hemisphere (Fig. S4.9).

**Optimal Control**

Our final broad goal was to provide a theoretical explanation for why certain networks support BCI learning. We hypothesized that these regularized networks might have structures that make it easier for the brain to modulate the patterns of activity that are necessary for BCI control. This hypothesis motivated us to formulate and validate a model to explain how the sparse statistical relationships characteristic of each subgraph could support the production of brain activity patterns implicated in BCI learning (Gu et al. (2017); Betzel et al. (2016b)). Additionally, this model should account for the brain’s ability to reach these patterns of activity in the context of the BCI task, where there is increased volitional modulation of the left motor cortex. Here, we use tools from network control theory to satisfy these conditions (Pasqualetti et al. (2014a)). Specifically, we characterize the theoretical brain activity at each sensor as a vector \( x(t) \), and we use the adjacency matrix \( A \) of a subgraph to quantify the ease with which that activity can affect other regions. We then incorporate volitional input control as input into the brain \( u(t) \) at a specific region (given by \( B \)). Then, by stipulating

\[
\dot{x}(t) = Ax(t) + Bu(t),
\]

we model the linear spread of activity along the connections in \( A \) in the context of input to regions given in \( B \). We note that these dynamics are simple, and we do not expect them to fully capture the richness of observed signals; nevertheless, simple models have the notable advantages of interpretability and flexibility.

With this model of network dynamics, optimal control trajectories can be formalized and
identified by developing a cost function that seeks to minimize two terms: (i) the distance of
the current state from the target state and (ii) the energy required for control. Specifically,
we solve the following minimization problem:

\[
\min_{u} \int_{0}^{T} (x_T - x(t))^T (x_T - x(t)) + \rho u_{\kappa}(t)^T u_{\kappa} \, dt,
\]

\[
\text{s.t. \quad } \dot{x} = Ax(t) + Bu(t), \quad x(0) = x_0, \quad \text{and} \quad x(T) = x_T,
\]

where \( \rho \) is a free parameter that weights the input constraint, \( x_T \) is the target state, and
\( T \) is the control horizon, which is a free parameter that defines the finite amount of time
given to reach the target state. During BCI control, there is specific, targeted control to a
specific area of the brain (here, the left motor cortex) in addition to other ongoing control
and sensory processes. We wished for our selection of the input matrix \( B \) to reflect this
richness and also allow for computationally tractable calculations of optimal control, which is
difficult for sparse control sets. Therefore, we constructed the input matrix \( B \) so as to allow
input that was dominated by the BCI control site, while maintaining minor contributions
from other areas. More specifically, rather than being characterized by binary state values,
channels other than the one located over left motor cortex were given the smallest non-zeros
value that assured low error calculations, approximately \( 5 \times 10^{-5} \) at their corresponding
diagonal entry in \( B \). See Supplement for the full derivation from (Gu et al. (2017)).

It is important to note that in general the tools from linear controllability theory are not
applicable to the functional networks commonly derived from neuroimaging data for two
reasons. The first reason is that the model which the tools are built upon stipulates a
time-dependent propagation of activity along edges; such a propagation is physically true
for structural connections derived from white matter, but is not generally true for other
types of connections used in network models, such as morphometric similarity or most
common functional connectivity measures. The second reason is that the model assumes
that interactions between nodes ‘a’ and ‘c’ are not due to node ‘b’, an assumption that is violated by measures of statistical similarity such as the Pearson correlation coefficient which is the measure of functional connectivity most commonly employed in neuroimaging studies. Because we are using neither structural connectivity nor common measures of functional connectivity, it was necessary for us to first prove that the networks we are studying are consistent with our model. To address the first point regarding the propagation of activity along edges, we demonstrate that the structure of the subgraphs used have utility in predicting empirical brain state transitions, and that the relative contribution of each subgraph is related to its temporal expression (Fig. S4.1C-D). It is only in light of these validations that we are able to interpret our results as a potential model for driving brain activity. To address the second point regarding isolation of pairwise relations not due to third party effects, we note that the matrix $A$ that we study reflects statistical similarity in phase after strict regularization that removes redundant statistical relationships (Fig. S4.1A-B).

**Target state definition**

A central hypothesis in this work is that certain regularized subgraphs are better suited to drive the brain to patterns of activity that are beneficial for BCI control than others. To test this hypothesis, we create target states that reflect these beneficial patterns, based on previous literature. Target states for motor imagery and attention are obtained for each band individually from references (Bamdadian et al. (2014); Grosse-Wentrup et al. (2011); Frey et al. (2013)), and can be briefly described as follows: $\alpha$ contralateral motor suppression for motor imagery and parietal suppression for attention, $\beta$ contralateral motor suppression and ipsilateral motor activation for motor imagery and vertex suppression for attention, and $\gamma$ contralateral motor activation for motor imagery and motor cortex suppression with frontal and occipital activation for attention (Fig. S4.10). While acknowledging the limits of anatomical inference from sensor data, we sought to approximate these true functional systems at the sensor level by dividing channels into lobes using standard montages provided by Brainstorm (Tadel et al. (2011) software (neuroim-
age.usc.edu/brainstorm/Tutorials/MontageEditor). The target state of channels in brain regions where we did not have specific hypotheses for their activity were set to zero; the target state of channels with activation were set to 1 and that of channels with deactivation were set to -1. Initial states were set to 0 for all channels. We then calculate the optimal energy (using the optimal control equation described above) required to reach each of these target states to test the hypothesis that subgraphs that support learning will have lower energy requirements than those that do not.

Statistical Analyses

Much of our analyses involve testing differences in distributions across subjects for different subgraphs or sessions, both for phase-randomized and empirical data. We also compare these distributions to subject learning rate defined as the slope of performance over time. For the results displayed in Fig. 4.2 here in the main manuscript, we used a repeated measures ANOVA to test for the presence of a main effect across conditions given that the distributions of performances were normal (see Fig. S4.11). In Fig. 4.3 here in the main manuscript, we sought to associate learning rate with ranked performance loading. After plotting quantile-quantile plots (see Fig. S4.12-S4.14) for the learning rate, and each of the performance loadings, it became clear that the lowest loadings were not normally distributed. Therefore, we used a linear model combined with non-parametric testing utilizing 5000 permutations (lmPerm package in R https://cran.r-project.org/web/packages/lmPerm). Standardized coefficients were calculated using the lm.beta package in R (https://cran.r-project.org/web/packages/lm.beta/lm.beta.pdf). We use a Bonferroni correction to control false positive errors due to multiple comparisons across all 6 predictors ($\alpha = 0.008$). To obtain an estimate of how sensitive our results are to our specific sample, we also plot summary statistics from 500 models obtained from bootstrapping a sample of equal size ($N = 60$, 3 band and 20 subjects). To examine differences in consistency (Fig. 4.4 here in the main manuscript), we use a linear model ($\text{consistency} \sim \text{band} + \text{dataType} + \text{rank}$) to test for a main effect of data type (null or empirical), band, and subgraph on consistency (see Fig. S4.15). We next sought to determine
if different subgraphs had consistently different temporal expression for null and empirical
data (Fig. 4.5 here in the main manuscript). We also used a repeated measures ANOVA
to test for a main effect of subgraph across bands, and paired t-tests to test for differences
amongst individual subgraphs (Fig. S4.16). Lastly, for the results shown in Fig. 4.6 here
in the main manuscript, we test the relationship between learning rate and optimal control
energy differences for several different models. Pearson’s correlations were used, given that
the data appears normally distributed and has few outliers (see Fig. S4.17-S4.20).

Data and Code

Code for analyses unique to this manuscript are available at github.com/jastiso/netBCI.
Code for the NMF algorithm and the NMF parameter selection is available at
github.com/akhambhati/Echobase/tree/master/Echobase/Network/Partitioning/Subgraph.
Code for optimal control analyses is available at github.com/jastiso/NetworkControl. Data
necessary to reproduce each figure will be made available upon request.

Results

BCI Learning Performance

Broadly, our goal was to examine the properties of dynamic functional connectivity during
BCI learning, and to offer a theoretical explanation for why a certain pattern of connec-
tivity would support individual differences in learning performance. We hypothesized that
decomposing dynamic functional connectivity into additive $N \times N$ subgraphs would reveal
unique networks that are well suited to drive the brain to patterns of activity associated
with successful BCI control. We use MEG data from 20 healthy adult individuals who
learned to control a motor-imagery based BCI over four separate sessions spanning a two
week period. Consistent with prior reports of this experiment (Corsi et al. (2018)), we
find a significant improvement in performance across the four sessions (one-way ANOVA
$F(3,57) = 13.8, p = 6.8^{-7}$) (Fig. 4.2). At the conclusion of training, subjects reached a
mean performance of 68%, which is above chance (approximately 55 - 60%) level for this
task (Müller-Putz et al. (2008)).
Dynamic patterns of functional connectivity supporting performance

To better understand the neural basis of learning performance, we detected and studied the accompanying patterns of dynamic functional connectivity. First, we calculated single trial phase-based connectivity in MEG data in three frequency bands: \( \alpha \) (7-14 Hz), \( \beta \) (15-25 Hz), and \( \gamma \) (30-45 Hz). We then used non-negative matrix factorization (NMF) – a matrix decomposition method – to separate the time-varying functional connectivity into a soft partition of additive subgraphs. We found that the selected parameters led to an average of 7.4 subgraphs, with a range of 6 to 9, and that all frequency bands had a decomposition error lower than 0.47 (mean \( \alpha \) error = 0.352, mean \( \beta \) error = 0.379, mean \( \gamma \) error = 0.465) (Fig. S4.2). The error is the Frobenius norm of the squared difference between our observed and estimated connectivity matrices (with dimensions 5152 × 384) and takes values between 0 and 1. For each band, the error value is low, giving us confidence that we have fairly accurately reconstructed relevant neural dynamics. To determine whether any properties of the identified subgraphs were trivially due to preprocessing choices, NMF parameters, or time-invariant autocorrelation in neural activity, we repeated the full decomposition process after permuting the phases of all time series uniformly at random. We found that the statistics of subgraph number and decomposition error were similar for the uniformly phase randomized data, indicating that any differences in subgraph and temporal expression between null and empirical data is not due to the NMF algorithm’s inability to find a good decomposition, but rather due to the structure of the chosen decomposition (Fig. S4.2).

We quantified the similarity between each subgraph’s temporal expression and the time course of performance, and we refer to this quantity as the subgraph’s performance loading (Fig. 4.1). Here, performance is calculated as the percentage of accurate trials over a run of 32 trials. We hypothesized that the ranked performance loading would be associated with task learning, as operationalized by the slope of performance over time. It is important to note the distinction between performance and learning: performance is defined as task accuracy and therefore varies over time, while learning is defined as the linear rate of change in that performance over the course of the experiment (384 trials over 4 days).
We tested whether learning was correlated with the performance loading of subgraphs. Because the minimum number of subgraphs in a given subject was 6, we decided to investigate the top four highest performance loading subgraphs, and the smallest and second smallest nonzero loading subgraphs. We found a general trend that the performance loading from high loading subgraphs was negatively associated with learning rate, and the performance loading from low loading subgraphs was positively associated with learning rate (Fig. 4.3AB). We assessed the statistical significance of these trends and found that only the third highest loading subgraph displayed a performance loading that was significantly correlated with learning rate after Bonferroni correction for multiple comparisons (linear model with permutation tests $slope \sim loading3 + band : p = 0.005$). Performance loading from uniformly phase randomized surrogate data for this subgraph was not associated with learning rate ($p = 0.292$). The direction of the observed effect in the empirical data is notable; subjects with lower loading onto high loading subgraphs learned the task better, suggesting that learning is facilitated by a dynamic interplay between several subnetworks. It is also notable that the highest loading subgraphs do not have the strongest associations with learning, indicating that the subgraphs that most closely track performance are not the same as the subgraphs that track changes in performance.

**Spatial properties of dynamic patterns of functional connectivity**

Next we sought to better understand why the third highest loading subgraph was most robustly associated with learning. We hypothesized that because of this subgraph’s strong association across subjects, it might recruit sensors near consistent brain regions and reflect the involvement of specific cognitive systems across subjects. To evaluate this hypothesis, we began by investigating the shared spatial properties of this subgraph in comparison to the others. To identify shared spatial features we grouped subgraphs together by their ranked performance loading, and then quantified how consistent edges were across participants (Roberts et al. (2017)) (see Methods). We found that the average consistency varied by frequency band, and differed between the empirical and surrogate data, but not across ranked subgraphs (linear model $consistency \sim band + rank + data : F_{band}(2,17) = 90.36$,
Figure 4.3: **Performance loading is associated with learning.** *(A)* Here we show the $p$-values for empirical (green) and uniformly phase randomized (grey) data for linear models relating the slope of performance with ranked performance loading from each frequency band. The black line corresponds to $p = 0.05$, while the red dashed line corresponds to the Bonferroni corrected $\alpha = 0.008$. Error bars show the standard error and median of $p$-values from 500 models with bootstrapped samples. *(B)* The standardized regression coefficients for the same models. Error bars show the standard error and mean of coefficients from 500 models with bootstrapped samples.
\( p_{\text{band}} = 9.00 \times 10^{-10}, F_{\text{data}}(1,17) = 41.8, p_{\text{data}} = 5.78 \times 10^{-6} \). The \( \alpha \) band had the most consistent edges, followed by the \( \gamma \) band, and then the \( \beta \) band \( (t_{\alpha \beta} = -12.68, p_{\alpha \beta} = 4.3 \times 10^{-10}, t_{\alpha \gamma} = -10.41, p_{\alpha \gamma} = 1.2 \times 10^{-8}) \). In the uniformly phase randomized surrogate data, we observed less consistent subgraphs than those observed in the empirical data \( (t = -6.47, p = 5.78 \times 10^{-6}) \). These observations support the conclusion that across the population, despite heterogeneous performance, similar regions interact to support performance and learning to varying degrees.

In order to approximate system-level activation with sensor level data, we used lobe montages provided by Brainstorm (see Methods). Spatially, subgraphs were dominated by connectivity in the frontal lobe sensors, with subtle differences in the pattern of connections from the frontal lobe sensors to sensors located in other areas of the brain (Fig. 4.4). To determine which functional edges were most consistent in each subgraph and frequency band, we calculated the average consistency over each lobe and motor cortex in both hemispheres (for the same analysis in surrogate data, see Fig. S4.6). In the \( \alpha \) band, the most consistent edges on average were located in the left frontal lobe in the highest performance loading subgraph, in the left occipital lobe in the second highest performance loading subgraph, between right frontal and right motor in the third highest performance loading subgraph, and between left frontal lobe and right parietal lobe in the lowest performance loading subgraph. In the \( \beta \) band, the most consistent edges were located between right and left frontal lobe for the highest and second highest performance loading subgraph, between left frontal lobe and right motor for the third highest performance loading subgraph, and between left and right frontal lobe for the lowest performance loading subgraph. In the \( \gamma \) band, the most consistent edges were located in the left frontal and right frontal lobes for the highest performance loading subgraph, in the left frontal lobe and right motor for the second highest performance loading subgraph, and in left frontal and right frontal lobe for the third highest and lowest performance loading subgraphs. We wished to demonstrate that the consistent involvement of more frontal sensors across subgraphs was not due to the presence of electro-oculogram (EOG) artifacts that persisted after removal of eye blinks.
with ICA. We therefore calculated the weighted phase-locking index between both vertical and horizontal EOG sensors and all neural sensors. Qualitatively, we did not observe any consistently strong connectivity between EOG channels and more frontal sensors, indicating that the frontal connectivity identified in our analysis is likely not due to residual artifacts from eye movements (Fig. S4.7). We also note that the most consistent individual edges for each subgraph are still only present in 10-12 individuals, indicating a high amount of individual variability. Collectively, these observations suggest widespread individual variability in the spatial composition of ranked subgraphs, with the most consistent connectivity being located in the frontal lobe during BCI learning.

**Temporal properties of dynamic patterns of functional connectivity**

Importantly, subgraphs can be characterized not only by their spatial properties, but also by their temporal expression. We therefore next examined the temporal properties of each subgraph to better understand why the third highest performance loading subgraph was most robustly associated with learning. As a summary marker of temporal expression, we calculated the total energy of the time series operationalized as the sum of squared values, as well as the time of the peak value of the time series. Across frequency bands, we found no significant dependence between energy and subgraph ranking. We did find a significant effect of rank for the peak time of temporal expression obtained from the empirical data (repeated measures ANOVA peak ∼ rank + band: $F_{rank}(3, 215) = 6.67, p_{rank} = 2.53 \times 10^{-4}$) but not from the uniformly phase randomized surrogate data ($F_{rank}(3, 215) = 1.28, p = 0.282$).

Overall, peak times are widely distributed across individuals. However we find that across bands, the highest performance loading subgraph has a later peak, which is intuitive since performance is generally increasing over time and these subgraphs most strongly track performance.

We then performed post-hoc paired $t$-tests corrected for multiple comparisons (Bonferroni correction $\alpha = 0.006$) between the highest performance loading subgraph and all other ranked subgraphs in each band. In the $\alpha$ band, the highest performance loading subgraph only peaked significantly later than the lowest (paired $t$-test $N = 20, t_{low} = 8.06, p_{low} =$
Figure 4.4: **Spatial distribution of subgraph edges that are consistent across participants.** Consistent edges for each frequency band and for each ranked subgraph. Left images show individual edges plotted on a topographical map of the brain. Right images show the mean edge weight over sensors for a given region. We studied 10 regions, including the frontal lobe, temporal lobe, parietal lobe, occipital lobe, and motor cortex in both hemispheres. The weight of the edge corresponds to the number of individual participants for whom the edge was among the 25% strongest for that subgraph.
1.49 × 10⁻⁷) after Bonferroni correction (α = 0.006). In the β band, the highest performance loading subgraph peaked significantly later than all others (paired t-test N = 20, t₂H = 10.9, p₂H = 1.39 × 10⁻⁹; t₃H = 7.56, p₃H = 3.57 × 10⁻⁷; tₗow = 8.07, pₗow = 1.49⁻¹⁰⁻⁷). In the γ band, the highest performance loading subgraph peaked significantly later than the second highest, and lowest loading subgraphs (paired t-test N = 20, t₂H = 4.50, p₂H = 2.46 × 10⁻⁴; tₗow = 8.06, pₗow = 1.49 × 10⁻⁷). (Fig. 4.5). Finally, we asked whether the time of the peak in the third highest performance loading subgraph was associated with learning. We did not find a relationship between peak time and learning in any frequency band (Pearson’s correlation: α : r = 0.005, p = 0.98, β : r = 0.047, p = 0.84, γ : r = −0.21, p = 0.037). To summarize these findings, we note that across participants and especially in the β band, subgraphs that support performance are highly expressed late in learning, when performance tends to be highest. However, subgraphs that support learning do not have consistent peaks across subjects, and each individual’s peak does not relate to their learning rate, indicating that some other feature of these subgraphs must explain their role in learning.

Explaining dynamic patterns of functional connectivity supporting BCI learning via network control theory

Lastly we asked how the third highest loading subgraph could facilitate successful BCI performance, as shown in Fig. 4.3. Here, we considered an edge – extracted under penalties of spatial and temporal sparsity – as a potential path for a brain region to affect a change in the activity of another brain region (Weigand et al. (2018); Ferreri et al. (2014)). Assuming the true connectivity structure is sparse, the regularization applied in the NMF algorithm can remove large statistical relationships between regions that are not directly connected, but might receive common input from a third region (Das et al. (2017)) (see Methods for addition discussion, and see Fig. S4.1A-B for the effect of regularization on the prevalence of triangles). We hypothesized that the pattern of edges in this subgraph would facilitate brain states, or patterns of activity, that were predictive of BCI literacy. Specifically, we expected that when the brain mirrored the connectivity of the third subgraph, the brain could more easily reach states of sustained motor imagery or sustained attention than when
Figure 4.5: **Temporal expression of ranked subgraphs.** The peak temporal expression for every subject (black data point), for each frequency band (indicated by color) and for each subgraph (ordered vertically). Violin plots show the density distribution of all subjects' peaks. The median is marked with a solid line through the violin plot.
the brain mirrored the connectivity of the lowest performance loading subgraph. To operationalize these hypotheses from sensor level data, we identified sensors near motor and attention areas with montages from Brainstorm and set those as targets (see Methods). We also hypothesized that the magnitude of this difference would be associated with each subject’s learning rate. To test these hypotheses, we used mathematical models from network control theory to quantitatively estimate the ease with which the brain can reach a desired pattern of activity given a pattern of connectivity (see Methods and Fig. S4.1C-D for analyses demonstrating the efficacy of the regularized subgraphs in linearly predicting changes in activity). Specifically we calculated the optimal control energy required to reach a target state (either sustained motor imagery or sustained attention) from an initial state when input is applied primarily to the left motor cortex, which was the site of BCI control (Fig. 64.A-B).

We tested whether the third highest performance loading subgraph supported the transition to states of sustained motor imagery or sustained attention with smaller energy requirements than other subgraphs that did not support learning in the same way. We chose the lowest performance loading subgraph for comparison because it was the only subgraph with a large positive standardized regression coefficient for fitting learning, which contrasts sharply with the large negative coefficient for the third subgraph. For both states (motor imagery and attention), we found no population level differences in energy requirements by the two subgraphs (paired $t$-test $N = 20$, motor imagery: $t_\alpha = -0.005$, $p_\alpha = 0.565$, $t_\beta = 1.38$, $p_\beta = 0.184$, $t_\gamma = -1.00$, $p_\gamma = 0.329$. attention: $t_\alpha = -1.35$, $p_\alpha = 0.193$, $t_\beta = -0.344$, $p_\beta = 0.735$, $t_\gamma = -0.937$, $p_\gamma = 0.360$). We next tested whether the magnitude of the difference in energy required by the two subgraphs to reach a given state tracked with learning rate. In the $\beta$ band, we observed a significant correlation between the magnitude of the energy difference to reach attentional states and learning rate over subjects (Pearson’s correlation coefficient $r = 0.560$, $p = 0.0103$, Bonferroni corrected for multiple comparisons across frequency bands; Fig. 4.6). Notably, the relationship remained significant when controlling for subgraph density (linear model $slope \sim energy\_difference + density\_difference$: $t_{energy} = 2.68$, $p_{energy} = 0.0103$, Bonferroni corrected for multiple comparisons across frequency bands; Fig. 4.6).
\( p_{\text{energy}} = 0.0158, t_{\text{density}} = -0.266, p_{\text{density}} = 0.794 \). When using subgraphs derived from the uniformly phase randomized surrogate data, the relationship was not observed (Pearson’s correlation \( r = -0.0568, p = 0.819 \)). We next asked which subgraph contributed most to this effect. We found no significant relationship between learning rate and the energy required to reach the attentional state by the third highest performance loading subgraph (Pearson’s correlation \( r = -0.389, p = 0.702 \)) or by the lowest performance loading subgraph (Pearson’s correlation \( r = 0.227, p = 0.335 \)). This finding suggests that learning rate depends on the relative differences between subgraphs, rather than the energy conserving architecture of one alone. As a final test of specificity, we assessed whether this difference was selective to the third highest and lowest performance loading subgraph. We found no significant relationship when testing the difference of the highest with the third highest performance loading subgraph (Pearson’s correlation \( r = -0.554, p = 0.586 \)), the highest with the lowest performance loading subgraph (Pearson’s correlation \( r = 0.40, p = 0.077 \)), the second highest with the third highest performance loading subgraph (Pearson’s correlation \( r = 0.266, p = 0.257 \)), or the second highest with the lowest performance loading subgraph (Pearson’s correlation \( r = -0.072, p = 0.764 \)). This pattern of null results underscores the specificity of our finding.

Reliability and specificity of inferences from network control theory

Collectively, our findings are consistent with the hypothesis that during BCI learning, one subnetwork of neural activity arises, separates from other ongoing processes, and facilitates sustained attention. An alternative hypothesis is that our results are due to trivial factors related to the magnitude of the attentional state, or could have just as easily been found if we had placed input to a randomly chosen region of the brain, rather than to the left motor cortex which was the actual site of the BCI control. To determine whether these less interesting factors could explain our results, we performed the same network control calculation but with a spatially non-overlapping target state, and then – in a separate simulation – with a mirrored input region (right motor cortex rather than left motor cortex). We performed the spatial shifting by ordering the nodes anatomically (to preserve spatial contiguity), and
Figure 4.6: **Separation of the ability to modulate attention is associated with learning.** Different patterns of connections will facilitate transitions to different patterns of brain activity. We hypothesize that the ease with which connections in certain regularized subgraphs facilitate transitions to patterns of activity that support either motor imagery (A) or attention (B) will be associated with learning rate. We use network control theory to test this hypothesis. We model how much energy ($u(t)$) is required to navigate through state space from some initial pattern of activity $x(0)$ to a final pattern of activity $x(T)$. Some networks (e.g., the brown network in panel A) will require very little energy (schematized here with a smaller, solid colored arrow) to reach patterns that support motor imagery, while other networks (e.g., the pink network in panel B) will have small energy requirement to reach patterns of activity that support attention. (C) The relationship between learning rate and the difference in energy required to reach the attention state when the underlying network takes the form of the lowest versus third highest performance loading subgraphs for empirical data (green) and uniformly phase randomized surrogate data (grey). (D) The relationship between the learning rate and the energy required to reach the attention state when the underlying network takes the form of the lowest performance loading subgraph, or when the underlying network takes the form of the third highest performance loading subgraph.
then circular shifting the attention target state by a random number between 1 and \( N - 1 \). For 500 circularly shifted states, only 3 (0.6\%) had a correlation value equal to or stronger than the one observed (Fig. S4.8). Furthermore, we found no significant relationship between learning rate and the difference in energy required by the two subgraphs to reach the true attention state when input was applied to the right motor cortex instead of the left motor cortex (Pearson’s correlation \( t = 0.711, p = 0.313 \)). Together, these two findings suggest that the relationship identified is specific to BCI control.

Finally, we assessed the robustness of our results to choices in modeling parameters. First we performed the computational modeling with two different sets of control parameter values (see Supplement). In both cases, the significant relationship remained between learning rate and the difference in energy required by the two subgraphs to reach the attentional state (set one Pearson’s correlation coefficient \( r = 0.476, p = 0.0338; \) set two Pearson’s correlation coefficient \( r = 0.514, p = 0.0204 \)). Second, since our target states were defined from prior literature, there was some flexibility in stipulating features of those states. To ensure that our results were not unduly influenced by these choices, we tested whether ideologically similar states would provide similar results. Namely, we assessed (i) the impact of varying the magnitude of (de)activation by changing \((-1)\) to \((-2)\), (ii) the impact of the neutral state by changing 0 to 1, and (iii) the impact of negative states by changing -1, 0 and 1 to 1, 2, and 3. We found a consistent relationship between learning rate and the difference in energy required by the two subgraphs to reach the attentional state when we changed the magnitude of activation/deactivation (Pearson’s correlation coefficient \( r = 0.560, p = 0.0103 \)), as well as when we changed the neutral state (Pearson’s correlation coefficient \( r = 0.520, p = 0.0188 \)). However, we found no significant relationship when removing negative states (Pearson’s correlation coefficient \( r = 0.350, p = 0.130 \)), indicating that this result is dependent on our choice to operationalize deactivation as a negative state value. After performing these robustness checks, we conclude that a selective separation of the third highest and lowest performance loading subgraphs impacts their ability to drive the brain to patterns of sustained attention in the \( \beta \) band in the context of BCI control.
This result is robust to most of our parameter choices, is selective for biologically observed states, and is not observed in surrogate data.

Discussion

In this work, we use a minimally constrained decomposition of dynamic functional connectivity during BCI learning to investigate which groups of phase locked brain regions (subgraphs) support BCI control. The performance loading onto these subgraphs favors the theory that dynamic involvement of several subgraphs during learning supports successful control, rather than extremely strong expression of a single subgraph. Additionally, we find a unique association for the third highest loading subgraph with learning at the population level. This result shows that learning is not simply explained by the subset of edges that has the most similar temporal expression to behavior, but rather that a subnetwork with a middling range of similarity has the strongest relationship with performance improvement. While the spatiotemporal distribution of this subgraph was variable across individuals, we did observe some consistencies at the group level. Spatially, the third highest loading subgraph showed strong edges between left frontal and right motor cortices for low frequencies, and left frontal and left motor cortices for the γ band. Lower frequencies showed stronger connectivity to the ipsilateral (to imagined movement) motor cortex, suggesting a possible role in suppression for selective control. This subgraph also showed the highest expression earlier than the other ranked subgraphs we investigated, perhaps linking it to the transition from volitional to automatic control.

We next wished to posit a theory of how these subgraphs fit with previously identified neural processes important for learning, despite their heterogeneity across subjects. After quantifying the extent to which NMF regularization removed potentially redundant relationships between regions (Fig. S4.1A-B), we suggested that the regularized pattern of statistical relationships identified in this subgraph could comprise an avenue through which brain activity could be modulated via cognitive control or external input. We then hypothesized that these networks would be better suited to modulate activity in either regions implicated in attention or in motor imagery than other subgraphs, and further that indi-
viduals whose networks better modulated activity in these regions would display greater task learning (Jeunet et al. (2016)). We chose to operationalize the “ease of modulation” with a metric from network control theory called optimal control energy. Optimal control energy quantifies the minimum input needed to drive the brain from an initial pattern of activity to a final pattern of activity, while also assuring that the pattern of activity stays close to the target state at every point in time. This last constraint ensures that we are unlikely to pass through biologically unfeasible patterns of activity. The notion of optimal control energy that we use here assumes a particular linear model of how neural dynamics change given potential avenues of communication between regions. Importantly, in the supplement (Fig. S4.1C-D) we show that our subgraphs predict empirical brain state changes according to this model, and that the contribution of each subgraph to empirical changes in brain state is related to its temporal expression. Using this model, we did not find any population differences in optimal control energy when the simulation was enacted on the third highest performance loading subgraph compared to the lowest performance loading subgraph. However, we did find that the magnitude of this difference was associated with learning in individual subjects. This result was specific to the $\beta$ band and to brain regions implicated in attention. Critically, the relation to learning could not be explained by the energy of either subgraph alone, was not present in surrogate data derived from a uniformly phase randomize null model, and was robust to parameter choices. Overall, the observations support our hypothesis that in the $\beta$ band the subgraphs we identified that support learning are well suited to modulate activity in brain regions associated with attention.

A delicate balance of interactions is required for BCI learning

Our initial analysis explored the relationship between performance loading and learning. It is important to note the behavioral difference between performance and learning: we use the term performance to refer to task accuracy over time, whereas we use the term learning to refer to how well a subject is able to increase that accuracy. With that distinction in mind, we aimed to better understand how subgraphs that vary similarly to performance (those with high performance loading) relate to learning. We found that the subgraph with the
third highest performance loading was most strongly associated with learning and that a narrow distribution of performance loading across all subgraphs was associated with better learning. Together, these two observations are in line with previous research in motor and spatial learning, which shows that some brain structures display differential activity during learning that is independent of performance (Shelton and Gabrieli (2004); Purushotham et al. (2002)). Our work adds to this literature by demonstrating that in addition to targeted differences in individual brain regions or networks, a minimally constrained decomposition of dynamic functional connectivity across the whole brain reveals that separable processes are most associated with performance and with learning.

Additionally, we find that BCI learning is not explained simply by the processes most strongly associated with performance and learning individually, but by a distributed loading across many different subgraphs. This notion is supported by the sign of beta value for ranked subgraphs. Generally, subgraphs with higher ranked loading were negative betas, while subgraphs with lower ranked loading were positive betas. A wealth of whole brain connectivity analyses have similarly shown that the interaction between systems is an important component of skill learning specifically, and other domains of learning more generally (Bassett et al. (2015); Altman and Krzywinski (2017)). While we observed marked interactions between many regions, the majority were located in the frontal lobe for all frequency bands. Even for $\alpha$ and $\beta$ frequencies in the highest loading subgraph, we see involvement of frontal regions and heterogeneity across individuals. This suggests that the NMF method did not extract a network that was trivially related to the deterministic mapping between brain activity and cursor location determined by the BCI2000 software. Previous work has also demonstrated changes in frontal-motor (Karim et al. (2017)) and fronto-parietal (Lin et al. (2012)) connectivity during motor skill learning. In BCI learning specifically, the strength of white matter connectivity between frontal and occipital regions predicts control of motor imagery based BCIs (Sitaram et al. (2013)). Additionally, analyses of this same experiment have shown task-related changes in functional connectivity were spatially diffuse, and found in frontal, temporal, and occipital regions in the $\alpha$ band (Corsi et al.
Our results add to these findings by demonstrating that the most consistent regions that covary in their functional connectivity are interactions between the frontal lobe and other regions. Our work shows that broad motifs like the dynamic integration of multiple systems (including cognitive systems involving the frontal lobe) found in other types of learning are also important for BCI learning. Additionally, we add to previous work on BCI learning specifically by quantifying the structure of covarying subgraphs of connectivity.

**BCI learning is heterogenous across individuals**

We find population level consistencies in spatial and temporal properties of ranked subgraphs despite having no constraint to assure consistency across individuals. However, we also note that there is a high degree of variability in both of these measures. The variability is mirrored in the subjects’ performance, with final performances varying from 38.1 % to 89.3 %. Our observations are in line with previous literature demonstrating variability in subjects’ performance and learning for psychological, cognitive, and neurological predictors (Jeunet et al. (2016); Halder et al. (2010)). Such pervasive and marked individual differences present a challenge for the use of BCIs clinically (Brunner et al. (2010)). To address this challenge, researchers have explored ways to optimize BCI features and algorithms for neurofeedback itself (Vidaurre et al. (2011); Kubler et al. (2015)) and to identify selection criteria for BCI based therapies (Jeunet et al. (2015); Halder et al. (2010)). The results of our study support the idea that different individuals will have slightly different neural correlates of both performance and learning based on a variety of features such as demographics (Schumacher et al. (2015)), spatial manipulation skills (Vuckovic and Osuagwu (2013)), relationship with the technology (Brosnan (1998)), and attention span (Grosse-Wentrup et al. (2011); Grosse-Wentrup and Schölkopf (2012)). Our findings also highlight the importance of studying models fit to each individual when searching for selection criteria for BCI therapies. Here, despite temporal and edge level heterogeneity, our minimally constrained, individual specific method of brain connectivity decomposition revealed a robust association with learning with a theoretical role that aligns well with previous literature.
Further development and expansion of this model to incorporate resting state neuroimaging data and other physiological predictors could be a promising direction for selection of candidates for BCI therapies before training.

**Role of beta oscillations in BCI learning**

Prominent theories describing the neural processes that give rise to cognition and shape our behavior often involve integration of complex multimodal information using a combination of top-down predictions (built from prior experience) and bottom-up, sensory-driven representations of the dynamic world around us (Talsma (2015); Kok et al. (2012); Sussman et al. (2016)). These generalized frameworks, in turn, require the precise coordination of ensemble neural activity both within and between brain regions. Several theoretical approaches have examined how these two scales of functional activity may harmonize to produce the desired behavior (Riddle et al. (2019)), and empirical research has shown that there is consistent cross-talk between these scales (Richter et al. (2017)). Within human neuroimaging work, synchronous oscillations have been critical to understanding this complex coordination, where cortico-cortical propagation delays and membrane potentials give rise to observed oscillatory activity in the brain (Bastos et al. (2012); Singer et al. (2001)). Here, we study the time varying connectivity within $\alpha$, $\beta$, and $\gamma$ bands. Much like how specialized functions arise from different brain regions, different narrowband oscillations have been implicated in diverse but specialized processes, where some generalizable theories suggest a role for $\alpha$ in disengagement of task irrelevant areas or a lack of sensory processing (Palva and Palva (2007)), $\beta$ in sustaining the current cognitive state (Engel and Fries (2010)) and $\gamma$ in task active local cortical computation (Fries (2009)). Specifically in the context of motor imagery based BCIs, $\alpha$ and $\beta$ bands have prominent signatures in motor imagery (McFarland et al. (2000)). Our results show that only the $\beta$ band’s functional connectivity is well suited to modulate patterns of activity that support sustained attention (not motor imagery), which is a critical process for BCI control. While our results are in line with generalized theories on the role of oscillations in cognition, the specificity of the $\beta$ band in our results extends classic studies that discuss the role of this oscillation in attention (Pfurtscheller A’b’ et al.
and in maintaining the current cognitive state (Engel and Fries (2010)). Our results suggest that this maintenance, a consistent control (or attention to) internally generated activity, may play a crucial role in longterm BCI use.

Methodological Considerations

NMF Non-negative matrix factorization is a machine learning technique for separating, in our case, a multimodal configuration matrix into a soft-partition of subgraphs with time-varying expression. This process has several advantages, such as being able to link behavioral and neural data, and creating a quantification of mesoscale structure where brain regions can participate in multiple functional groups. Nevertheless, the method also faces several limitations that are common to other large-scale machine learning techniques. NMF yields a low rank approximation of a large configuration matrix, and can sometimes be rank deficient for large number of subgraphs, for very large datasets, or for datasets with high covariance. Because of this sensitivity, we were not able to test our data against independently phase randomized null models.

MEG Functional Connectivity We chose to complete our analyses in sensor, rather than source space. Ultimately, this choice was motivated by the fact that if any of our findings could be applicable to clinicians monitoring learning during real-time BCI learning they would need to be obtained in the sensor space. However, this choice has two major methodological consequences: (1) it limits the anatomical resolution of our data, and therefore the specificity of the claims that we can make about the spatial distribution of the regions involved and (2) it does not protect as well against false positive connectivity estimates (Palva et al. (2018); Zumer et al. (2008)). We were not interested in the finer anatomical resolution of the identified subgraphs, but more in the process of identifying them, in validating the hypothesis that features of these subgraphs are associated with learning, and in their theoretical functions. We used montages provided by Brainstorm to approximate lobes and systems at the sensor level; however, we acknowledge that even claims made about specific systems (motor, and attention) at the source level are best interpreted in
light of controls. Our use of spatial permutation tests is thus particularly important, because they demonstrate that similar contiguous states do not show the same relationship between energy and learning. Additionally, we cannot fully eliminate the possibility that parts of our data are due to false positive interaction from signal spread, and our conclusions should be interpreted in light of this fact. That being said, we have taken several steps to reduce the influence of false positives in our connectivity estimates. First, we use a connectivity estimate that does not include zero-phase lag contributions that could arise from signal spread (Vinck et al. (2011)). However, removing zero-phase lag contributions on its own is not enough to prevent against false positive from source spread from true connections (Palva et al. (2018)). While source reconstruction partially addresses this problem, it does not eliminate it entirely (Palva et al. (2018)), and it additionally requires many parameter choices and has potentially confounding effects on estimates of functional connectivity (Brookes et al. (2011); Hillebrand et al. (2012); Colclough et al. (2016)). Secondly, all results of interest are compared to a phase-randomized null model with the same static covariance structure as the original data, which should lessen the effect of spurious connectivity estimates.

Optimal Control We chose to use tools from network control theory to quantify the ease with which each network can modulate brain activity. Network control theory relies on several assumptions that should be considered when interpreting these results (Tu et al.). First, the model of dynamics that we employ is linear and noise free, unlike the brain (Gu et al. (2015a)), but has proven useful in gaining intuitions about the behavior of nonlinear systems (Muldoon et al. (2016); Honey et al. (2007)). However, we still sought to quantify the ability of this linear model to explain empirical changes in brain state. Specifically, we asked two questions: (1) do the regularized subgraphs used in our analyses have the ability to predict state transitions, and do they do so better than randomly rewired networks, and (2) is the contribution of each subgraph to explaining a given state transition proportional to its temporal expression, and is it more proportional than a different subgraph’s temporal expression? To evaluate these questions, we generated brain states for every trial (band
specific power at each channel) and simulated Eq. 4.5 (see Ch. 4 Supplement). Regarding the similarity of predicted and empirical state transitions, we find modest correlation values (mean Pearson’s $r = 0.25$) that are significantly greater than the correlations observed from randomized networks. Similarly for our second question, we found small but positive correlations between the contribution of each subgraph to a given transition and its temporal expression (mean Pearson’s $r = 0.03$), which was also significantly greater than correlations to temporal expression from mismatched subgraphs. While it is unsurprising that our linear model did not fully capture neural dynamics across a three second trial, it is worth considering extensions that can maximize this similarity for future analyses investigating how connections between regions facilitate changes to activity. One option is to use effective connectivity (Liu and Aviyente (2012); Neumaier and Schneider (2001)) – that solve for a network of connections that best predicts the evolution of brain states in time. However, effective connectivity matrices are often sparse, and therefore not well suited to the NMF matrix decomposition used in the present work. Alternatively, one could use non-linear models of dynamics (Jirsa and Haken (1996)) and non-linear control theory (Zanudo et al. (2017)) to capture a wider range of dynamic behaviors, although non-linear control does not currently support the same scope of tools available for linear control theory. Lastly, future work could use functional approximation (Brunton et al. (2016)) in order to identify a set of simple basis functions that well approximate the data. If a sparse approximation can be found, it supports the idea that the underlying non-linear dynamics can be captured with linear combinations of these basis functions, and therefore are suitable to be modeled with simplified linear models.

Additionally, network control is typically applied to time invariant, structural connections that have a clear role as an avenue along which brain activity can propagate. Here we used functional connectivity (weighted phase locking) which is a statistical relationship that (1) does not imply the presence of a physical connection and (2) is not time invariant. Due to (1), our original functional connectivity matrix can have large values between two regions that are not directly connected, but might both connect to the same region. This situation
would lead to a triangle composed of three connections in a functional connectivity matrix where in reality there are only two connections. However, the regularization applied by the NMF algorithm mitigates this concern in a manner that is similar to the regularization applied in effective connectivity metrics (Liu and Aviyente (2012); Das et al. (2017)). We also explicitly quantify the effect of regularization on triangles in our subgraphs and find a dramatic reduction from the original functional connectivity (Fig. S4.1A-B). This quantification, along with the two validations discussed above, show that our model is a suitable way to evaluate the role of regularized subgraphs in modulating different patterns of activity. In relation to (2), we note that functional connectivity in not time-invariant, unlike the state matrix more commonly employed in linear control models. However, it is important to note that NMF identifies subgraphs that are separable from their temporal expression, and that we expect that the hypothesized role in control would only be prominent when the subgraph was highly expressed.

Conclusion and Future Directions

Future research that builds on this work could explore ways to increase sensitivity to an individual’s learning rate. Given that EEG and MEG sensors capture some unique information (Lopes Da Silva et al. (1991)) and provide increased discriminability in clinical applications including BCIs (Corsi et al. (2019); Chowdhury et al. (2015)), it would be interesting to investigate whether the concurrently collected EEG data in this study better captures relevant neural dynamics for performance and learning, respectively. Such an effort, combined with source reconstruction, would be a useful next step in basic scientific inquiries directed towards characterizing these separable networks involved in learning. However, combining EEG and MEG sources would greatly increase the number of variables relative to the number of observations in the connection matrix to be decomposed, and would make the NMF algorithm less likely to converge. It may thus be necessary to use connectivity estimates from smaller time windows. Clinical utility could potentially be achieved if similar methods could be applied to resting state data to identify network properties that separate individuals by their learning rate, thereby eliminating the need for any BCI training. Finally,
confirmatory studies with a larger sample of individuals would both validate the current results, and provide a better assessment of potential clinical utility.

In conclusion, we use a minimally constrained method of matrix decomposition that is specific to each human participant to investigate the dynamic neural networks that support BCI learning. We find that the subgraphs that most tightly mirror performance are not the same subgraphs that most strongly support learning. Additionally, we find that the interaction between many different neural processes is important for BCI learning. While the subgraphs identified are heterogeneous (as is subject performance), we find consistent involvement of frontal and motor cortices in subgraphs that support learning. We also observe differential temporal expression amongst subgraphs, and perhaps most notably that the subgraphs that vary more similarly with performance reach their highest expression later in learning. Lastly, we test the hypothesis that subgraphs that support learning are better suited to modulate activity in brain regions important for attention than other subgraphs. We find evidence to support this hypothesis in the β band specifically, ultimately suggesting that the separation of processes for maintaining attention is important for successful BCI learning. Our results align with prior work from dynamic functional connectivity in other types of skill learning, and also highlight a method for identifying individual predictors of successful BCI control with theoretical support.
Supporting information

Supplemental Methods

Non-negative Matrix Facroization

In the main text, we describe NMF algorithm as a method for finding a soft partition of edges into subgraphs. Here, we discuss specifically how that solution is obtained. We solve for $W$ and $H$ such that:

$$\min_{W,H} \frac{1}{2} \| \hat{A} - WH \|_F^2 + \alpha \|W\|_F^2 + \beta \sum_{t=1}^{T} \|H(:,t)\|_1^2,$$  \hspace{1cm} (4.6)

where $\beta$ is the penalty to impose sparse basis weights, and $\alpha$ is the regularization for the basis set. Regularization is frequently used in machine learning algorithms to avoid overfitting data, which is especially important when employing these techniques to examine highly variable single trial estimates of functional connectivity (Kim and Park (2011)). Additionally, selecting for sparsity will encourage the characterization of local neural processes where many edges do not contribute (Khambhati et al. (2018b)). From many such local processes arises the diversity of cognitive functions involved in complex tasks such as BCI control (Jeunet et al. (2015)).

To solve the NMF equation, we use an alternating non-negative least squares with block-pivoting method with 100 iterations for fast and efficient factorization of large matrices, where $W$ and $H$ with non-negative weights are drawn from a uniform random distribution on the interval $[0,1]$ (Kim et al. (2014)). The parameter $m$ is drawn from the range $(2,20)$, and $\alpha$ and $\beta$ are drawn from the range $(0.001,2)$. We select for parameters that will both minimize the residual error, and maximize the temporal and subgraph sparsity (Khambhati et al. (2018b)). Specifically, we select the optimal parameters $\bar{m}$, $\bar{\alpha}$, and $\bar{\beta}$ that are in the lowest 25th percentile for residual error, and the highest 25th percentile for temporal and subgraph sparsity. This procedure resulted in an average $\bar{m}$ of 7.4, an average $\bar{\alpha}$ of 0.46, and an average $\bar{\beta}$ of 0.45. Distributions of parameters and reliability across runs are shown.
in Fig. S4.2 and S4.3.

Given the non-deterministic nature of this approach, we also test for the stability of our identified clusters using a consensus clustering algorithm (Greene et al. (2008)). Our procedure was comprised of the following ordered steps: (1) run the NMF algorithm $r = 100$ times per multimodal configuration matrix, (2) concatenate the subgraph matrix $W$ across $r$ runs into an aggregate matrix with dimensions $E \times (r \times \bar{m})$, and (3) apply NMF to the aggregate matrix to determine a final set of subgraphs and expression coefficients (Khambhati et al. (2018b)). While the implementation is heuristic in nature, we found that across two runs of the algorithm, we obtain highly consistent selections for parameters (see Supplement), bolstering confidence in the robustness of the subsequent analyses.

**Optimal Control**

In the methods section of the main text, we describe the notion of optimal control trajectories in rather broad strokes. Here in the supplement, we provide a more formal description. We begin by noting that from the formulation in Eq (4.6) in the main text, we can see that the term $(x_T - x(t))^T(x_T - x(t))$ constrains the trajectories of a subset of nodes by preventing the system from traveling too far from the target state. We can also see that the term $\rho u_{\kappa}(t)^T u_{\kappa}$ constrains the amount of input used to reach the target state, which is a requirement dictated by the underlying biology including metabolic demands and tissue sensitivities.

To solve the minimization problem stated in Eq (4.6) in the main text, we follow the derivation from (Gu et al. (2017)).

**Optimal Control Energy of the System** To quantify differences in the ease of controlling the system to a certain target state, we calculated a single measure of total energy for node $i$, defined as follows:

$$E_{i0x_T} = \int_0^T u_{ix_0x_T} \, dt.$$  \hspace{1cm} (4.7)

A single measure of energy for the entire system was calculated by summing over all nodes.
in the network.

**Metric for Simulation Error**

Because optimal control is a computationally difficult problem, we also calculate the numerical error associated with each computation. The numerical error is calculated as

$$n_{err} = \left\| \begin{pmatrix} E_{12} \\ E_{22} \end{pmatrix} p^* + \begin{pmatrix} E_{11} \\ E_{21} \end{pmatrix} x^*(0) + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} x(T) \\ 0 \end{pmatrix} \right\|.$$  \quad (4.8)

We minimize this error metric when choosing values for the free parameters $\rho$ and $T$, as described in more detail below.

**Parameter Selection**

Our optimal control framework has two free parameters: $\rho$, the relative importance of the input constraint over the distance constraint, and $T$, the control horizon, or the amount of time given for the system to reach the target. Intuitively, choosing a lower value of the parameter $\rho$ corresponds to relaxing the constraint on the minimal energy, leading to larger energies but smaller errors. The final parameter $T$ determines how quickly the system is required to reach the target state. At small values of $T$, the system is difficult to control, leading to large errors and high energy requirements. At moderately large values of $T$, the system has more time to reach the target state, and simulations typically produce smaller errors. At very large values of $T$, it is difficult to calculate the matrix exponentials, and simulations typically produce large errors.

Because we lack direct biological data that would inform the choice of these parameter values, we explored a range of values for both parameters, and we chose values that minimized the numerical error of the simulation. For each parameter, we first calculated the error of the simulations for parameter values that were logarithmically spaced between $1 \times 10^{-4}$ and 1 for $\rho$ and between 0.01 and 1 for $T$. We then selected the parameters that produced minimal error. Specifically, the parameters selected were $T = 0.1$ and $\rho = 0.1$. For the
purposes of reliability and reproducibility, we also report results for two different sets of parameters and we note that these two sets also produced low error. The two additional sets used were $T = 0.2$ and $\rho = 0.01$, and $T = 0.07$, and $\rho = 1$.

Model Validation

In this work, we sought to test the hypothesis that regularized subgraphs of co-varying functional connectivity are well suited to modulate certain patterns of connectivity by positing a model of how activity would spread across connections defined by those subgraphs. It is important to note that the functional connectivity we study here reflects statistical dependencies between regions rather than causal interactions. However, since the subgraphs are regularized and were selected to maximize sparsity, they could be interpreted as possible paths of influence that are only weakly modulated by redundant relationships. Here we therefore investigated the validity of using these regularized subgraphs in our theoretical model of activity spread.

We operationalized this investigation by considering three properties of these subgraphs that would be required in order for them to be suitable for use in the control theoretic model. First, we sought to demonstrate that the regularization imposed by NMF leads to subgraphs with fewer triangles than functional connectivity matrices obtained without regularization (Das et al. (2017)). Second, our model assumes that brain activity, at least on short time scales, evolves linearly along the connections of the network. This assumption leads to the testable prediction that simulating state transitions using Eq. 4.5 in the main text would yield states that bear some resemblance to empirical states. Third and finally, we wish to validate our claim that each subgraph contributed to a different part of the observed changes in brain state, and that that contribution is larger when that subgraph’s temporal expression is higher relative to the other subgraphs.

To test the first prediction – that regularized subgraphs will have fewer triangles than the original functional connectivity matrices – we calculate the fraction of possible triangles present in the subgraphs as well as the average weight of each triangle. The first metric is
calculated from a binarized version of the subgraph, while the second metric is calculated from the weighted version of the subgraph. For each subject, we report average values of each metric over subgraphs. Since the original functional connectivity matrices (prior to decomposition with regularization) are fully weighted and fully connected, they contain the maximum number of possible triangles. In contrast, subgraphs have approximately 30% of the possible triangles (Fig. S4.1A). Moreover, we observed a drastic reduction in the weight of triangles in the subgraphs compared to that observed in the original functional connectivity matrices (Fig. S4.1B). Together, these two results indicate that the process of regularization enacted by NMF removes many of the redundant relationships present in functional connectivity matrices.

We now turn to our second prediction that simulating state transitions using Eq. 4.5 in the main text would yield states that bear some resemblance to empirical states. It is unreasonable to expect noise free, time invariant linear models to fully capture neural dynamics at this time scale. Instead, we would like to ask if the features of network topology we are interested in provide additional explanatory power than similarly simple models that do not contain these features. We operationalize this prediction in the specific context of our experiment by stating that the third highest performance loading subgraph and the lowest performance loading subgraph should explain some of the changes in brain activity according to the following model of dynamics:

\[ \dot{x}(t) = Ax(t) + Bu(t), \]  

where \( x \) is the brain state at each sensor, \( A \) is the subgraph that we wish to test, and \( Bu \) represents input to the left motor region during the BCI task.

To test our prediction, we use the above equation to obtain simulated brain states for each trial \( t \) where \( x(0) \) is the brain state at trial \( t - 1 \), and we compare those simulated states to the observed brain states. Consistent with prior work (Stiso et al. (2018)), we define a brain state as a vector of power estimates in each channel, and we obtain such
states separately for each frequency band. More specifically, for each trial we use FieldTrip (http://www.fieldtriptoolbox.org/) to implement a multtaper Fourier transform (method = ‘mtmfft’) with half taper smoothing. We then log transform the calculated power spectra, and z-score across trials and within sessions to obtain a brain state for each trial. For each trial, we then simulate the above equation with either the third highest or lowest performance loading subgraph as \( A \), the same \( B \) that was used in the main text (ones at regions in the left motor cortex, and a constant, smaller number at other regions), and \( u(t) = 1 \). The time in the simulation is defined in arbitrary units, and thus it is not clear which time point to select that would be comparable to the 3 s trial interval. Here, we select 1000 arbitrary time steps to ensure that the system’s response has stabilized. Then we select the time point at which the simulated state is most similar to the empirical brain state, where state similarity is given by the Pearson’s correlation coefficient between the simulated state and the observed state. This procedure provides a measure of the best possible prediction that our model is capable of generating. We then average maximum correlations across trials to obtain one correlation value per subject. To ensure that our results were dependent on the true network topology of the empirical subgraphs, we repeated the same process with randomized networks that preserve the edge weight distribution, number of nodes, and number of edges of the original networks. These random networks are obtained using the ‘randmio_und.m’ function with 1000 swaps per edge from the Brain Connectivity Toolbox (Rubinov and Sporns (2010)). For both the third highest (paired \( t \)-test \( t = 4.49, p = 3.67 \times 10^{-4} \) ) and lowest (paired \( t \)-test \( t = 4.84, p = 1.75 \times 10^{-4} \) ) performance loading subgraphs, correlations are significantly larger than those obtained from the null model. Together, these results indicate that simulating state transitions using Eq. 4.5 in the main text does indeed yield states that are statistically similar to empirical states, and that that similarity is greater than expected in appropriate random network null models.

Lastly, we wished to test the final prediction that each subgraph will contribute more to future brain states when its expression is relatively high. We performed simulations using the same model described above, but this time we used each subgraph individually. More
specifically, for every subgraph $C_i$, we simulate the dynamics

$$\dot{x}(t) = C_i x(t) + Bu(t). \quad (4.10)$$

Ultimately, this process gives us the predicted states for every subgraph and every trial $\bar{X}$, an $N \times m \times t$ vector where $N$ is the number of nodes, $m$ is the number of subgraphs, and $t$ is the number of trials. We can then estimate the weight of each prediction on the true state by solving for $d$ in

$$\dot{x}(t) = \bar{X}(t)d(t), \quad (4.11)$$

where $d$ is a $1 \times m$ vector. We can then test the similarity of $d$ to the temporal expression obtained from the NMF algorithm. Specifically, we z-score both $d$ and $W$ across subgraphs to scale the data, and then assess the correlation coefficient between $d_i$ and $W_i$ for each subgraph $i \in m$. We then average the correlation values over subgraphs. Lastly, we compare the observed mean correlations to those obtained from comparing $d_i$ to $W_j$, where $j$ is drawn randomly with replacement from $m$. We find that correlations for the empirical data are significantly greater than 0 (permutation test: $p = 0.038$), and significantly greater than correlations obtained from unmatched subgraphs (permutation test, $p = 0.022$) (Fig. S4.1D). Together, these results indicate that a subgraph will contribute more to future brain states when its expression is relatively high, consistent with our prediction.

Collectively, the validation of our three predictions provides empirical support for our use of NMF-derived subgraphs as adjacency matrices in the network control framework.
Supplemental Figures
Figure S4.1: **Effect of Regularization on Triangles.** (A) The fraction of present triangles over possible triangles in binarized graphs for each subgraph, and for the original functional connectivity (FC) matrices prior to NMF decomposition with regularization. (B) The average weight of triangles for each subgraph, and for the original FC matrices prior to NMF decomposition with regularization. The inset shows all subgraphs on a different scale. (C) Similarity between simulated and empirical brain states for third highest and lowest performance loading subgraphs. Each data point reflects the average of the maximum correlation reached over all trials (N = 384 if no trials were removed). Correlations from empirical subgraphs are shown in green, and correlations from randomized subgraphs are shown in grey. (D) The similarity between (i) the weighted contribution of each subgraph to predicting the next brain state and (ii) the temporal expression of that subgraph derived from NMF. Correlations from matched state prediction weights and temporal expression are shown in green, and correlations from mismatched temporal expression are shown in grey.
Figure S4.2: **NMF Parameters.** Distributions of all three NMF parameters (panel A, $\beta$; panel C, $\alpha$; panel D, $m$) and error (panel B) for empirical data ($emp$) and null data ($upr$, indicating uniformly phase randomized null model). Each band is shown in a different color: $\alpha$ band (red), $\beta$ band (grey-blue), and $\gamma$ band (cream).
Figure S4.3: **NMF Parameter Consistency.** The consistency of NMF parameters across two randomly chosen runs of the algorithm: run $i$ marked by the x-axis, and run $j$ marked by the y-axis. Each parameter is given in a different row: $\alpha$ (top row), $\beta$ (second row), rank (third row), and error (bottom row). Each band is shown in a different color: $\alpha$ band (red, left), $\beta$ band (grey-blue, middle), and $\gamma$ band (cream, right). Each data point indicates the selected parameter for a single subject.
Figure S4.4: Mean Subgraphs. Example histograms from a representative subject showing how one subgraph \((m,\text{ shown in purple})\) has drastically different distributions of \(A\) temporal expression and \(B\) coefficients than the others. These outlier subgraphs were removed from further analysis to allow us to investigate differences among the remaining more local subgraphs. Different subgraphs are plotted in different colors.
Figure S4.5: **Consistency Across Thresholds.** To quantify spatial features common to each subgraph, we calculate the average consistency of each edge, or the number of subjects that had a given edge in their X% strongest connections. Here, we plot the difference in average consistency between empirical and null data changes with the choice of different thresholds (choices of X). Notice that for any choice of threshold, the relationship between null and empirical graphs for a given band has the same sign, and that generally null data is less consistent (except for the $\gamma$ band in the second highest subgraph). Each frequency band is shown in a different color: $\alpha$ band (red), $\beta$ band (grey-blue), and $\gamma$ band (cream). Results are shown for the highest performance loading subgraph (top left), second highest performance loading subgraph (top right), third highest performance loading subgraph (bottom left), and lowest performance loading subgraph (right).
Figure S4.6: Consistency for Null Data. The consistent subgraphs identified for the uniformly phase randomized null data (*UPR*) and for the empirical data (*Empirical*). Each frequency band is shown in a different color: α band (red, two leftmost columns), β band (grey-blue, two middle columns), and γ band (cream, two rightmost columns). The color bar reflects consistency, or the number of subjects that had a given edge in their 25% strongest connections.
Figure S4.7: Connectivity to Electro-occulogram (EOG) Channels. To address the possibility that consistent involvement of more frontal sensors in subgraphs was due to residual EOG artifacts, we calculated weighted phase-locking index between vertical (Left) and horizontal (Right) EOG channels, and all other sensors using the same methods as in the main manuscript. Since the strong frontal interactions are present in almost all subgraphs, and therefore should be present at all time points, we then averaged across trials to obtain a single horizontal and vertical EOG connectivity pattern for all subjects. Similar to the main manuscript, we then wished to investigate which edges were consistent across subjects. We thresholded each connectivity vector to only contain the 25% strongest edges. The colorbar shows how many subjects retained each edge after thresholding. The top, middle, and bottom rows show the $\alpha$, $\beta$, and $\gamma$ bands respectively. Note that there is not a clear increase in connectivity to frontal sensors, unlike what is seen in the main text in Figure 4.
Figure S4.8: **Alternate target states with the same magnitude.** A histogram of the correlation between the energy difference for low and third highest subgraphs to reach a spatially shifted brain state with the same magnitude as the attention state, and the learning rate for 500 randomly shifted target states. The red line indicates the observed $p$-value.
Figure S4.9: Montages. Masks of the montages used to define regions for consistency analyses. Black dots indicate sensor locations. Yellow areas indicate regions included in the montage; dark blue areas indicate regions not included in the montage. Columns indicate hemisphere. Rows indicate lobe or swath: frontal (top), motor (second), parietal (third), temporal (fourth), and occipital (bottom).
Figure S4.10: Optimal Control Target States. Masks of the target states used for the optimal control analyses. Each state is a linear combination of the montages displayed in the previous figure. Color indicates state value. The left column shows the attention state for the $\alpha$ band (top), $\beta$ band (middle), and $\gamma$ band (bottom); the right column shows the motor state for the $\alpha$ band (top), $\beta$ band (middle), and $\gamma$ band (bottom).
Figure S4.11: **Normality of Performance.** Quantile-quantile plot for the performance in each session across subjects. Each data point indicates a subject. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
Figure S4.12: Normality of Performance Loading for Empirical Data. Quantile-quantile plots for performance loading across subjects. Each data point indicates a subject and band. Each panel displays data from a different subgraph: the highest performance loading subgraphs (top left), the second highest performance loading subgraph (top right), the third highest performance loading subgraph (middle left), the fourth highest performance loading subgraph (middle right), the lowest performance loading subgraph (bottom left), and the second lowest performance loading subgraph (bottom right). Note that some of the data does not fall near the red line, indicating a non-normal distribution. In each panel, the dotted red lines represent the 95% confidence interval.
Figure S4.13: **Normality of Performance Loading for Null Data.** Quantile-quantile plots for performance loading across subjects from the uniformly phase randomized null data. Each data point indicates a subject and band. Each panel displays data from a different subgraph: the highest performance loading subgraphs (top left), the second highest performance loading subgraph (top right), the third highest performance loading subgraph (middle left), the fourth highest performance loading subgraph (middle right), the lowest performance loading subgraph (bottom left), and the second lowest performance loading subgraph (bottom right). Note that some of the data does not fall near the red line, indicating a non-normal distribution. In each panel, the dotted red lines represent the 95% confidence interval.
Figure S4.14: Normality of Consistency Data. Quantile-quantile plot for the consistency across subjects. Each data point indicates a subject and band. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
Figure S4.15: **Normality of Peak Temporal Expression Data.** Quantile-quantile plots for the peak expression across subjects, bands, and subgraphs. Each data point indicates a subgraph for a given subject and band. Empirical data is shown on the left and uniformly phase randomized data is shown on the right. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
Figure S4.16: **Normality of Energy Difference to Attention States.** Quantile-quantile plots for the optimal energy difference in attention states across subjects, separately for the α band (left), β band (middle), and γ band (right). Each data point indicates a subject. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
Figure S4.17: Normality of Energy Difference to Motor Imagery States. Quantile-quantile plots for the optimal energy difference of motor imagery states across subjects, separately for the $\alpha$ band (left), $\beta$ band (middle), and $\gamma$ band (right). Each data point indicates a subject. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
Figure S4.18: Normality of Energy Difference for Alternate Control Set and for Null Data. Quantile-quantile plots for the optimal energy difference of null analyses across subjects. The left panel shows data for the alternative control set and the right panel shows data for the uniformly phase randomized null. Each data point indicates a subject. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
Figure S4.19: **Normality of Energy Difference for Various State Controls.** Quantile-quantile plots for the optimal energy difference of state control analyses across subjects. These analyses include a control for state magnitude (top left), a control for larger magnitude (top right), a control changing the state intervals (bottom left), and a control for the mean state (bottom right). Each data point indicates a subject. Most of the data points fall near the red line, indicating a normal distribution. Dotted red lines represent the 95% confidence interval.
CHAPTER 5: White matter network architecture guides direct electrical 
stimulation through optimal state transitions

This chapter contains work from Stiso, J., Khambhati, A. N., Menara, T., Kahn, A. E., 
Abstract

Electrical brain stimulation is currently being investigated as a potential therapy for neurological disease. However, opportunities to optimize and personalize such therapies are challenged by the fact that the beneficial impact (and potential side effects) of focal stimulation on both neighboring and distant regions is not well understood. Here, we use network control theory to build a formal model of brain network function that makes explicit predictions about how stimulation spreads through the brain’s white matter network and influences large-scale dynamics. We test these predictions using combined electrocorticography (ECoG) and diffusion weighted imaging (DWI) data from patients with medically refractory epilepsy undergoing evaluation for resective surgery, and who volunteered to participate in an extensive stimulation regimen. We posit a specific model-based manner in which white matter tracts constrain stimulation, defining its capacity to drive the brain to new states, including states associated with successful memory encoding. In a first validation of our model, we find that the true pattern of white matter tracts can be used to more accurately predict the state transitions induced by direct electrical stimulation than the artificial patterns of a topological or spatial network null model. We then use a targeted optimal control framework to solve for the optimal energy required to drive the brain to a given state. We show that, intuitively, our model predicts larger energy requirements when starting from states that are farther away from a target memory state. We then posit testable hypotheses regarding which structural properties will lead to efficient stimulation for improving memory encoding based on energy requirements. We show that the strength and homogeneity of edges between controlled and uncontrolled nodes, as well as the persistent modal controllability of the stimulated region, predict energy requirements. Our work demonstrates that white matter architecture plays a vital role in guiding the dynamics of direct electrical stimulation, more generally offering empirical support for the utility of network control theoretic models of brain response to stimulation.
Introduction

Direct electrical stimulation has demonstrated clinical utility in detecting brain abnormalities during surgery (Li et al. (2011)) as well as in mitigating symptoms of epilepsy, essential tremor, and dystonia (Sironi (2011); Perlmutter and Mink (2006); Lozano and Lipsman (2013)). Apart from clinical diagnosis and treatment, direct electrical stimulation has also been used to isolate areas responsible for complex higher-order cognitive functions (Desmurget et al. (2013)) including awareness of actions (Fornia et al. (2020)), semantic memory (Mankin and Fried (2020)), language (Mani et al. (2008)) and face perception (Parvizi et al. (2012)). An open and important question is whether such stimulation can be used to reliably enhance cognitive function, and if so, whether stimulation parameters (e.g., intensity and location) can be optimized and personalized based on individual brain anatomy and physiology. While some studies demonstrate enhancements in spatial learning (Lee et al. (2017)) and memory (Ezzyat et al. (2018); Laxton et al. (2010); Ezzyat et al. (2017); Suthana et al. (2012)) following direct electrical stimulation, others show decrements (Jacobs et al. (2016); Kim et al. (2018c) (for a review, see (Kim et al. (2016))). Such conflicting evidence is also present in the literature on other types of stimulation, including transcranial magnetic stimulation. Proposed explanations range from variations in stimulation intensity (Reichenbach et al. (2011); Mohan et al. (2020)) to individual differences in brain connectivity (Downar et al. (2014)).

A key challenge in circumscribing the utility of stimulation for cognitive enhancement or clinical intervention is the fact that we do not have a fundamental understanding of how an arbitrary stimulation paradigm applied to one brain area alters distributed neural activity in neighboring and distant brain areas (Johnson et al. (2013); Laxton et al. (2010); Lozano and Lipsman (2013)). Models of stimulation propagation through brain tissue range in complexity and biophysical realism (McIntyre et al. (2004b)), from those that only model the region being targeted to those that use finite element models (Yousif and Liu (2009)) or dynamical systems (Steinhardt et al. (2020)) to expand predictions throughout different tissue types. Even in the simpler simulations of the effects of stimulation on a local cell
Figure 5.1: Schematic of Methods. (A) Depiction of network construction and definition of brain state. 
(Left) We segment subjects’ diffusion weighted imaging data into \( N = 234 \) regions of interest using a Lausanne atlas (Cammoun et al. (2012)). We treat each region as a node in a whole-brain network, irrespective of whether the region contains an electrode. Edges between nodes represent mean quantitative anisotropy (Yeh et al. (2013)) along the streamlines connecting them. 
(Right, Top) Practically, we summarize the network in an \( N \times N \) adjacency matrix. 
(Right, Bottom) A brain state is defined as the \( N \times 1 \) vector comprising activity across the \( N \) regions. Any element of the vector corresponding to a region with an electrode is defined as the band-limited power of ECoG activity measured by that electrode. Each brain state is also associated with an estimated probability of being in a good memory state, using a previously validated machine learning classifier approach (Ezzyat et al. (2017)).
(B) A schematic of a single stimulation trial. First, ECoG data is collected for 500 ms. Then, stimulation is applied to a given electrode for a (250 - 1000ms). Finally, ECoG data is again collected after the stimulation.
(C) A schematic of the open loop and optimal control paradigms. In the open loop design, energy \( u(t) \) is applied \textit{in silico} at the stimulation site to the initial, pre-stimulation brain state \( x(0) \). The system will travel to some other state \( x(T) \) as stipulated by our model of neural dynamics, and we will measure the similarity between that predicted state and the empirically observed post-stimulation state. In the optimal control design, the initial brain state \( x(0) \) has some position in space that evolves over time towards a predefined target state \( x(T) \). At every time point, we calculate the optimal energy \( (u(t)) \) required at the stimulating electrode to propel the system to the target state.
population, there are challenges in accounting for the orientation of cells, and the distance from the axon hillock, which can lead to strikingly different circuit behaviors (McIntyre et al. (2004b)). In the more expansive studies of the effects of stimulation across the brain, it has been noted empirically that minute differences in electrode location can generate substantial differences in which white matter pathways are directly activated (Mohan et al. (2020); Riva-Posse et al. (2014)), and that an individual’s white matter connectivity can predict the behavioral effects of stimulation (Horn et al. (2017); Mankin and Fried (2020)). These differences are particularly important in predicting response to therapy, given recent observations that stimulation to white matter may be particularly efficacious in treating depression (Riva-Posse et al. (2013); Mayberg et al. (2005)) and epilepsy (Toprani and Durand (2013)). Despite these critical observations, a first-principles intuition regarding how the effects of stimulation might depend on the pattern of white matter connectivity present in a single human brain has remained elusive.

Network control theory provides a potentially powerful approach for modeling direct electrical stimulation in humans (Tang and Bassett (2017)). Building on recent advances in physics and engineering, network control theory characterizes a complex system as composed of nodes interconnected by edges (Newman (2010)), and then specifies a model of network dynamics to determine how external input affects the nodes’ time-varying activity (Liu et al. (2011)). Drawing on canonical results from linear systems and structural controllability (Kailath (1980)), this approach was originally developed in the context of technological, mechanical, and other man-made systems (Pasqualetti et al. (2014b)), but has notable relevance for the study of natural processes from cell signaling (Cornelius et al. (2013)) to gene regulation (Zanudo et al. (2017)). In applying such a theory to the human brain, one first represents the brain as a network of nodes (brain regions) interconnected by structural edges (white matter tracts) (Bassett and Sporns (2017)), and then one posits a model of system dynamics that specifies how control input affects neural dynamics via propagation along the tracts (Gu et al. (2015b)). Formal approaches built on this model address questions of where control points are positioned in the system (Gu et al. (2015b);
Tang et al. (2017); Muldoon et al. (2016); Wu-Yan et al. (2017)), as well as how to define spatiotemporal patterns of control input to move the system along a trajectory from an initial state to a desired final state (Gu et al. (2017); Betzel et al. (2016b)). Intuitively, these approaches may be particularly useful in probing the effects of stimulation (Muldoon et al. (2016)) and pharmacogenetic activation or inactivation (Grayson et al. (2016)) for the purposes of guiding transitions between cognitive states or treating abnormalities of brain network dynamics such as epilepsy (Ching et al. (2012); Ehrens et al. (2015); Taylor et al. (2015)), psychosis (Braun et al. (2018b)), or bipolar disorder (Jeganathan et al. (2018)). However, this intuition has not yet been validated with direct electrical stimulation data.

Here, we posit a simple theory of brain network control, and we test its biological validity and utility in combined electrocorticography (ECoG) and diffusion weighted imaging (DWI) data from patients with medically refractory epilepsy undergoing evaluation for resective surgery. For each subject, we constructed a structural brain network where nodes represented regions of the Lausanne atlas (Cammoun et al. (2012)) and where edges represented quantitative anisotropy between these regions estimated from diffusion tractography (Yeh et al. (2013)) (Fig. 5.1A). Upon this network, we stipulated a noise-free, linear, continuous-time, and time-invariant model of network dynamics (Gu et al. (2015b); Betzel et al. (2016b); Tang et al. (2017); Gu et al. (2017); Kim et al. (2018b)), from which we built predictions about how regional activity would deviate from its initial state in the presence of exogenous control input to any given node. Using ECoG data acquired from the same individuals during an extensive direct electrical stimulation regimen (Fig. 5.1B), we test these theoretical predictions by representing (i) regional activity as an electrode’s power in a given frequency band, (ii) the pre-stimulation brain state as the power prior to stimulation, and (iii) the post-stimulation brain state as the power after stimulation (Fig. 5.1C). After quantifying the relative accuracy of our theoretical predictions, we next use the model to make more specific predictions about the control energy required to optimally guide the brain from a pre-stimulation state to a specific target state. Here, we select a target state associated with successful memory encoding, though the model could be applied to
any desired target. We quantify successful encoding states using subject-level power-based biomarkers of good memory encoding extracted with a multivariate classifier from ECoG data collected during a verbal memory task (Ezzyat et al. (2017)). Finally, we investigate how certain topological (Kim et al. (2018b)) and spatial (Roberts et al. (2016)) properties of a subject’s network alter its response to direct electrical stimulation, and we ask whether that response is also modulated by control properties of the area being stimulated (Gu et al. (2015b); Muldoon et al. (2016)). Essentially, our study posits and empirically tests a simple theory of brain network control, demonstrating its utility in predicting response to direct electrical stimulation.

Results

Our model assumes the time-invariant network dynamics

\[ \dot{x}(t) = Ax(t) + Bu(t), \]

where the time-dependent state \( x \) is an \( N \times 1 \) vector \((N = 234)\) whose \( i^{th} \) element gives the band-specific ECoG power in sensor \( i \) if \( i \) contained an electrode \((x_i = 1 \text{ otherwise})\), \( A \) is the \( N \times N \) adjacency matrix estimated from DWI data, \( B \) is an \( N \times N \) matrix that selects the control set \( K = u_1, \ldots, u_p \) where \( p \) is the number of regions that receive exogenous control input (in most cases, \( p = 1 \)). In our data, the stimulation site was typically the temporal lobe or cingulate (see Supplemental Figure S5.11 and Supplemental Table T1 for further details regarding electrode location). The input is constant in time and given by \( u(t) = \beta \times I \times \log(\omega) \times (\Delta t) \), where \( I \) is the empirical stimulation amplitude in amperes (range 0.5 - 3 mA), \( \omega \) is the empirical stimulation frequency in hertz (range 10 - 200 Hz), and \( \Delta t \) is the number of simulated samples (here, 950) divided by the empirical stimulation duration (range 250 - 1000 ms) in seconds. Note that since our model is in arbitrary time units with no clear mapping onto physical units of time (i.e. seconds) we incorporate the duration of stimulation into the energy term – following the intuition that longer stimulation sessions add more total energy – rather than incorporating it into the number of time units. The free parameter \( \beta \) scales the input to match the units of \( x \). Biologically, \( \beta \) reflects the
relationship between activity in a cell population and the current from an electrode, which in turn can be influenced by the orientation of the cells, the proximity of the cell body or axons to the electrode, and the quality of the electrode (McIntyre et al. (2004a)) (see Materials and Methods). Intuitively, this model formalizes the hypothesis that white matter tracts constrain how stimulation affects brain state and that those effects can be quantified using network control theory.

Predicting Post-Stimulation States by Open Loop Control

We begin by exercising the model to determine whether our theory accurately predicts changes in brain state induced by direct electrical stimulation. Specifically, we simulate Eq. 5.1 to predict how stimulation alone (independent of other ongoing intrinsic dynamics) will alter brain state, given the structural adjacency matrix $A$ and the initial state $x(0)$ comprised of the ECoG power at every node recorded pre-stimulation ($x_i = 1$ if node $i$ is a region without electrodes, and the $z$-scored power otherwise; see Fig S5.6 for further details). For each stimulation event, we calculate the Pearson’s correlation coefficient between the empirically observed post-stimulation state of ROIs with electrode coverage (an electrode by frequency matrix) and the predicted post-stimulation state at every time point in the simulated trajectory $x(t)$. Furthermore, $x_i = 1$ if node $i$ is a region without electrodes, and the $z$-scored power otherwise (including stimulating and non-stimulating electrodes) (see Fig S5.6 for further details). To measure the capacity of the model simulation to predict the post-stimulation state, we measure the maximum correlation achieved across the model simulation time of arbitrary units. Since there is no clear mapping of stimulated time steps onto physical units of time (i.e. seconds), we chose a number of time steps that was sufficient to allow correlation values to stabilize (Fig S5.1). In the supplement, we provide evidence that results are highly consistent across different time step sizes as long as this stability has been reached. Accordingly, we compute a maximum correlation value across simulated time points between the model prediction and the empirically observed post-stimulation state for each stimulation trial (mean = 0.036 standard deviation = 0.019; Fig. 5.2A). We observe that the mean of the maximum correlation values is significantly
greater than zero ($t$-test $N = 16$, $t = 5.83$, $p = 3.31 \times 10^{-5}$). We note that this correlation represents the impact of stimulation alone on linear dynamics, and does not take into account any other incoming stimuli from the surrounding environment, any ongoing cognitive or metabolic processes, nonlinear dependencies, or inter-frequency interactions (Canolty and Knight (2010); Buzsáki et al. (2012)). Complementing this estimate, we were also interested in the time point (measured in arbitrary units) at which the trial reached its largest magnitude correlation (positive or negative) before decaying towards zero. In our model, the white matter networks define the dynamics of how brain states evolve in time. In addition to affecting the amount that each region changes its activity, the pattern of connections also affects the dynamics of brain states, and how quickly input to the system will dissipate. Intuitively, if more time points are required (and the peak time is large), energy needs longer to spread, and needs to spread across higher order connections compared to when the peak time is small. We observed that the time at which the peak magnitude occurred differed across trials, having a mean of 298 a.u. with a standard deviation of 114 a.u. (Fig. 5.2B).

To determine the influence of network geometry on our model predictions, we compared the empirical observations to those obtained by replacing $A$ in the simulation with one of two null model networks, each designed to independently remove specific geometric features of the structural network (see Fig S5.14 for examples). First for each trial, we constructed a topological null: a randomly rewired network that preserved the edge distribution, number of nodes, and number of edges. Second, we constructed a spatial null: a randomly rewired network that additionally preserved the relationship between edge strength and Euclidean distance. Intuitively, if the observed correlations are due to unique features of human white matter tracts (and not the number edges and their strength, or patterns of connectivity that arise from the spatial embedding of the brain) then we would expect smaller correlations from the null models. Similarly, if the observed peak correlation times are due to the need for energy to spread to unique higher order connections in human white matter networks, we would expect earlier peak times in the null models. Using a repeated measures ANOVA,
Figure 5.2: **Post-Stimulation Brain State Depends on White Matter Network Architecture.** (A) Box plots depicting the average maximum correlation between the empirically observed post-stimulation state and the predicted post-stimulation state at every time point in the simulated trajectory $x(t)$. Box plots indicate the median (solid horizontal black line) and quartiles of the data. Each data point represents a single subject, averaged over all trials (with different stimulation parameters). (B) Box plots depicting the average time to reach the peak magnitude (positive or negative) correlation between the empirically observed post-stimulation state and the theoretically predicted post-stimulation state at every time point in the simulated trajectory $x(t)$. Time is measured in arbitrary units (a.u.). Color indicates theoretical predictions from Eq. 5.1 where $A$ is (i) the empirical network (purple) estimated from the diffusion imaging data, (ii) the topological null network (dark charcoal), and (iii) the spatial null network (light charcoal).
Figure 5.3: **Longer-Distance Trajectories Require More Stimulation Energy.** *(A)* The normalized energy required to transition between the initial state and the post-stimulation state, as a function of the Frobenius norm between the initial state and the post-stimulation state. The black solid line represents the best linear fit (with grey representing standard error), and is provided simply as a guide to the eye. Normalization is also performed to enhance visual clarity. *(B)* The energy required to transition to a good memory state, as a function of the initial probability of being in a good memory state. *(C)* The energy required to transition to a good memory state as a function of the empirical change in memory state resulting from stimulation. *(D)* In three experimental sessions that included both sham and stimulation trials, we calculated the energy required to reach the post-stimulation state or the post-sham state, rather than a target good memory state. Here we show the difference in energy required for sham state transitions in comparison to stimulation state transitions. Error bars indicate standard error of the mean across trials. Across all four panels, different shades of blue indicate different experimental sessions and subjects.
we find a significant main effect across null models ($F(2, 20) = 20.6, p = 1.37 \times 10^5$), and the time at which the maximum correlation values occur ($F(2, 20) = 21.78, p = 9.50 \times 10^{-6}$). We then performed post-hoc analyses and found that the topological null produced significantly weaker maximum correlations between the empirically observed post-stimulation state and the simulated states (paired $t$-test: $N = 11, t = 4.82$, uncorrected $p = 7.04 \times 10^{-4}$), which also peaked significantly earlier in time than the true data ($N = 11, t = 6.68$, uncorrected $p = 5.47 \times 10^{-5}$). The spatial null model also produced significantly weaker maximum correlations between the empirically observed post-stimulation state and the predicted post-stimulation states (permutation test $N = 11, t = 4.27$, uncorrected $p = 1.65 \times 10^{-3}$), which also occurred significantly earlier in time than that observed in the true data ($N = 11, t = 2.83$, uncorrected $p = 0.018$). We observed consistent results in individual subjects (after correcting for multiple comparisons, and with medium to large effect sizes) (see Supplement), across all frequency bands (Fig S5.3), with different values of $\beta$ (Fig S5.2) and when using a smaller resolution atlas (see supplement) for whole brain parcellation. The only exception was that spatial null models did not peak significantly earlier than empirical models after Bonferroni correction for individual frequency bands (see Supplemental Methods). Considering individual variability in DWI estimates, we next asked whether our model would more accurately predict transitions with an individual’s own connectivity, compared to the connectivity of another subject in the same cohort. We did not find a significant difference (paired $t$-test $N = 11, t = -0.40$, $p = 0.70$) (see Supplemental Analyses), indicating that our model generalizes across the subjects in this cohort and does not either depend upon or capitalize upon individual differences in connectivity. Overall, these observations support the notion that structural connections facilitate a rich repertoire of system dynamics following cortical stimulation, and directly constrain the dynamic propagation of stimulation energy in the human brain in a manner consistent with a simple linear model of network dynamics.
Simulating State Transitions by Optimal Network Control

We next sought to use the model to better understand the principles constraining brain state transitions in the service of cognitive function, and their response to exogenous perturbations in the form of direct electrical stimulation. Building on the network dynamics stipulated in Eq. 5.1, we used an optimal control framework to calculate the optimal amount of external input $u$ to deliver to the control set $K$ containing the stimulating electrode, driving the system from a specific pre-stimulation state towards a target post-stimulation state (Fig. 5.1C). Put differently, rather than predicting the brain state changes associated with empirical stimulation for input as we did with our open-loop control model, the optimal control model will analytically solve for the optimal input to get to a specific state. Because this model will necessarily reach the target state that is specified, the optimal control model is better suited to make theoretical predictions about where and when to stimulate rather than to predict state changes based on a certain stimulation paradigm. Here, the specific (or target) post-stimulation state was defined as a period with high probability of successfully encoding a memory, and was operationalized using a previously validated classifier constructed from ECoG data from the same subjects during the performance of a verbal memory task (Ezzyat et al. (2017)) (Fig. 5.1A). We use this target state as a simple, data-driven estimate of a single behaviorally relevant state for illustrative purposes rather than as an exhaustive account of successful memory processes. To determine the optimal input, we use a cost function that minimizes both the energy and the difference of the current state from the target state:

$$
\min_u \int_0^T (x_T - x(t))'S(x_T - x(t)) + \rho u(t)'u(t)dt,
$$

(5.2)

where $x_T$ is the target state, $S$ is a diagonal $N \times N$ matrix that selects a subset of states to constrain (here, $S$ is the identity and all diagonal entries are equal to 1), $\rho$ is the importance of the input penalty relative to the state penalty, $T$ is the time allotted for the simulation, and ('') indicates a matrix transposition (see Fig S5.12 and STAR Methods for details about parameter selection). Since the input $u(t)$ is being solved for rather than defined by the
user, we do not differentiate between the different stimulation parameters used on different trials. Practically, we note that optimizing the cost function in Eq. 5.2 necessarily identifies simulated optimal control trajectories from the pre-stimulation state to a good memory state reasonably close to the target (final distance from target mean = 0.12 standard deviation = 0.06) with minimal error (range from $3.65 \times 10^{-5}$ to $5.19 \times 10^{-4}$).

We begin by addressing the hypothesis that greater energy should be required to reach the target state when it is farther from the initial state. We operationalize this notion by defining distance in four different ways. First, we define distance as the Frobenius norm of the difference between initial and target states. We fit a linear mixed effects model to the integral of the input squared, or energy (here, $Bu$) in every trial, treating the Frobenius norm distance between initial and final state as a fixed effect, and treating subject as a random effect. We find that the distance between initial and final state is positively related to the energy required for the transition ($\beta = 8.3 \times 10^{-3}$, $t(7547) = 18.11$, $p < 2 \times 10^{-16}$) (Fig. 5.3A). Although this result is fairly intuitive, it is also important to consider other measurements of distance that are more informed by biological intuitions about the energy landscape of the brain. For this purpose, we next define distance by the memory capacity in the initial state. It is important to keep in mind that this memory state is defined by a previously trained and validated classifier, and not by task performance during stimulation. We fit a linear mixed effects model to the integral of the input squared in every trial, treating the initial state’s probability of successfully encoding a memory as a fixed effect, and treating subject as a random effect. We find that the initial state’s probability of successfully encoding a memory is negatively related to the energy required for the transition ($\beta = -0.18$, $t(7547) = -14.4$, $p < 2 \times 10^{-16}$) (Fig. 5.3B), suggesting that states that begin closer to the target require less energy to reach the target. Third, we define distance as the observed change in memory state resulting from stimulation. We fit a linear mixed effects model to the input squared in every trial, treating the change in memory state as a fixed effect, and treating subject as a random effect. We find that the change in memory state is positively related to the energy required for the transition ($\beta = 9.5 \times 10^{-2}$, $t(7547) = 8.43$,
$p < 2 \times 10^{-16}$) (Fig. 5.3C).

Taken together, this set of results serves as a basic validation that transitions between nearby brain states will generally require less energy than transitions between distant states. This finding holds whether distance is defined in terms of the difference in Frobenius norm between matrices of regional power, or in terms of the estimated probability to support the cognitive process of memory encoding. In specificity analyses, we also determined whether these relationships were expected in appropriate random network null models. We observed that the relationships were significantly attenuated in theoretical predictions from Eq. 5.1 where $A$ is either the topological null network ($N = 7547, p = 6.1 \times 10^{-4}$) or the spatial null network ($N = 7547, p = 0.0017$) (Fig S5.7). Interestingly, we also found that the largest differences between the empirical relationships and those expected in the null networks were observed in the context of biological measures of distance (e.g., initial probability and change in probability), with only modest differences seen in the statistical measure of distance (the Frobenius norm).

As a fourth and final test of the biological relevance of these findings, we considered sham trials, where no stimulation was delivered, as compared to stimulation trials. Intuitively, we expect that the state that the brain reaches after stimulation is farther away from the initial state than the state that the brain reaches naturally at the conclusion of a sham trial. We first examine this expectation in the context of the Frobenius norm distance discussed above. We observed that 2 out of the 3 experimental sessions that included sham stimulation displayed significantly larger distances (measured by the Frobenius norm) between pre-and post-stimulation states for stimulation conditions than for sham conditions (permutation test, $N > 192, p < 6.8 \times 10^{-3}$ for all subjects). We next tested whether more energy would be required to simulate the transition from the initial pre-stimulation state to the post-stimulation state, than from the initial pre-sham state to the post-sham state. We found consistently greater energy for stimulation trials compared to sham trials in all datasets (paired $t$-test, $N = 3, p = 0.01$; Fig. 5.3D). We further confirmed this finding
with a non-parametric permutation test assessing differences in the distribution of energy values across trials for sham conditions and the distribution of energy values across trials for stimulation conditions (permutation test, \( N > 192, p < 2 \times 10^{-16} \) for all subjects). These observations support the notion that transitions between nearby brain states occur without stimulation (sham) and require little predicted energy, whereas transitions between distant brain states occur with stimulation and require greater predicted energy.

The Role of Network Topology in Stimulation-Based Control

While it is natural to posit that the distance between brain states is an important constraint on the ease of a state transition, there are other important principles that are also likely to play a critical role. Paramount among them is the architecture of the network available for the transmission of control signals. We therefore now turn to the question of which features of the network predict the amount of energy required for each transition from the pre-stimulation state to a good memory state. To address this question, we considered the empirical networks as well as the topological and spatial null model networks discussed earlier. We find that the optimal control input energy required for these state transitions differs across network types (one-way repeated measures ANOVA \( F(2, 20) = 14.75, p = 1.06 \times 10^{-4} \)). In post-hoc testing, we found that the optimal control energy was significantly different between the empirical network and the topological null network \( (N = 11, t = 3.64, p = 4.6 \times 10^{-3}) \) (Fig. 5.4A), but not between the empirical network and the spatial null network \( (N = 11, t = -1.80, p = 0.10) \). This observation suggests that the spatial embedding that characterizes both the real network and the spatial null network may increase the difficulty of control. In supplemental analyses, we test two additional spatially embedded null models that further preserve degree distribution and strength sequence, and we find similar average energies to the empirical and spatial null models discussed here (see Supplement). We hypothesized that the difference in optimal control energy could be mechanistically explained by the determinant ratio, a recently proposed metric quantifying the trade-off between connection strength (facilitating control) and connection homogeneity (hampering control) (Kim et al. (2018b)). Intuitively, a network with a high determinant
ratio will have weak, homogenous connections between the control nodes and nodes being controlled. We found that across all networks the determinant ratio explains a significant amount of variance in energy after accounting for network type (linear mixed effects model with network type and determinant ratio as fixed effects: $\chi^2(2, N = 33) = 13.3, p = 2.65 \times 10^{-5}$) (Fig. 5.4B). These results support the notion that spatial embedding could impose energy barriers by compromising the trade-off between the strength and homogeneity of connections emanating from the stimulating electrode (see Fig S5.9 for extensions to other spatially embedded null models).

**Characteristics of Efficient Regional Controllers**

Thus far, we have seen that the distance of the state transition and the architecture of the network available for the transmission of control signals both impact the energy required. However, neither of these factors address the potential importance of anatomical characteristics specific to the region being stimulated. Such regional effects are salient in the one subject (S8 - 3 stimulation sessions across 7 unique electrodes) in our patient sample who had multiple empirical stimulation sites spanning the same number of ROIs. Since both sites span the same number of ROIs, we know that any differences in energy cannot be due to differences in the control set used in the stimulation. In this patient, we found that transitions from the observed initial state to a good memory state required significantly greater energy when stimulation was delivered to electrodes in the middle temporal region than when stimulation was delivered to the inferior temporal region (permutation test, $N = 555, p < 2 \times 10^{-16}$) (Fig. 5.5A). We hypothesized that this sensitivity to anatomical location could be mechanistically explained by regional persistent and transient modal controllability, which quantify the degree to which specific eigenmodes of the network’s dynamics can be influenced by input applied to that region (Fig S5.13). Energetic input to nodes with high persistent controllability will result in large perturbations to slowly decaying modes of the system, while energetic input to nodes with high transient controllability will result in large perturbations to quickly decaying modes of the system.

To test our hypothesis, we simulated optimal trajectories from the initial state to a good
Figure 5.4: **Topological and Spatial Constraints on the Energy Required for Stimulation-Based Control.** *(A)* Average input energy required for each transition from the pre-stimulation state to a good memory state, as theoretically predicted from Eq. 5.1 where $A$ is (i) the empirical network (purple) estimated from the diffusion imaging data, (ii) the topological null network (dark charcoal), and (iii) the spatial null network (light charcoal). *(B)* The relationship between the determinant ratio and the energy required for the transition from the pre-stimulation state to a good memory state. Note: The color scheme is identical to that used in panel *(A).*
Figure 5.5: Role of Local Topology Around the Region Being Stimulated. 
(A) Transitions from the observed initial state to a good memory state required significantly
greater energy when affected by the middle temporal sensors than when affected by the
inferior temporal sensors. Every color is a subject (N = 11), and every dot is a different
simulated stimulation site. (B) Relationship between persistent (top) or transient (bottom)
controllability of the stimulated region and the energy predicted from optimal transitions
from the initial state to a good memory state. We only allow energy to be injected into a
single electrode-containing region, and we consider a broadband state matrix. (C) As in
panel (B) but when considering the $\alpha$ band state vector only.

memory state while only allowing energy to be injected into a single electrode-containing re-
gion (irrespective of whether or not empirical stimulation was applied there). We then com-
pared the energy predicted from these simulations to the regional controllability. We found
a significant relationship between persistent (but not transient) modal controllability of the
region being stimulated and the input energy of the state transition (linear mixed effects
model accounting for subject: persistent controllability $\chi^2(1, 374) = 3.89$, $p = 0.049$, tran-
sient controllability $\chi^2(1, 374) = 1.69$, $p = 0.19$) (Fig. 5.5B). We note that the strength
of the region being stimulated was not a significant predictor of energy (linear mixed effects
model $\chi^2(1, 374) = 3.5$, $p = 0.061$), though there is only a small difference between the
predictive power of strength, and persistent controllability. Additionally, in the one subject
that had two empirical stimulation locations, we observed that the middle temporal stim-
ulation site with larger energy requirements had smaller persistent controllability (0.058)
than the inferior temporal site with smaller energy (0.072). Given this modest effect for
broad-band state transitions, we next asked whether the influence of regional controllability
varied based on the specific frequency band being controlled. Notably, we found that both
Transient and persistent controllability showed strong relationships to energy in the $\alpha$ band (linear mixed effects model: persistent controllability $\chi^2(1, 374) = 13.8, p = 2.00 \times 10^{-4}$, transient controllability $\chi^2(1, 374) = 11.4, p = 7.5 \times 10^{-4}$; Bonferroni corrected for multiple comparisons across frequency bands) (Fig. 5.5C). Persistent controllability alone also showed a statistically significant relationship for the high gamma band (linear mixed effects model: persistent controllability $\chi^2(1, 374) = 12.2, p = 4.67 \times 10^{-4}$) (Fig S5.10). These findings suggest that the local white matter architecture of stimulated regions can support the selective control of slowly damping dynamics.

Effective Prediction of Energy Requirements

In the previous section, we presented a series of analyses with the goal of elucidating what aspects of brain state and white matter connectivity affect the energy requirements predicted by our model, in an effort to better understand the network wide effects of direct electrical stimulation. Here, we conclude by synthesizing these results into a single model to predict the energy requirements of a stimulation paradigm, given the persistent controllability of the region to be stimulated, the determinant ratio of the network to be controlled, and the probability of encoding a memory at the time of stimulation (Fig5.6A). We fit a random forest model to predict energy given these inputs from our data, and we compared the performance of this model to the performance of a distribution of 1000 models in which the association between energy values and predictors was permuted uniformly at random. We found that our model had an out-of-bag mean squared error of $9.28 \times 10^{-3}$, substantially lower than the null distribution (mean = $9.62 \times 10^{-3}$ and standard deviation = $2.97 \times 10^{-5}$). We also found that our model explained 93.2% of the variance in the predicted energy of the state transition. Random forest models also produce a measure of variable importance, which represents the degree to which including these variables tends to reduce the prediction error. We found that the determinant ratio was the most important (increased node purity = 627), followed by the persistent controllability (320), followed by the initial probability of encoding a memory (23.0). Broadly, these results suggest that the energy requirements for a specific state transition can be accurately predicted given simple features of the connectome.
and the current brain state.

Discussion

While direct electrical stimulation has great therapeutic potential, its optimization and personalization remains challenging, in part due to a lack of understanding of how focal stimulation impacts the state of both neighboring and distant regions. Here we use network control theory to test the hypothesis that the effect of direct electrical stimulation on brain dynamics is constrained by an individual’s white matter connectivity. By stipulating a simplified noise-free, linear, continuous-time, and time-invariant model of neural dynamics, we demonstrate that time-varying changes in the pattern of ECoG power across brain regions is better predicted by an individual’s true white matter connectivity than either topological or spatial network null models. We build on this observation by positing a model for exact brain state transitions in which the energy required for the state transition is minimized, as is the length of the trajectory through the available state space. We use this model to make theoretical predictions about how white matter architecture and brain states make stimulation to these specific states easier. We demonstrate that transitions between more distant states are predicted to require greater energy than transitions between nearby states; these results are particularly salient when distance is defined based on differences in the probability with which a cross-regional pattern of ECoG power supports memory encoding. In addition to the distance between initial and target states, we also find that regional and global characteristics of the network topology predict the energy required for the state transition: networks with smaller determinant ratios (stronger, less homogeneous connections), and stimulation regions with higher persistent controllability, tend to demand less energy. Finally, we demonstrate that these two topological features in combination with the initial brain state explain 93% of the variance in required energy across subjects. Overall, our study supports the notion that control theoretic models of brain network dynamics provide biologically grounded, individualized hypotheses of response to direct electrical stimulation by accounting for how white matter connections constrain state transitions.
Figure 5.6: **Network Topology and Brain State Predict Energy Requirements.** (A) Schematic of the three topology and state features included in the random forest model that we built to predict energy requirements. Network level effects (tan) are captured by the determinant ratio, regional effects (brown) are captured by persistent controllability, and state-dependent effects (red) are captured by the initial memory state. (B) Comparison of the out-of-bag mean squared error for a model where each subject’s determinant ratio, persistent controllability, and initial memory state are used to predict their required energy. We compared the performance of this model to the performance of a distribution of 1000 models in which the association between energy values and predictors was permuted uniformly at random.
Developing theories, models, and methods for the control of neural systems is not a new goal in neuroscience. Whether in support of basic science (e.g., seminal experiments from Hodgkin and Huxley) or in support of clinical therapies (e.g., technological development in brain-machine interfaces or deep brain stimulation), efforts to control neural activity have produced a plethora of experimental tools with varying levels of complexity (Schiff (2011)). Building on these empirical advances, the development of a theory for control of distributed circuits is a logical next step. Network control theory is one particularly promising option. In assimilating brain state and connectivity in a mathematical model (Schiff (2011)), network control theory offers a first-principles approach to modeling neural dynamics, predicting its response to perturbations, and optimizing those perturbations to produce a desired outcome. In cellular neuroscience, network control theory has offered predictions of the functional role of individual neurons in *C. elegans*, and those predictions have been validated by perturbative experiments (Yan et al. (2017a)). While the theory has also offered predictions in humans (Gu et al. (2015b); Muldoon et al. (2016); Ching et al. (2012); Taylor et al. (2015); Jeganathan et al. (2018)), these predictions have not been validated in accompanying perturbative experiments. Here we address this gap by examining the utility of network control theory in predicting empirically recorded brain states, and by validating the fundamental assumption that state transitions are constrained by an individual’s white matter connectivity. The work provides theoretical support for emerging empirical observations that structural connectivity can predict the behavioral effects of stimulation (Horn et al. (2017); Ellmore et al. (2009)), thus constituting an important first step in establishing the promise and utility of control theoretic models of brain stimulation.

**The Principle of Optimal Control in Brain State Transitions**

By positing a model for optimal brain state transitions, we relate expected energy expenditures to a simple, validated estimate of memory encoding, directly relating the theory to a desired behavioral feature. This portion of the investigation was made possible by an important modeling advance addressing the challenge of simulating a trajectory whose
control is dominated by a single node: the stimulating electrode. This type of control is an intuitive way to model stimulation, where one wishes to capture changes resulting from a single input source. However, prior work has demonstrated that while the brain is theoretically controllable from a single point, the amount of energy required can be so large as to make the control strategy impractical (Gu et al. (2015b)). Here we extend prior models of optimal control (Betzel et al. (2016b); Gu et al. (2017)) by relaxing the input matrix \( B \) such that it allows large input to stimulated regions, but also allows small, randomly generated amounts of input at other nodes in the network. Practically, this approach greatly lowers the error of the calculation and also produces narrowly distributed trajectories for the same inputs (see Supplemental Methods).

*Topological Influencers of Control*

Beyond the distance of the state transition, we found that both local and global features of the network topology were important predictors of control energy. In line with previous work investigating controllability radii (Menara et al. (2018b)), energy requirements were lower for randomly rewired networks. Both empirical and topological graphs share the common feature of modularity (Chen et al. (2013)), which is destroyed in random topological null models (Roberts et al. (2016)). Prior theoretical work has demonstrated that modularity is one way in which to decrease the energy of control by decreasing the determinant ratio, a quantification of the relationship between the strength and heterogeneity of direct connections from the controlling node to others (Kim et al. (2018b)). Here we confirmed that the determinant ratio accurately predicted the required energy, while leaving a small amount of variance unexplained. We expected that this unexplained variance could be somewhat accounted for by features of the local network topology surrounding the stimulated node (Tang et al. (2017)). Consistent with our expectation, we found that persistent controllability was the only significant predictor of energy across all frequency bands, indicating a specific role of slow modes in these state transitions. The effect was particularly salient in two bands with consistent (yet different) activity patterns in memory encoding – the \( \alpha \)-band and the high \( \gamma \)-band (Fell et al. (2011); Buzsáki and Moser (2013)). Future avenues for
research could include a comprehensive investigation of whether and why different regional topologies facilitate the control of frequency bands with distinct characteristic changes.

**Clinical Implications**

Our study represents a first step towards developing a control theoretic model to answer two pressing questions in optimizing direct electrical stimulation to meet clinical needs: (i) what changes in the brain after a specified stimulation event, and (ii) which regions are most effective to stimulate. Network control is by no means the only candidate model for answering these questions (McIntyre et al. (2004b); Yousif and Liu (2009); Kim et al. (2011)). Nevertheless, it is a particularly promising model in that it can account for global changes to focal events, is generalizable across any initial and target brain state, and is specific to each individual and their white matter architecture. Unsurprisingly, the linear model of dynamics only captures a small amount of variance observed after stimulation, but stands to benefit from an expansion of the model to non-linear models of dynamics, to time-varying changes in connectivity, and to field spread of stimulation. Importantly, we also show that the optimal control energy for a given transition captures intuitions about the energy landscape of the brain despite being based on simplified linear dynamics. This metric was then used to identify features of white matter architecture that could facilitate control. Investigation into whether metrics could be incorporated into existing multimodal predictions of stimulation outcome is a logical next step in developing a tool for clinical selection of stimulation regions. Finally, an evaluation of long-term efficacy of specific stimulation paradigms informed by principles of network control is warranted, and would benefit from work in non-human animal models where precise measurements of plasticity are accessible.

**Methodological Considerations**

**Primary Data**

As with any model of complex biological systems, our results must be interpreted in the context of the underlying data. First, we note that DWI data provides an incomplete picture of white matter organization, and even state-of-the-art tractography algorithms can identify
spurious connections (Thomas et al. (2014)). As higher resolution imaging, reconstruction, and tractography methods emerge, it will be important to replicate the results we report here. Second, while ECoG data provides high temporal resolution, it is collected from patients with epilepsy and results might not generalize to a healthy population (Parvizi and Kastner (2017)). However, it is worth noting that recent work has shown that tissue damage resulting from recurrent seizures can be minimal (Rossini et al. (2017)), and most electrodes are not placed in epileptic tissue (Parvizi and Kastner (2017)). Nevertheless, this population can display atypical physiological signatures of memory (Glowinski (1973)), as well as atypical white matter connectivity (Gross et al. (2006)). It will be important in future to extend this work to non-invasive techniques accessible to healthy individuals.

Modeling Assumptions

Our results must also be interpreted in light of model assumptions. First, we consider a relaxed input matrix to ensure that state transitions are primarily influenced by the set of stimulating electrodes and to a lesser extent non-stimulating electrodes. This choice is not a true representation of single point control, but instead reflects the fact that the system is constantly modulated by endogenous sources (Gu et al. (2017); Betzel et al. (2016b)). Second, our model uses a time-invariant connectivity matrix. While DWI data is relatively stable over short time-scales, repeated stimulation can result in dynamic changes in plasticity that are not captured here (Malenka and Bear (2004)).

Lastly, we note that our model assumes linear network dynamics. While the brain is not a linear system, such simplified approximations can predict features of fMRI data (Honey et al. (2007)), predict the control response of nonlinear systems of coupled oscillators (Muldoon et al. (2016)), and more generally provide enhanced interpretability over nonlinear models (Kim and Bassett (2019)). Nevertheless, considering control in nonlinear models of neural dynamics will constitute an important next step for two reasons. First, nonlinear models of brain dynamics can capture a richer repertoire of brain states that is more consistent with the repertoire observed in neural data (Jirsa et al. (2014); Jirsa and Haken...
(1996); Breakspear et al. (2003); Messe et al. (2015, 2014); Hansen et al. (2015)). Second, nonlinear approaches offer distinct types of control strategies. Specifically, linear control is frequently used to examine the transition between an initial state and a final state. Yet, some hypotheses about neural function might benefit from nonlinear control approaches – such as feedback vertex set control (Zanudo et al. (2017); Cornelius et al. (2013)) – that allow one to examine the transition from one manifold of activity to another (Slotine and Li (1991); Sontag (2013)). Such attractor-based control seems intuitively appropriate for the study of complex behaviors that are not well-characterized by a single pattern of activity, but rather by a different trajectory through many states. Despite some progress, nonlinear approaches still lag far behind linear control approaches in their applicability and capability, and thus further theoretical work is needed (Slotine and Li (1991); Sontag (2013)).

Defining Brain States

In our model, a brain state represents the z-scored power across electrodes in eight logarithmically spaced frequency bands from 1 to 200 Hz. This choice was guided by (i) the goal of maintaining consistency with the brain states on which the memory classifier was trained, and (ii) the fact that power spectra are well-documented behavioral analogs for memory (Ezzyat et al. (2017); Fell et al. (2011); Buzsáki and Moser (2013)). Yet, since many power calculations require convolution with a sine wave, power is insensitive to non-sinusoidal and phase-dependent features of the signal (Schalk et al. (2017); Cole et al. (2017); Vinck et al. (2011)). It would be interesting in future to explore transitions in other state spaces, such as instantaneous voltage (Schalk et al. (2017)). Lastly, it is important to note that our algorithm controls each frequency band independently, although incorporating inter-frequency coupling (Bonnefond et al. (2017); Canolty and Knight (2010)) could be an interesting direction for future work. These considerations involving brain state also affect the interpretation of our target state as a good memory state. While our selection of target state does not exhaustively sample patterns of brain activity in which successful encoding can occur, and only makes claims about a narrow range of all memory processes (encoding specifically), for the purposes of exploring the utility of network control theory in modeling
direct stimulation for the first time this classifier provides an important, if relatively narrow, behavioral link.

Conclusions and Future Directions

Our study begins to explore the role of white matter connectivity in guiding direct electrical stimulation, with the goal of driving brain dynamics towards states with a high probability of memory encoding. We demonstrate that our model of targeted direct electrical stimulation tracks well with biological intuitions, and is influenced by both regional and global topological properties of underlying white matter connectivity. Overall, we show that our control theoretic model is a promising method that has potential to inform hypotheses about the outcome of direct electrical stimulation.

STAR Methods

Contact for Reagent and Resource Sharing

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Danielle Bassett (dsb@upenn.edu).

Experimental Model and Subject Details

Diffusion Weighted Imaging – DWI

Diffusion imaging data were acquired from either the Hospital for the University of Pennsylvania (HUP), or Jefferson University Hospital. At HUP, all scans were acquired on a 3T Siemens TIM Trio scanner with a 32-channel phased-array head coil. Each data acquisition session included both a DWI scan as well as a high-resolution T1-weighted anatomical scan. The structural scan was conducted with an echo planar diffusion weighted technique acquired with iPAT using an acceleration factor of 2. The diffusion scan had a b value of 2000 s/mm$^2$ and TE/TR = 117/4180 ms. The slice number was 92. Field of view read was 210 mm and slice thickness was 1.5 mm. Acquisition time per DWI scan was 8:26 min. The anatomical scan was a high-resolution 3D T1-weighted sagittal whole-brain image using an MPRAGE sequence. It was acquired with TR = 2400 ms; TE = 2.21 ms; flip angle = 8 degrees; 208 slices; 0.8 mm thickness. At Jefferson University Hospital, all scans were acquired on a 3T Philips Acheiva scanner. Each data acquisition session included both a
DWI scan as well as a high-resolution T1-weighted anatomical scan. The diffusion scan was 61-directional with a b value of 3000 s/mm² and TE/TR = 7517/98 ms, in addition to 1 b0 images. Matrix size was 96 × 96 with a slice number of 52. Field of view was 230 × 130 × 230 mm² and slice thickness was 2.5 mm. Acquisition time per DWI scan was just over 9 min. The anatomical scan was a high-resolution 3D T1-weighted sagittal whole-brain image using an MPRAGE sequence. It was collected in sagittal orientation with in-plane resolution of 256 × 256 and 1 mm slice thickness (isotropic voxels of 1 mm³, 170 slices, TR = 650 ms, TE = 3.2 ms, Field of view 256 mm, flip angle 8 degrees, SENSE factor = 1, duration = 5 min).

Diffusion volumes were skull-stripped using FSL’s BET, v5.0.10. Volumes were subsequently corrected for eddy currents and motion using FSL’s EDDY tool, v.5.0.10 (Andersson and Sotiropoulos (2016)). Anatomical scans were processed with FreeSurfer v6.0.0. Surface reconstructions were used to generate subject-specific parcellations based on the Lausanne atlas from the Connectome Mapper Toolbox (Daducci et al. (2012)). Each parcel was then individually warped into the subject’s native diffusion space. Using DSI-Studio, orientation density functions (ODFs) within each voxel were reconstructed from the corrected scans using GQI (Yeh et al. (2013)). We then used the reconstructed ODFs to perform a whole-brain deterministic tractography using the derived QA values in DSI-Studio (Yeh et al. (2013)). We generated 1,000,000 streamlines per subject, with a maximum turning angle of 35 degrees and a maximum length of 500 mm (Cieslak and Grafton (2014)). We hold the number of streamlines between participants constant (Griffa et al. (2013)).

**Electrocorticography – ECoG**

Electrocorticography data were collected on eleven subjects (age 32 ± 10 years, 63.6% male and 36.4% female) as part of a multi-center project designed to assess the effects of electrical stimulation on memory-related brain function. Data were collected at Thomas Jefferson University Hospital and the Hospital of the University of Pennsylvania. The research protocol was approved by the institutional review board (IRB) at each hospital.
and informed consent was obtained from each participant. Electrophysiological data were collected from electrodes implanted subdurally on the cortical surface as well as deep within the brain parenchyma. In each case, the clinical team determined the placement of the electrodes to best localize epileptogenic regions. Subdural contacts were arranged in both strip and grid configurations with an inter-contact spacing of 10 mm. Depth electrodes had 8-12 contacts per electrode, with 3.5 mm spacing.

Electrodes were anatomically localized using separate processing pipelines for surface and depth electrodes. To localize depth electrodes we first labeled hippocampal subfields and medial temporal lobe cortices in a pre-implant, 2 mm thick, coronal T2-weighted MRI using the automatic segmentation of hippocampal subfields (ASHS) multi-atlas segmentation method (Yushkevich et al. (2015)). We additionally used whole brain segmentation to localize depth electrodes not in medial temporal lobe cortices. We next co-registered a post-implant CT with the pre-implant MRI using Advanced Normalization Tools (ANTs) (Avants et al. (2008)). Electrodes visible in the CT were then localized within subregions of the medial temporal lobe by a pair of neuroradiologists with expertise in medial temporal lobe anatomy. The neuroradiologists performed quality checks on the output of the ASHS/ANTs pipeline. To localize subdural electrodes, we first extracted the cortical surface from a pre-implant, volumetric, T1-weighted MRI using Freesurfer (Fischl et al. (2004)). We next co-registered and localized subdural electrodes to cortical regions using an energy minimization algorithm. For patient imaging in which automatic localization failed, the neuroradiologists performed manual localization of the electrodes.

Intracranial data were recorded using one of the following clinical electroencephalogram (EEG) systems (depending on the site of data collection): Nihon Kohden EEG-1200, Natus XLTek EMU 128, or Grass Aura-LTM64. Depending on the amplifier and the preference of the clinical team, the signals were sampled at either 500 Hz, 1000 Hz or 1600 Hz and were referenced to a common contact placed either intracranially, on the scalp, or on the mastoid process. Intracranial electrophysiological data were filtered to attenuate line noise
(5 Hz band-stop fourth order Butterworth, centered on 60 Hz). To eliminate potentially confounding large-scale artifacts and noise on the reference channel, we re-referenced the data using a bipolar montage. To do so, we identified all pairs of immediately adjacent contacts on every depth electrode, strip electrode, and grid electrode, and we took the difference between the signals recorded in each pair. The resulting bipolar timeseries was treated as a virtual electrode and used in all subsequent analysis. We performed spectral decomposition of the signal into 8 logarithmically spaced frequencies from 3 to 180 Hz. Power was estimated with a Morlet wavelet, in which the envelope of the wavelet was defined with a Gaussian kernel that allowed for 5 oscillations of the frequency of interest (one of 8, from 3-180 Hz). This kernel was then convolved with 500 ms epochs of ECoG data before and after stimulation to obtain estimates of power. The resulting time-frequency data were then log-transformed, and z-scored within session and within frequency band across events.

Method Details

Stimulation Protocol

During each stimulation trial, we delivered stimulation using charge-balanced, biphasic, rectangular pulses with a pulse width of 300 µs. We cycled over the following parameters in consecutive trials: pulse frequency (10–200 Hz), pulse amplitude (0.5–3.0 mA), stimulation duration (0–1 sec), and inter-stimulation interval (2.75–3.25 sec). These stimulation parameter ranges were chosen to be well below the accepted safety limits for charge density, and ECoG was continuously monitored for after-discharges by a trained neurologist. Some subjects (N = 8) only received stimulation to one set of regions, while other received stimulation to multiple sets of regions (N = 3) (Table T1). Each subject’s stimulating electrodes are shown in Fig S5.11. Most electrodes were in the temporal lobe, with some in the cingulate and frontal lobe.

Memory State Classification and Good Memory State Definition

Prior to collecting the data used in this study, each subject had a memory classifier trained based on their performance during a verbal memory task. The input data that we used
was the spectral power averaged across the time dimension for each word encoding epoch (0-1600ms relative to word onset). Each subject’s personalized classifier was then used to return a likelihood of being in a good memory state for each pre- and post-stimulation recording. For more information about the classifier and the task design, see (Ezzyat et al. (2017)). A good memory state was defined for each subject using this classifier output. The target state was defined as the average of the top 5% of states with the largest probabilities (returned from the classifier) associated with them. The threshold of 5% was chosen as the smallest threshold that reliably included sufficient trials in the average (minimum number of trials was 192). The probabilities associated with these final target states ranged from 0.61 to 0.74.

The Mathematical Model - Open Loop Control

We use network control theory to model the effect of stimulation on brain dynamics because it accounts for systems level properties of brain states alongside external input. The theory requires us to stipulate a model of brain dynamics as well as a formulation of the network connecting brain areas whose time-varying state in response to stimulation we wish to understand. As described in the main manuscript, we use a linear time invariant model:

$$\dot{x}(t) = Ax(t) + Bu(t),$$  

where $x(t)$ is a $N \times 1$ vector that represents the brain state at a given time, and $N$ is the number of regions ($N = 234$). More specifically $x(t)$ is the $z$-scored power at time $t$ in $m$ regions containing electrodes. The $N-m$ regions without electrodes are assigned an initial and target state equal to 1. In the network adjacency matrix $A$, each $ij$th element gives the quantitative anisotropy between region $i$ and region $j$. Note that we scale $A$ by dividing it by its largest eigenvector and then we subtract the identity matrix; these choices assure that $A$ is stable.

The $N \times 1$ input vector $u(t)$ represents the input required to control the system. Lastly, $B$ is the $N \times N$ input matrix whose diagonal entries select the regions that will receive
input, and this set of selected regions is referred to as the control set \( \kappa \). Here \( \mathbf{B} \) will be selected to assure that the input energy is concentrated at the stimulating electrode; to increase the computational tractability of the control calculation, \( \mathbf{B} \) will also be selected to include additional control points. Specifically, if \( i \) represents the index of the stimulating electrode, then \( \mathbf{B}(i, i) = 1 \). If \( j \) is the index of a region containing a different electrode, then \( \mathbf{B}(j, j) = 0 \). Lastly, if \( k \) is the index of a region that does not contain an electrode, then \( \mathbf{B}(k, k) = \alpha \), where \( \alpha \) is randomly drawn from a normal distribution with mean \( 5 \times 10^{-4} \) and standard deviation \( 5 \times 10^{-5} \). The distribution was chosen specifically to give a narrow range of values with a relative standard deviation of 10\%, and a mean that was small enough to allow stimulation control to dominate the dynamics, but large enough to improve the computational tractability of the problem.

### The Mathematical Model - Optimal Control

Our longterm goal is to use the model described above to predict optimal parameters for stimulation. To take an initial step towards that goal, we seek to estimate the optimal energy required to reach a state that is beneficial for cognition, and we therefore define the following optimization problem:

\[
\min_{\mathbf{u}} \int_{0}^{T} \left( (\mathbf{x}_T - \mathbf{x}(t))^T \mathbf{S} (\mathbf{x}_T - \mathbf{x}(t)) + \rho \mathbf{u}_\kappa(t)^T \mathbf{u}_\kappa \right) dt,
\]

s.t. \( \dot{\mathbf{x}} = \mathbf{A} \mathbf{x}(t) + \mathbf{B} \mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \quad \text{and} \quad \mathbf{x}(T) = \mathbf{x}_T, \quad (5.4) \]

where \( \mathbf{x}_T \) is the target state, \( T \) is the control horizon, a free parameter that defines the finite amount of time given to reach the target state, and \( \rho \) is a free parameter that weights the input constraint. We also define \( \mathbf{S} \) to be equal to the identity matrix, in order to constrain all nodes to physiological activity values. The input matrix \( \mathbf{B} \) was defined to allow input that was dominated by the stimulation ROI. More specifically, rather than being characterized by binary state values, regions without electrodes were given a value of approximately \( 5 \times 10^{-5} \) at their corresponding diagonal entry in \( \mathbf{B} \). This additional input ensured that the calculation of optimal energy was computationally tractable (which is not the case for input
applied to a very small control set). With these definitions, two constraints emerge from our optimization problem. First, \((x_T - x(t))^T S(x_T - x(t))\) constrains the trajectories of a subset of nodes by preventing the system from traveling too far from the target state. Second, \(\rho u_\kappa(t)^T u_\kappa\) constrains the amount of input used to reach the target state, a requirement for biological systems, which are limited by metabolic demands and tissue sensitivities.

To compute an optimal \(u^*\) that induces a transition from the initial state \(Sx(0)\) to the target state \(Sx(T)\), we define the Hamiltonian as

\[
H(p, x, u_\kappa, t) = x' Sx + \rho u_\kappa'u + p(Ax + Bu_\kappa).
\] (5.5)

According to the Pontryagin minimization principle, if \(u^*_\kappa\) is a solution with the optimal trajectory \(x^*\), then there exists a \(p^*\) such that

\[
\frac{\partial H}{\partial x} = -2S(x_T - x^*) + A' p^* = -\dot{p}^*;
\]

\[
\frac{\partial H}{\partial p} = Ax^* + Bu_\kappa;
\]

\[
\frac{\partial H}{\partial u_\kappa} = 2\rho u^*_\kappa + B' p^* = 0.
\]

From Eqs. (5.4)-(5.6), we can derive that

\[
u^*_\kappa = -\frac{1}{2\rho} B' p^*, \tag{5.6}\]

\[
\dot{x}^* = Ax^* - \frac{1}{2\rho} BB' p^*, \tag{5.7}\]

such that the only unknown is now \(p^*\). Next, we can rewrite Eqs. (5.4) and (5.8) as

\[
\begin{bmatrix}
\dot{x}^* \\
\dot{p}^*
\end{bmatrix} =
\begin{bmatrix}
A & \frac{1}{2\rho} BB' \\
-2S & -A'
\end{bmatrix}
\begin{bmatrix}
x^* \\
p^*
\end{bmatrix} +
\begin{bmatrix}
0 \\
2S
\end{bmatrix} x_T.
\] (5.8)
Let us define

\[ \tilde{A} = \begin{bmatrix} A & \frac{1}{2p}BB' \\ -2S & -A' \end{bmatrix}, \]
\[ \tilde{x} = \begin{bmatrix} x^* \\ p^* \end{bmatrix}, \]
\[ \tilde{b} = \begin{bmatrix} 0 \\ 2S \end{bmatrix}x_T, \]

so that Eq. (5.9) can be rewritten as

\[ \dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{b}, \quad (5.9) \]

which can be solved as

\[ \tilde{x}(t) = e^{\tilde{A}t}\tilde{x}(0) + \tilde{A}^{-1}(e^{\tilde{A}t}\tilde{x}(0) - I)\tilde{b}. \quad (5.10) \]

Let

\[ c = \tilde{A}^{-1}(e^{\tilde{A}t}\tilde{x}(0) - I)\tilde{b}, \quad (5.11) \]

and

\[ e^{\tilde{A}T} = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}. \quad (5.12) \]

Then, by fixing \( t = T \), we can rewrite Eq. (5.10) as

\[ \begin{bmatrix} \dot{x}^*(T) \\ \dot{p}^*(T) \end{bmatrix} = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \begin{bmatrix} \dot{x}^*(0) \\ \dot{p}^*(0) \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \quad (5.13) \]

From this expression we can obtain

\[ x^*(T) = E_{11}x^*(0) + E_{12}p^s(0) + c_1. \quad (5.14) \]
Moreover, if we let $\bar{S} = I - S$, then as a known result in optimal control theory (Bryson (1996)), $\bar{S}p^*(T) = 0$. Therefore,

$$\bar{S}p^*(T) = \bar{S}E_{21}x^*(0) + \bar{S}E_{22}p^* + \bar{S}c_2 = 0. \quad (5.15)$$

We can now solve for $p^*(0)$ as follows:

$$p^*(0) = \begin{bmatrix} SE_{12} \\ SE_{22} \end{bmatrix}^+ \left( - \begin{bmatrix} SE_{11} \\ SE_{21} \end{bmatrix} x^*(0) - \begin{bmatrix} Sc_1 \\ Sc_2 \end{bmatrix} + \begin{bmatrix} Sx(T) \\ 0 \end{bmatrix} \right), \quad (5.16)$$

where $\mathbb{X}^+$ indicates the Moore-Penrose pseudoinverse of a matrix. Now that we have obtained $p^*(0)$, we can use it and $x^*$ (or $x(0)$) to solve for $\tilde{x}$ via forward integration. To solve for $u^*_\kappa$, we simply take $p^*$ from our solution for $\tilde{x}$ and solve Eq. (5.7).

**Parameter Selection**

Our optimal control framework has three free parameters: $\gamma$, the scaling of the matrix $A$, $\rho$, the relative importance of the input constraint over the distance constraint, and $T$, the control horizon, or amount of time given for the system to converge. Intuitively, $\gamma$, which is only applied after the matrix has been scaled to be stable, controls the time scale of the dynamics of the system: large values down-weight the smaller eigenmodes, causing them to damp out more quickly. Very large values of this parameter tend to increase the computational complexity of estimating the matrix exponentials. Lower values of the parameter $\rho$ correspond to relaxing the constraint on minimal energy, leading to larger energies but lower error values. The final parameter $T$ determines how quickly the system is required to converge. Small values of $T$ will make the system difficult to control, and likely lead to larger error and energy. Moderately large values of $T$ will give the system more time to converge, and will typically lower the error. However, very large values of $T$ will also increase the difficulty of calculating the matrix exponentials, and will lead to high error values.
Because we lack strong, biologically motivated hypotheses to help us in choosing values for these parameters, we explored a range of values for all three parameters, and found the set that produced the smallest error in the optimal control calculation. We chose this approach rather than the alternative of fitting the model to resting state data for two reasons. First, solving optimal control problems can easily become computationally intractable for large matrices with sparse control sets, both of which are features of our model. This inherent difficulty decreases our confidence in fitting the model to resting state data, and increases the expected uncertainty in parameter estimates derived therefrom. Second, since we are explicitly modeling exogenous control and our parameters relate directly to that exogenous input, we expect that the parameters that best fit resting state data would be very different from those that best fit stimulation data. For each parameter, we first calculated the error of the simulations for parameter values that were logarithmically spaced between 0.001 and 100. We then selected a subspace of those parameter values that produced small error values. From this subspace, we calculated the $z$-score of each error value, and we identified the region in the three-dimensional space in which the $z$-score was less than or equal to $-1$. We then took the average coordinate in this space across subjects, and the 3 parameter values specified by this coordinate became our parameter set of interest for all main analyses presented in our study. This process is illustrated in Fig. S5.12. Specifically, the parameters selected were $\gamma = 4$, $T = 0.7$, and $\rho = 0.3$. For the purposes of reliability and reproducibility, here in the supplement we also report several results for key analyses when using two different sets of parameters that also produced low error. The two additional sets used were $\gamma = 7$, $T = 0.4$, and $\rho = 0.1$, and $\gamma = 3$, $T = 0.9$, and $\rho = 0.5$.

Quantification and Statistical Analysis

Post-Stimulation State Correlations - Open Loop Control

We simulated stimulation to a given region in the Lausanne atlas from the observed pre-stimulation state ($x(i)$ is the $z$-scored power if $i$ is a region with an electrode, $x(i) = 1$ otherwise). We then calculated the two-dimensional Pearson’s correlation coefficient between the empirically observed post-stimulation state and the predicted post-stimulation
state at time points \( t = 5 \) to \( t = T \) in the simulated trajectory \( x(t) \). The time points \( t < 5 \) were excluded to prevent the initial state, or the trajectory very near to it, from being considered as the peak. We calculated two statistics of interest: the maximum correlation reached and the time at which the largest magnitude (positive or negative) correlation occurred.

**Metrics for Energy and Simulation Error - Optimal Control**

We calculated trajectories for each of 8 logarithmically spaced frequency bands spanning 1 to 200 Hz, and then we combined them into a single state matrix for most analyses reported in the main manuscript. Then we calculated distances between the initial and final states using only the \( m-p \) regions that had variable states.

*Energy:* To quantify differences in trajectories, and the ease of controlling the system, we calculated a single measure of energy for every trajectory. We used a measure of total energy that incorporates the weights of \( B \) in addition to the energy \( u \):

\[
E_{x_0, x_T} = \int_0^T \| B_{x_0} u_{x_T} \|_2^2 dt . \tag{5.17}
\]

Our decision to define \( B \) as a weighted, rather than binary matrix made the problem of optimal control much more tractable, but also necessitated the incorporation of \( B \) into the calculation of energy for a more representative estimate. Trajectories were simulated for each frequency band, and these trajectories were combined into a single state matrix for all analyses, unless otherwise specified (e.g., as in Fig. 5.5C and in some figures in the Supplementary Materials). More specifically, comparisons of brain state were calculated as the two-dimensional Pearson’s correlation coefficient between simulated region-by-frequency matrices and empirical region-by-frequency matrices (Fig. 5.2). Only regions with electrodes were included in correlations, as they were the only regions with initial state measurements. Energy in all optimal control analyses was calculated in each band independently, and then summarized in a region-by-frequency matrix at each time point (Fig. 5.3 - 5.6). A single measure of energy for a trial was calculated by integrating the
Frobenius norm of the energy matrix over time.

**Numerical Error:** Because optimal control is a computationally difficult problem, we also calculate the numerical error associated with each computation. The numerical error is calculated as

\[ n_{\text{err}} = \left\| \left( \begin{bmatrix} \bar{SE}_{12} \\ \bar{SE}_{22} \end{bmatrix} \right) p^* + \left( \begin{bmatrix} \bar{SE}_{11} \\ \bar{SE}_{21} \end{bmatrix} x^*(0) + \begin{bmatrix} \bar{Sc}_1 \\ \bar{Sc}_2 \end{bmatrix} - \begin{bmatrix} Sx(T) \\ 0 \end{bmatrix} \right) \right\|. \] (5.18)

**Network Statistics**

To probe the role of graph architecture in the energy required for optimal control trajectories, we calculated the determinant ratio, which is defined as the ratio of the strength to the homogeneity of the connections between the first degree driver (anything with a non-zero entry in \( B \)) and the non-driver (anything with a zero entry in \( B \)) (Kim et al. (2018b)). This metric was derived assuming that a system has a greater number of driver nodes than non-driver nodes, and that the initial and final states are distributed around zero. Quantitatively, the trade-off between strength and homogeneity is embodied in the ratio between the determinant of the Gram matrix of all driver to non-driver connections, and the determinant of that same matrix with each non-driver node removed iteratively. The gram matrix here is the inner product of the vectors giving connections from driver nodes to and non-driver nodes. More specifically, if \( C \) is the Gram matrix of all driver to non-driver connections, and \( C_k \) is the matrix of all connections from driver nodes to all but the \( k \)th non-driver node, the determinant ratio is defined by \( N^{-1} \sum_{k=1}^{N} \frac{\text{det}(C_k)}{\text{det}(C)} \). Since the calculation of the determinant of large matrices can be computationally challenging, we use the equivalent estimate of the trace of the inverse of the Gram matrix, Trace\((C^{-1})\), to calculate the average determinant ratio (see Kim et. al. for a full derivation) (Kim et al. (2018b)).

To understand the expected differences in stimulation-induced dynamics based on which region is actually being stimulated, we calculated two network control statistics: the *persis-
tent modal controllability and the transient modal controllability. Intuitively, the persistent (transient) controllability is high in nodes where the addition of energy will result in large perturbations to the slow (fast) modes of the system (Gu et al. (2015b)). Typically, modal controllability is computed from the eigenvector matrix \( V = [v_{ij}] \) of the adjacency matrix \( A \). The \( j \)th mode of the system is poorly controllable from node \( i \) if the entry for \( v_{ij} \) is small. Modal controllability is then calculated as \( \phi_i = \sum_{j=1}^{N} (1 - \lambda_j^2(A)) v_{ij}^2 \). We adapt this discrete-time estimate to continuous-time by defining modal controllability to be \( \phi_i = \sum_{j=1}^{N} (1 - (e^{\lambda_j(A)\delta t})^2) v_{ij}^2 \). Here, \( \delta t \) is the time step of the trajectory and \( e^{\lambda_j(A)\delta t} \) is the conversion from continuous to discrete eigenvalues of the system. Persistent (transient) modal controllability are computed in the same way, but using only the 10% largest (smallest) eigenvalues of the system. We chose 10% as a strict (allowing few modes to be considered) cutoff, that also showed a large amount of variance across nodes for both metrics (Fig S5.6).

Here we complement the regional metric analysis reported in the main manuscript by also testing two additional metrics: average controllability and communicability. Intuitively, average controllability is proportional to the average input energy needed for a certain set of nodes to drive the system to all possible target states (though this was only proven mathematically using a full control set). This metric is interpreted as a node’s ability to push the network to many easy-to-reach states (Gu et al. (2015b)). Average controllability is proportional to the Trace\( (W^{-1}) \), where \( W^{-1} \), the inverse of the controllability Gramian, is defined as \( W_\kappa = \sum_{\tau=1}^{\infty} A^\tau B_\kappa B_\kappa^T A^\tau \). Here, \( B_\kappa \) identifies a specific control set \( \kappa \). Following prior work, we calculate average controllability as Trace\( (W_\kappa) \), because the inverse is often poorly conditioned (Gu et al. (2015b)).

Intuitively, communicability is a measure of how well a node communicates with every other node in the network. It is similar to network efficiency (Latora and Marchiori (2001)), but considers all paths and walks between two nodes, rather than only the shortest paths. This feature is useful because, biologically, non-shortest paths (such as thalamocortical loops)
can be important in many computations (Crofts et al. (2010)). The metric is weighted such that shorter paths carry more weight. Specifically, we calculated weighted communicability as $G = e^{D^{-1/2}AD^{-1/2}}$ and the average communicability for each node as $g_i = \frac{1}{N} \sum_{j=1}^{N} G_{i,j}$. Here $N$ is the number of nodes in the network, and $D$ is the diagonal weighted degree matrix where $D_{i,i} = d_i$. We have chosen a measure of communicability where longer paths are weighted by a factor of $1/k!$ because it is a standard measure in the field, and because it can be justified by arguments from statistical mechanics (Crofts et al. (2010)); however other weighting schemes could also be used.

### Null Models

We compared the empirically observed values – of the maximum correlation reached and the time at which the largest magnitude correlation occurred – to those expected under two null models: (i) a topological null model that preserved only the number of edges and their total strength, and (ii) a spatially embedded null model that also preserved the relationship between edge strength and edge distance. Instantiations of the topological null model were generated using the Brain Connectivity Toolbox. The rewiring algorithm begins by randomly choosing two pairs of edges ($i \rightarrow j$ and $k \rightarrow l$) and continues by swapping their origin and termination points ($i \rightarrow k$ and $j \rightarrow l$). Here, we performed $2 \times 10^4$ bidirectional edge swaps per network. Instantiations of the spatially embedded model were generated using code from (Roberts et al. (2016)). The rewiring algorithm begins by calculating the Euclidean distance between the average coordinates of all regions in the Lausanne atlas, and continues by removing the effect of distance on the mean and variance of the edge weights, randomly rewiring, and then adding the effect of distance back to the newly rewired graph. For both topological and spatial null model analyses, a new random graph was generated for every trial (minimum number of trials was 192). Null models were created from the stabilized rather than raw versions of the structural matrices, and – in the optimal control analyses – were also scaled by a parameter $\gamma$ to reduce the error of the calculation.

To further explore the role of spatial embedding in optimal control efficiency, we tested two
additional null models: (i) a spatially embedded null model that also preserves the degree distribution, and (ii) a spatially embedded null model that further preserves the strength sequence. Exemplar spatially embedded null model graphs were generated using code from (Roberts et al. (2016)). Similarly to the spatially embedded null model described above, all calculations for the additional spatially embedded null model graphs begin with a calculation of the Euclidean distance between the average coordinates of all regions in the Lausanne atlas. Next, we remove the effect of distance on the mean and variance of the edge weights. Pairs of edges are then swapped uniformly at random, and the effects of distance are added back in to the matrix. While these measures of Euclidean distance ignore the curvilinear character of white matter tracts, the true fiber length and the Euclidean distance are highly correlated (Roberts et al. (2016)). In the strength distribution preserving null model, both the row and column sums are then iteratively updated to converge to the empirical strength distribution. The strength sequence preserving null model graph was defined similarly, but with a convergence to the strength sequence rather than to the strength distribution. The strength distance relationships were then added back into the graph. More details about these processes can be found in (Roberts et al. (2016)). For these analyses, a new exemplar null model graph was generated for every trial (minimum number of trials was 192). Null models were created from stable matrices, scaled by the parameter $\gamma$. Examples of null models used in the main text and supplement are shown in Fig S5.14. Note that with the exception of the randomly rewired null model, other models look qualitatively similar to empirical connectivity matrices.

**Random Forest Models**

Random forest models are constructed by averaging predictions over a large number of decision trees (here: 500), where each branch in the tree splits one of the predictors into two groups, the means of which are used as a predicted value for observations in each branch (Liaw and Wiener (2002)). Splits are selected to reduce prediction error. Random forest models rely on bootstrapping data for each split, and a random selection of the variable to split on to avoid overfitting the data. Out-of-bag mean squared
error is calculated as the prediction error of the samples that were not included in boot-
strapped selection for each tree, and therefore are samples that the model has not been
trained on (Liaw and Wiener (2002)). For our last analysis, we built a random forest
model that included one global predictor, one regional predictor, and one state predic-
tor. To test the efficacy of this model, we also simulated 1000 null models, where each
subject’s true energy on every trial in R with the randomForest package (https://cran.r-
project.org/web/packages/randomForest/randomForest.pdf) (Liaw and Wiener (2002)). Five-
hundred trees were used with $mtry = 1$ for each model.

Data and Software Availability

Code for simulations and select metrics is available at https://github.com/jastiso/NetworkControl.
Data will be made available upon request.
Supporting information

Figure S5.1: Effects of Different Parameter Choices in Open Loop Control. A: Effect of the Number of Time Steps Used: (i) For a representative subject, correlation coefficients between the empirical post-stimulation state and the simulated state for every time step in the simulation. Colored lines represent different trials. (ii) The maximum correlation coefficient values for empirical white matter connectivity for every subject when simulations were run with two different numbers of time steps: 750 time steps and 950 time steps. (iii, iv) Here we show the results of performing the same analysis displayed in panel (ii) but for topological and spatial null models. B: The effect of the scaling parameter $\beta$ on the correlations observed between the true post-stimulation state and the predicted post-stimulation state. Here we show the maximum correlations reached as a function of the $\beta$ value used in the model. In the main manuscript, we report results for $\beta = 1$, while here in the supplement we report qualitatively similar results for $\beta = 0.1$. We first observe that empirical graphs reach a maximum correlation that is significantly different from zero ($N = 11$, $t$-test $t = 5.07$, $p = 4.83 \times 10^{-4}$). In the topological null model, we observe smaller maximum correlations (paired $t$-test: $N = 11$, $t = 4.76$, $p = 7.64 \times 10^{-4}$), and earlier peaks ($N = 11$, $t = 7.40$, $p = 2.33 \times 10^{-5}$). Similarly in the spatial null model, we observe smaller maximum correlations ($N = 11$, $t = 4.10$, $p = 2.15 \times 10^{-3}$), and earlier peaks ($N = 11$, $t = 3.41$, $p = 6.70 \times 10^{-3}$).
Figure S5.2: Alternative Ways to Operationalize Brain States and Dynamic Properties. A: Results for simulations in which brain state represents power in single frequency bands. (i) Results for paired \( t \)-tests comparing the maximum correlation observed between the true post-stimulation state and the predicted post-stimulation state, for the spatial null model (top) and the topological null model (bottom) for each of the 8 frequency bands (x-axis). (ii) Results for paired \( t \)-tests comparing the time to inflection for the spatial null model (top) and the topological null model (bottom) for each of the 8 frequency bands (x-axis). In both panels (A) and (B), we display the associated p-values in the color of each square; values that pass Bonferroni correction for multiple comparisons are indicated with an asterisk. B: Time to Maximum (rather than peak) Correlation: Box plots depicting the average time to reach the maximum correlation between the empirically observed post-stimulation state and the predicted post-stimulation state at every time point in the simulated trajectory \( x(t) \). Box plots indicate the median and quartiles of the data. We find a main effect for graph type (repeated measures ANOVA \( F(2,20) = 21.84, p = 9.34 \times 10^{-6} \)). We also find that the empirical networks peak significantly later than both the topological (paired \( t \)-test: \( N = 11, t = 5.82, p = 1.69 \times 10^{-4} \)) and spatial (paired \( t \)-test: \( N = 11, t = 3.49, p = 5.79 \times 10^{-3} \)) null models. C: Statistics for alternate atlas and within subject analyses: All analyses from the main text were replicated when using a smaller parcellation of the Lausanne (\( N = 463 \)) atlas. Analyses from the main text were also repeated within each dataset (subject and stimulation site combination) and across trials. The table gives the minimum number of trials in the test, an effect size threshold, a p-value threshold, and the number of datasets with larger effect sizes and smaller p-values than the given thresholds.
<table>
<thead>
<tr>
<th>Subject</th>
<th>Age at Implant</th>
<th>Gender</th>
<th>Handedness</th>
<th>Years of Education</th>
<th>Number of electrodes</th>
<th>Unique Stimulation Bipolar Pair ROIs</th>
<th>Stimulation Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>Male</td>
<td>Left</td>
<td>16</td>
<td>109</td>
<td>LH_hippocampus</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>Female</td>
<td>Right</td>
<td>12</td>
<td>59</td>
<td>RH_hippocampus</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>Female</td>
<td>Right</td>
<td>12</td>
<td>146</td>
<td>LH isthmus, RH precuneus</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
<td>Male</td>
<td>Right</td>
<td>16</td>
<td>45</td>
<td>LH entorhinal, LH hippocampus</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>45</td>
<td>Female</td>
<td>Right</td>
<td>12</td>
<td>127</td>
<td>LH hippocampus</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>39</td>
<td>Female</td>
<td>Ambidextrous</td>
<td>16</td>
<td>75</td>
<td>LH hippocampus</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>Male</td>
<td>Left</td>
<td>12</td>
<td>128</td>
<td>LH hippocampus, LH parahippocampal</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>Female</td>
<td>Right</td>
<td>16</td>
<td>109</td>
<td>LH middletemporal, LH fusiform, LH inferior temporal</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>49</td>
<td>Female</td>
<td>Right</td>
<td>12</td>
<td>101</td>
<td>LH hippocampus, LH fusiform, LH middletemporal, LH amygdala</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>Male</td>
<td>Right</td>
<td>12</td>
<td>97</td>
<td>LH banks, LH superior temporal</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>Female</td>
<td>Right</td>
<td>12</td>
<td>109</td>
<td>LH superiorfrontal</td>
<td>1</td>
</tr>
</tbody>
</table>

Table S5.1: **Subject Information.** Demographic and task relevant information about each subject. Stimulation locations are given as unique regions in which either the anode or cathode were located. Stimulation sessions refers to the number of times that the task was run. Each session utilized a different set of stimulation electrodes.
Figure S5.3: Robustness and Selection of Parameters for Optimal Control

A: Impact of the Choice of S on the Relation Between Energy and Error. When S is the identity matrix, none of the 16 data sets showed a significant correlation between energy and error. (i) For one representative subject, we show the scatterplot of log(energy) versus log(error) when S is the identity matrix. At the single-subject level, we observed that when S contains zeros along the diagonal at ROIs without electrodes present (and thus only constrains the states of regions with ECoG electrodes), then 13 of the 16 data sets displayed a significant correlation between energy and error (Bonferroni corrected p < 0.003). (ii) For the same subject, we show the scatterplot of log(energy) versus log(error) when S is selected to only constrain the states of regions with ECoG electrodes in them. Based on these findings, we selected S to be the identity matrix.

B: Impact of Initial State Value on the Relation Between Energy and Error. We test whether these state values change the relative energies between trials within individuals by calculating the Pearson’s correlation coefficient between the energy from all possible pairs of state values. We find that all subjects showed highly significant correlations between results obtained with the 0 state value and results obtained with the 1 state value, as well as highly significant correlations between results obtained with the 0 state value and results obtained with the −1 state value, after Bonferroni correction for multiple comparisons (p < 0.001). Fourteen out of 16 data sets showed significant correlations between results obtained with the −1 state value and results obtained with the 1 state value. (i) For one representative subject, we show the scatterplot of log(energy) when the state value was set to −1, versus when it was set to 0. (ii) For the same subject, we also show the scatterplot of log(energy) when the state value was set to −1, versus when it was set to 1. iii For the same subject, we show the scatterplot of log(energy) when the state value was set to 0, versus when it was set to 1. We next tested whether the average magnitude of the energy for each state value was different across subjects. A one-way ANOVA returned no significant differences across state values (F(2, 45) = 0.93, p = 0.40) (iv) The average energy for each dataset for all state values (−1, 0, 1). (C: Method for Parameter Selection.) We began by calculating the error values of the simulations for every subject for each value of ρ, γ, and T. After selecting a range with small enough error (left-center), error values are z-scored to select only parameter combinations with a z-score less than −1 (center-right). The center coordinate from this latter space was selected to dictate the final parameter choice values: γ = 4, T = 0.7, and ρ = 0.3.
Figure S5.4: Relationships Between Energy and Distance in Topological and Spatial Null Models. 

A Relationship between distance and energy for simulations using topological null model graphs, with (right) Frobenius norm distance \( (p = 7.5 \times 10^{-16}) \), (middle) initial probability of being in a good memory state \( (p = 3.4 \times 10^{-10}) \), and (left) change in probability of being in a good memory state \( (p = 6.1 \times 10^{-4}) \), resulting from stimulation. All statistics used linear mixed effects models. 

B Relationship between distance and energy for simulations using spatial null model graphs, with (right) Frobenius norm distance \( (p = 6.9 \times 10^{-13}) \), (middle) initial probability of being in a good memory state \( (p = 1.5 \times 10^{-11}) \), and (left) change in probability of being in a good memory state \( (p = 0.0017) \), resulting from stimulation. 

C Summary statistics obtained by repeating these analyses with two different sets of parameters for the optimal control model: the set of parameter values \( \gamma = 7, T = 0.4, \) and \( \rho = 0.1 \), and the set of parameter values \( \gamma = 3, T = 0.9, \) and \( \rho = 0.5 \).
Figure S5.5: Additional Spatially Embedded Null Models. (A: Effects on Average Energy Requirements) Before exploring the energy used by different graph models, it is important to note that the numerical error returned for every network was consistently less than 0.004, indicating that the trajectories were highly accurate. Across graph types, we found a significant difference in the energy required for the transition from the initial state to the good memory state (repeated measures ANOVA $F(4, 40) = 12.96$, $p = 7.42 \times 10^{-7}$). In post-hoc analyses, we observed that neither of the additional null models were significantly different from those observed in the empirical graphs (paired $t$-test for weight preserving: $N = 11$, $t = 0.071$, $p = 0.945$; strength sequence preserving: $N = 11$, $t = -1.71$, $p = 0.12$). Average energy required for each transition from the pre-stimulation state to a good memory state, as theoretically predicted from Eq. 4.1 where $A$ is (i) the empirical networks (purple) estimated from the diffusion imaging data, (ii) the topological null model graphs (black), (iii) the spatial null model graphs (dark charcoal), (iv) the spatial strength distribution preserving null model graphs (medium charcoal), and (v) the spatial strength sequence preserving null model graphs (white). (B: The relationship between the determinant ratio and the energy required for the transition from the pre-stimulation state to a good memory state.) Additionally, across all graphs, we still observe a significant relationship between energy and the determinant ratio (linear mixed effects model $N = 11$, $\chi^2 = 33.2$, $p = 8.34 \times 10^{-9}$). Note: The color scheme is identical to that used in panel (A). (C: Visualization of Null Models.) Visualizations of empirical connectivity matrices (top left) and all null models (random, spatial, strength sequence preserving, and degree distribution preserving). Color bar gives the connection weight (QA).
Figure S5.6: **Additional Analyses Related to Regional Topology.** (A: Additional Control Metrics Associated With Energy in the Alpha Band.) A heat map of the significance of the relationship between each regional metric and input energy (estimated with a linear mixed effects model) for each frequency band. (B: Change in Power Across Different Bands) The average difference in power between all initial states and target state for each frequency band. Note: the alpha band does not show strikingly greater difference, suggesting that the finding that all metrics are significant in the alpha band is not due to a difference in the magnitude of the transition. (C: Persistent and Transient Modal Controllability.) To select a cut-off for persistent and transient modal controllability calculations, we plotted the persistent (Left) and transient (Right) controllability value for every node as a function of threshold on the eigenvalues of the adjacency matrix $A$. Here we show the mean (black) and standard deviation across nodes (grey) from one representative, and randomly chosen, subject. A threshold of 10% was chosen as a value that would provide sufficient variance and measure a small number of modes.
Figure S5.7: **Electrode Coverage:** Visualization of each subject’s electrodes mapped to MNI space. Blue electrodes were only used for recording; red electrodes were used for stimulation.
CHAPTER 6: Conclusion

In this body of work we sought to test underlying principles of neural dynamics via quantification of observations and theory-based models of perturbations. We first review the body of work identifying and testing principles of neural architecture formation, as a template for a body of work that identifies rules governing the behavior of neural systems (Ch. 2). We then move on to empirical studies investigating properties of neural dynamics in response to perturbations. All of models used here were network models, which represent a system as a set of discrete units and the interactions between them. We then model data from perturbations that were both endogenous (Ch. 3 and Ch. 4) and exogenous (Ch. 5), and both pathological (Ch. 3) and non-pathological (Ch. 4 and Ch. 5). We first provide a detailed description of changes in functional interactions associated with a specific waveform seen in epilepsy and IED. We then describe the findings of 2 projects testing theory-based models that rely on the principle that endogenous activity and perturbations spread along connections to change the state of the system over time. We ultimately show the usefulness of this model for understanding learning, and predicting the efficacy of stimulation. We also lay the groundwork for future studies incorporating more principles and different perturbations.

Limitations and Methodological considerations

The broad conclusions and impact of the included studies should be interpreted in light of the limitations of this approach. The more theory-based projects (Chapter 4 and 5) include populations who have undergone extensive stimulation regimes or BCI training, which reduces the potential sample size for a given study. While we rigorously tested all results against null models with appropriate statistical tests, the small sample sizes could lead to spurious findings (Munafò et al. (2017)). Positive and negative group-level effects should be validated with future studies in an independent sample, or with meta-analyses of available data.

One potential limitation of our theory-based studies was the assumption that activity
spreads across connections in a linear, noise free manner. While neural dynamics at small scales are highly non-linear, dynamics on the macroscale can be fit well with linear models (Honey et al. (2007); Nozari et al. (2020)). Additionally, tools assuming linear dynamics have been shown to capture meaningful features of non-linear oscillator systems (Muldoon et al. (2016)). However, the application of non-linear control theory to neural systems, or the explicit testing of which order of functions are required to approximate the system are exciting and promising paths for future advancement of the theory that activity in the brain is constrained to spread along its network of connections (Zanudo et al. (2017)).

Insights from data-based models

In this work, we use data-based models to quantify changes to functional connectivity, quantify paths along functional and structural connections, and support theory-based models. Each unique use and context of data-based models brought unique insights about the nature of neural dynamics. Specifically, from changes in functional connectivity that are associated with IEDs, we learned that IEDs tend to occur with increases in functional connectivity. Current state of the art models of epilepsy do not recreate these observations (Courtiol et al. (2020)) and therefore this work adds a new benchmark for future modeling studies. Additionally, the wide variety of data-based models used allows us to conclude that changes in functional connectivity are associated with the amplitude, not the phase of signals, that high and low frequency activity are impacted differently, and that most changes are associated with the first IED in a sequence, and not subsequent IEDs. These additional insights also posit new dimensions of inquiry for future data- or theory-based modeling of IEDs.

We learn from quantifying paths that one subset of regularized functional connections supports BCI learning in our cohort. This subset is not the subset that most strongly tracks performance, indicating separable processes for performance and learning (Shelton and Gabrieli (2004)). Additionally, this subset of connections shows variable spatial and temporal properties across individuals that are not associated with learning. Here, our data-based models revealed associations between neural dynamics and behavior through a
minimally constrained data-driven process, but was not able to posit theory regarding how this set of connections supported learning. To fill this gap, we then use this data-based model to represent the set of possible paths by which activity can spread for theory-based modeling.

Insights from theory-based models

We used theory-based models to test one principle of neural dynamics in different contexts - activity and perturbations are constrained to spread linearly along connections. We use this model to better understand the response of neural systems to both endogenous (BCI control) and exogenous (direct electrical stimulation) perturbations. In the case of BCI control, we learn that subsets of connections associated with learning are especially well suited to support sustained attention according to this model, providing a potential explanation for why these subgraphs are important that could be difficult to identify from the purely data-based approach. This observation is consistent with behavioral reports placing attentional behavioral traits as a predictor of BCI learning (Jeunet et al. (2016)), and with further development could inform a neural test of BCI literacy before investing time into training.

We also explicitly quantify the extent to which this model explains neural dynamics in response to electrical stimulation. The total variance explained is low, indicating a need for refinement and additional principles. However, we find more variance explained than one would expect using null models of neural connections, indicating that the system’s response to stimulation is constrained to spread along white matter tracts. In the electrical stimulation example, we also use our model to make explicit, testable predictions for future experiments. We posit predictions that span both individual and group levels, regarding neural activity, global connectivity, and local connectivity. Some of these predictions can be assessed without the use of the theory-based model, and rely on data-based modeling for quantification of different properties.
Figure 6.1: **Cooperation between models.** When beginning research into a complex system, both theory- and data-based models are needed to derive principles. Observations from models at the data-based end of the spectrum are used to inform and validate theory-based models, and theory-based models can posit predictions and hypotheses that can be tested in data-based models. Work from this thesis samples different parts of this cycle, with Chapter 2 reviewing currently understood principles of connectome formation, Chapters 3 and 4 using data-based models to generate observations, and Chapters 4 and 5 using theory-based models to generate predictions.

Need for a spectrum of models

Understanding the underlying principles of complex systems involves continual cross-talk between models at different levels (Fig. 6.1). This process is often thought of as a progression from complexity to simplicity (Gao and Ganguli (2015)). Early models report and model parts of the system at levels of detail, and continual iteration and refinement of theory reveals which parts of the system are important for modeling the behavior of interest (Gao and Ganguli (2015)). It is important to note explicitly that this process of refinement can be advanced by models on all parts of the continuum. This process is exemplified in this work, which demonstrates the types of conclusions and advancements that can be drawn from models at different levels of theory.
Future directions
These studies lay the groundwork for future research directions to characterize neural dynamics and test the extent to which connections constrain those dynamics. One immediate future direction is the application of control-theoretic models to other types of perturbations. Concepts from network control theory have already been used to understand the effects of transcranial magnetic stimulation (Medaglia et al. (2018); Beynel et al. (2020)), though less work has examined the dynamical characteristics of pathological perturbations such as IEDs. This model could be used to test if spread along white matter connections recreates the observed increases in functional connectivity following IEDs, or if more principles are needed to recreate that observed behavior.

Another promising future direction is the testing of how connections constrain activity at different evolutionary scales. The theory implies that connections should constrain activity in all neural systems, not just humans. Some evidence for the evolutionary extent of this principle is evident through applications of network control theory to other systems. For example, it has been shown that across the evolutionary hierarchy, connectomes become better suited for energetically favorable control (Kim et al. (2018a)). Additionally, the function of neurons in C. elegans was inferred using insights from control theoretic models (Yan et al. (2017b)). Additionally, the brain states involved in habit learning in nonhuman primates (Szymula et al. (2020)) and gene coexpression changes during addiction in mice (Brynildsen et al. (2020)) are explained by the energetic properties of different regions. However, a direct quantification of the extent to which these models account for changes in neural dynamics has not been done in other species.

Lastly, extensions of the network control model could help researchers understand other principles in the context of activity spreading along connections. For example, connections could be made to vary in time to test theories of plasticity or neural development. Alternatively, the development of non-linear models or neural dynamics obtained from data would allow for the application of tools from non-linear control theory (Zanudo et al. (2017);
Hocker et al. (2017)). This extension would enable researchers to study not only how activity is constrained to move between different patterns of activity (Cornblath et al. (2020)), but to move between different manifolds of time evolving activity.

Here, we add observations and predictions to the body of literature characterizing neural dynamics in epilepsy and during stimulation, and posit a guiding principle that activity is constrained to spread along white matter tracts. Pursuing some of the studies detailed above would better define the extent and limitations of the constraint for activity to spread along connections.
APPENDIX

Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field (Dion et al. (2018); Caplar et al. (2017); Maliniak et al. (2013); Dworkin et al. (2020)). Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al. (2020); Zhou et al. (2020)). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 7.86% woman(first)/woman(last), 13.69% man/woman, 20.36% woman/man, and 58.09% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (Ambekar et al. (2009); Sood and Laohaprapanon (2018)). By this measure (and excluding self-citations), our references contain 10.28% author of color (first)/author of color (last), 13.79% white author/author of color, 22.14% author of color/white author, and 53.8% white author/white author. This method is limited in that a) names and Florida Voter Data to make the predictions may not be indicative of racial/ethnic identity, and b) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.


D. S. Bassett, C. H. Xia, and T. D. Satterthwaite. Understanding the emergence of neu-


networks: Accounting for the cost of wiring. *Network Neuroscience*, 2016c. ISSN 2472-1751.


G. Buzsáki and E. I. Moser. Memory, navigation and theta rhythm in the hippocampal-


K. W. Doron, D. S. Bassett, and M. S. Gazzaniga. Dynamic network structure of inter-


H. Glowinski. Cognitive deficits in temporal lobe epilepsy: an investigation of memory


J. A. Henderson and P. A. Robinson. Relations between the geometry of cortical gyrification


226


A. Vuckovic and B. A. Osuagwu. Using a motor imagery questionnaire to estimate the performance of a Brain-Computer Interface based on object oriented motor imagery.


