Thesis

Long term brain dynamics extend cognitive neuroscience to timescales relevant for health and physiology

Maxwell B. Wang

Committee

Avniel Ghuman (Chair, University of Pittsburgh)

Robert Kass (Chair, Carnegie Mellon University)

Timothy Verstynen (Carnegie Mellon University)

Russell Schwartz (Carnegie Mellon University)

Mark Richardson (Massachusetts General Hospital)

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Program of Neural Computation and the Department of Machine Learning at Carnegie Mellon University

Pittsburgh, PA

April 2023

Acknowledgments

I would first like the thank all of the patients of Pittsburgh and the surrounding areas that graciously agreed to participate in the various research endeavors taking place in this city. Without them, much of the work we do at both Carnegie Mellon and the University of Pittsburgh would not have been possible. Ultimately, we do all of this for them.

I would also thank the physicians and staff in the University of Pittsburgh Comprehensive Epilepsy Center at the University of Pittsburgh Medical Center, particularly Cheryl Plummer and Anto Bagic, without whom none of these data exist.

I would also like to thank the members of my dissertation committee for their time, patience, and intellectual contributions to my development. Rob Kass for being the unashamedly rigorous statistician he is. Mark Richardson for his very valued mentorship on how to learn to be both a good physician and scientist. Timothy Verstynen for his perspectives as a talented psychologist that pushed me to really elucidate the fundamental neuroscience questions underlying this body of work. Russell Schwartz for his mentorship throughout my journey in the MSTP program.

I especially would like the thank my thesis advisor, Avniel Ghuman, a passionate physicist studying the brain. At the time when I joined his lab, it was primarily a group focused on studying on how our brains process sight. As someone at the time who did not have a particularly strong interest in vision (which has changed during my time here), I am particularly thankful for him creating a place for me to explore a large range of ideas about the brain even if they did not directly align with the lab's stated goal at the

time. Before coming to his lab, I had spent most of my life studying and thinking as an engineer and applied mathematician. He was the one who taught me to be a scientist.

I'd also like to thank the other current and previous members of the lab, Arish, Mike, Yuanning, Matt, Katie, Jhair, and David for putting up with me occasionally crashing our server and being a general hazard to any well-functioning organization.

I would also like to thank my mentors before graduate school. Tom Anderson as the one who first taught me calculus. Who stayed on the phone with me for hours as he walked me through my mistakes. Lawrence Bush for getting me interested in engineering. ShiNung Ching for first getting me interested in the brain. Pratik Mukherjee who first got me interested in medicine. Howard Aizenstein who got me interested in interventional work.

I am extremely grateful to the Hertz Foundation and community for their support during graduate school. It has been a marvelous time to get to know the other fellows and to share ideas between people from a large range of disciplines across math, engineering, the physical sciences, and biology. I am especially thankful to Megan Blewett for her continuing advice as well as the good friends I have made in the program: Bailey, Ben, Hannah, Katherine, Maya, Lilly, Vikram, Sasha, Alex, Cole, and more.

I'd also like to thank my friends for their support during these past four years. K, Sean, Akasha, Oz, Jared, and John from my time in St. Louis. Allen, Ananya, and Zach in the Langsdorf community. Philip and Jeff particularly in the MSTP program. Jackie, Song, Bethany, Jinman, Ying, and Adam from medical school.

And finally I'd like to thank my family. My parents for being a continuing source of unfiltered advice and support. Jeffrey for being the inspiration he is. Allison for keeping me humble.

Contents

Abstract

In medicine, we define and treat diseases based on their causes. We classify infections as viral, fungal, parasitic, or bacterial with specialized treatments for each class. We define and treat tumors according to which genetic abnormalities allow them to proliferate uncontrollably. This tenet underlies almost every field of medicine except for the brain where many diseases are largely defined by their symptoms. As a result, with a few notable exceptions, we define treatments of the brain around which symptoms to use them for rather than what root cause they solve. But what if we could peer inside somebody's brain, identify the pathological circuits and activity that drives a patient's disease, and then gear our treatment towards that? While past attempts at this have shown initial promise, they have been limited by small sample sizes and difficulty in selecting appropriate study populations. In this thesis, I explore how both these problems can be addressed by a paradigm shift to study neural activity over very long timescales.

In medicine, we define and treat diseases based on their causes. We classify infections as viral, fungal, parasitic, or bacterial with specialized treatments for each class. We define and treat tumors according to which genetic abnormalities allow them to proliferate uncontrollably. This tenet underlies almost every field of medicine except for the brain where many diseases are largely defined by their symptoms. As a result, with a few notable exceptions, we define treatments of the brain around which symptoms to use them for rather than what root cause they solve. But what if we could peer inside somebody's brain, identify the pathological circuits and activity that drives a patient's disease, and then gear our treatment towards that? While past attempts at this have shown initial promise, they have been limited by small sample sizes and difficulty in selecting appropriate study populations. In this thesis, I explore how both these problems can be addressed by a paradigm shift to study neural activity over very long timescales.

CHAPTER 0

Introduction

The idea of identifying the anomalous brain activity driving a patient's disease arguably began in 1934-1935 when Fisher, Lowenbach, Gibbs, Davis, and Lennox used the recently invented electroencephalogram to describe neural spike waves causing epileptic seizures [1]. Their discovery sparked several revolutions in the fields of neurology and neurosurgery towards specifically identifying and targeting different kinds of neural activity, primarily within epilepsy and movement disorders with very recent advances beginning to be made in a few psychiatric conditions. These discoveries have been made possible by rapid development in several technologies ranging from more effective ways to record neural activity to therapies capable of precisely targeting and modulating it.

However, many diseases affecting the brain remain primarily defined as a collection of symptoms either in isolation or due to a root cause that is several steps removed from the clinical presentation. Major depressive disorder is defined as possessing five of eleven symptoms such as depressed mood, loss of energy, insomnia, trouble concentrating, and more [2]. Traumatic brain injuries are defined as the appearance of neurological symptoms such as loss of consciousness or confusion following a physical trauma to the head [3]. Alzheimer's disease is defined as progressive deterioration of cognition and memory we believe is related to amyloid and/or tau depositions [4].

What is missing from all of these diseases is the capability to look at an individual

patient and identify which abnormal neural activity in their brain is leading to the symptoms disrupting their life. We may understand on the cellular level how individual neurons degrade during these diseases, but how does that translate into the macroscopic picture of a brain consisting of nearly a 100 billion neurons losing specific functions? As a result, our understanding of these diseases is not unlike where cancer research was a few decades ago when we knew many cancers were caused by smoking, toxins, or radiation through some form of genetic mutation but were unable to readily detect which genetic mutations were driving a patient's tumor. But once we developed the capability to detect these mutations, they provided a new avenue of therapeutic targets that revolutionized the field of oncology. In the brain, we are already developing a host of new technologies to precisely modulate and control the activity of different neural populations, but where in the brain should we point them?

Artificial intelligence has been proposed as one piece of solving this puzzle due to its strengths in pattern recognition that have been demonstrated in many non-medical fields. For disorders of the brain, machine learning has seen increasing use over the past several years in interpreting functional neuroimaging or intracranial recordings to diagnose and classify the severity of various psychiatric disorders and predicting treatment response [5–7]. It is starting to be used as a way to control closed-loop brain stimulation [8, 9]. It is beginning to show efficacy in engineering new ways to treat depression [10, 11].

Despite these advances, a number of key challenges remain, one of the largest of which is extremely small sample sizes. Modern machine learning is typically performed on datasets that are orders of magnitude larger than the samples we conventionally use when studying brain activity. One of the largest leaps forward in computer vision occurred with the publication of the ImageNet database which originally contained over a million images when it was first showcased in 2009 [12]. AlphaFold was trained on over 170,000 protein structures [13]. The natural language processor GPT-3 was trained on roughly a billion pages of text [14]. In comparison, the average sample size of a functional MRI

paper published in highly-cited journals from 2017-2018 was 24 subjects [15]. If we define our analyses as "obtain a sample of brain activity from patients presenting with some disease and compare their activity to healthy controls", it will be highly non-trivial to obtain hundreds of thousands of samples.

As a result, when analyzing neural data, we typically limit ourselves to simpler, smaller models that are orders of magnitude less complex than what is commonly used in other non-medical fields. And while we have made many remarkable discoveries in neuroscience and medicine using these methods, this limitation will become a fundamental barrier if not one we have reached already.

A second key challenge is in selecting the populations that go into these studies. If we wanted to, for example, study "neural activity associated with major depressive disorder", a commonly used method would be to select a population of individuals meeting a diagnosis for it and then compare their neural activity to healthy controls. Training an algorithm to differentiate between the two is a logical method for asking "what neural signatures are present in the vast majority of people with depressive symptoms", a valuable line of inquiry that we can learn much from. But it is also a method that is prone to overlooking heterogeneity in these populations and the very possible situation that what we call "major depressive disorder" is in reality several different diseases masquerading with a very similar set of symptoms.

In this thesis, I argue that both of these problems can be elegantly solved by a paradigm shift in cognitive neuroscience to study brain dynamics in a space orders of magnitude slower than what has been conventionally done. I start by presenting two examples of efforts that fall within these two "key challenges". One project involving depression and another involving movement disorders. Two projects that were able to uncover interesting information about these diseases but also highlighted these fundamental limitations. I end by presenting two examples of how to study long term brain dynamics in natural situations in order to demonstrate how this new approach can be implemented.

References

[1] F. A. GIBBS, H. DAVIS, and W. G. LENNOX. The electro-encephalogram in epilepsy and in conditions of impaired consciousness. *Archives of Neurology & Psychiatry*, 34(6):1133–1148, 12 1935.

[2] Rudolf Uher, Jennifer L Payne, Barbara Pavlova, and Roy H Perlis. Major depressive disorder in dsm-5: Implications for clinical practice and research of changes from dsm-iv. *Depression and anxiety*, 31(6):459–471, 2014.

[3] David K Menon, Karen Schwab, David W Wright, Andrew I Maas, et al. Position statement: definition of traumatic brain injury. *Archives of physical medicine and rehabilitation*, 91(11):1637–1640, 2010.

[4] Bruno Dubois, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, et al. Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria. *The Lancet Neurology*, 13(6):614–629, 2014.

[5] Meenal J Patel, Alexander Khalaf, and Howard J Aizenstein. Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*, 10:115–123, 2016.

[6] Shuang Gao, Vince D Calhoun, and Jing Sui. Machine learning in major depression: From classification to treatment outcome prediction. *CNS neuroscience & therapeutics*, 24(11):1037–1052, 2018.

[7] Jiayang Xiao, Nicole R Provenza, Joseph Asfouri, John Myers, Raissa K Mathura, Brian Metzger, Joshua A Adkinson, Anusha B Allawala, Victoria Pirtle, Denise Oswalt, et al. Decoding depression severity from intracranial neural activity. *Biological Psychiatry*, 2023.

[8] Brady Houston, Margaret Thompson, Andrew Ko, and Howard Chizeck. A machine-

learning approach to volitional control of a closed-loop deep brain stimulation system. *Journal of neural engineering*, 16(1):016004, 2019.

[9] Qitong Gao, Michael Naumann, Ilija Jovanov, Vuk Lesi, Karthik Kamaravelu, Warren M Grill, and Miroslav Pajic. Model-based design of closed loop deep brain stimulation controller using reinforcement learning. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 108–118. IEEE, 2020.

[10] Katherine W Scangos, Ghassan S Makhoul, Leo P Sugrue, Edward F Chang, and Andrew D Krystal. State-dependent responses to intracranial brain stimulation in a patient with depression. *Nature medicine*, 27(2):229–231, 2021. Publisher: Nature Publishing Group.

[11] Sameer A Sheth, Kelly R Bijanki, Brian Metzger, Anusha Allawala, Victoria Pirtle, Joshua A Adkinson, John Myers, Raissa K Mathura, Denise Oswalt, and Evangelia Tsolaki. Deep brain stimulation for depression informed by intracranial recordings. *Biological Psychiatry*, 92(3):246–251, 2022. Publisher: Elsevier.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Denes Szucs and John PA Ioannidis. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage*, 221:117164, 2020.

CHAPTER 1

Acute trajectories of neural activation predict remission to pharmacotherapy in late-life depression

Pharmacological treatment of major depressive disorder (MDD) typically involves a lengthy trial and error process to identify an effective intervention since most antidepressants have a $\sim 50\%$ success rate but require two months to determine their efficacy. Previous work by the group demonstrated that changes in neural activation during resting state and emotional regulation tasks throughout a 12-week trial of venlafaxine (a commonly used antidepressant) could be measured using functional MRI. Here I showed that changes occurring within the first day of treatment, particularly within the cingulate and frontal cortex, could be used to predict the overall trials' efficacy. I also found that fMRI scans taken before treatment initiation could predict treatment response in isolation, albeit at slightly worse accuracy. Taken together, this demonstrates the potential utility of classifying patients with depression into subtypes based on their neural signature when parsing through potential treatment options.

1.1 Introduction

In late-life depression (LLD), the time between initiating treatment and clinical response generally takes 4-6 weeks. This delayed clinical effect is associated with prolonged suffering,

exacerbated medical comorbidities, and increased risk of suicide [1]. While many studies report changes in neural activation in depressed older adults during emotional reactivity or regulation tasks, relatively few studies have focused on the predictive utility of the identified changes or described changes within a timeframe that permits this information to be used clinically.

There is emerging evidence that neural markers may have predictive capacity toward reducing the number of trialed antidepressants and possibly improving antidepressant outcomes [2, 3]. In a subset of the sample used in this chapter, previous members in the group identified neural changes within a day of the first dose of antidepressants utilizing functional magnetic resonance imaging (fMRI), which was dependent on remission status [4, 5]. These early responses indicate that while behavioral changes often take weeks to manifest, a patient's underlying neural activity is quickly changed by antidepressant treatment in a detectable fashion.

In this study, I investigated the treatment response predictive capacity of three neural markers: activation during an emotion reactivity task, activation during an explicit emotion regulation task, and whole brain voxel-wise connectivity (eigenvector centrality) at rest in a sample of LLD participants (N = 49) receiving venlafaxine, a commonly used antidepressant. I tested the predictive capacity of pre-treatment neural activation and of the change in neural activation following a single dose of venlafaxine. I compared the predictive capacity of these markers (both separately and in unison) to the predictive capacity of baseline depression severity. At the time, this was one of few studies that had investigated the predictive utility of acute change in neural activation for treatment remission in LLD. Our group hypothesized that these markers would have greater predictive capacity than baseline depression severity and that early changes in these tasks would improve our predictive capacity more than pre-treatment neural markers alone.

1.2 Methods

1.2.1 Study design and participants

The Aizenstein group collected neuroimaging data as part of a larger 5-year multi-site study of treatment in LLD that collected neuroimaging data at one site (Pittsburgh, USA). Participants were recruited and prescribed with open-label venlafaxine (a serotonin and norepinephrine reuptake inhibitor). Participants were included if they were at least 50 years old, met Diagnostic and Statistical Manual for Mental Disorders-IV (SCID-IV) criteria for major depressive disorder (MDD) and had a Montgomery-Asberg Depression Rating Scale (MADRS) score of 15 or higher at baseline. Participants were excluded if they had a history of mania or psychosis, alcohol or substance abuse (within last 3 months), dementia or neurodegenerative disease as well as conditions with known effects on mood and cognition (e.g. stroke, multiple sclerosis, vasculitis, significant head trauma, and/or unstable hypertension). Informed consent was obtained from all participants prior to engaging in any research procedures, and the University of Pittsburgh Institutional Review Board approved this study.

All MRI scanning was conducted in the morning. Five MRI scans were collected during the treatment trial. Participants came in on the first day for a baseline scan (no medication). In the same evening, they were given a placebo, after which they returned the next day for another scan (placebo scan). The evening of that scan, they were given their first dose of venlafaxine (37.5 mg), after which they returned the next day for another scan (single dose scan). They continued their medication for one week and returned for another scan (week one scan). They returned a final time after the end of the treatment trial (12 weeks, end scan). This analysis does not utilize the week one or end of trial scans as we intended to understand the predictive capacity of the neuroimaging data over an acute period. Henceforth, we only describe the relevant scanning procedures and analyses. During the trial, participants returned for weekly or bi-weekly clinical visits

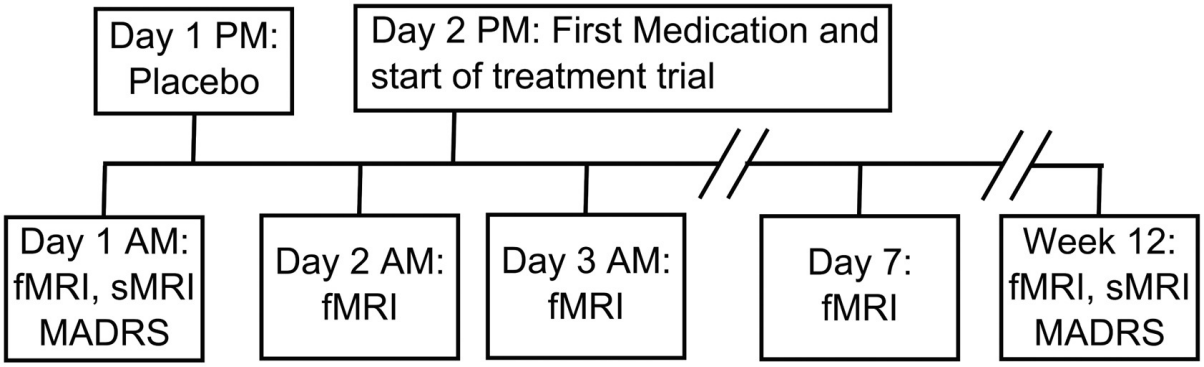

Figure 1.1: The study design protocol: Functional and structural magnetic resonance imaging (fMRI and sMRI, respectively) was performed in the morning during various times through the trial. On the first day, participants came in for an fMRI scan (baseline) and then were given a placebo following the scan. On the second day, they returned for another fMRI scan and then were started on venlafaxine following the scan. They returned the next day ($\sim$ 12h later) for another fMRI scan (1st dose change). They continued their medication as normal and came in for scans at the end of the first week and at the end of the trial. Only the fMRI scans at baseline and 1st dose change were used in this paper.

and the venlafaxine dosage was increased as necessary (up to a maximum of 150 mg/day by week 6). Participants who did not show signs of response by week 6 had the dosage increased up to a maximum of 300 mg/day. At the end of the study, participants were classified as remitters if they had a MADRS ≤ 10 for at least two weeks during the trial (and remained so until the end of the trial). Figure 1.1 summarizes the study timeline.

A total of 62 participants signed consent. Eleven were excluded due to: side effects of medication (N=2), non-adherence to protocol (N=2), inaccurate diagnosis of MDD (N=1), and inability to determine remission status due to lost/missing data (N=6). Among the remaining participants (N=51), two participants did not complete all MRI scanning but did complete the treatment trial. In summary, 49 participants were included in this analysis.

1.2.2 MRI protocols

All scanning was conducted at the University of Pittsburgh Medical Center on a Research dedicated 3 T Siemens Trio TIM scanner (Munich, Germany) using a 12-channel

head coil. The baseline and end scan protocol included both a structural and functional image, while other scans collected only functional sequences. In this manuscript, we limited our analysis to the functional sequences, which were a resting state sequence, an explicit emotion regulation task sequence, and an emotional reactivity (faces/shapes) sequence.

An axial, whole brain 3D magnetization prepared rapid gradient echo (MPRAGE) was collected with repetition time (TR) = 2300 ms, echo time (TE) = 3.43 ms, flip angle (FA) = 9 degrees, inversion time (TI) = 900 ms, field of view (FOV) = 256 × 224, 176 slices, 1 mm isotropic resolution and with Generalized Autocalibrating Partial Parallel Acquisition (GRAPPA) factor = 2. An axial, whole brain 2D fluid attenuated inversion recovery (FLAIR) was collected with TR = 9160 ms, TE = 90 ms, FA = 150 degrees, TI = 2500 ms, FOV = 256 × 212, 48 slices, and 1 × 1 × 3 mm resolution.

An axial, whole brain (excluding cerebellum) echo planar (EPI) T2*-weighted functional image was collected to measure the blood oxygen level dependent (BOLD) response with TR = 2000 ms, TE = 34 ms, FA = 90 degrees, FOV = 128 × 128, 28 slices, 2 × 2 × 4 mm resolution. The duration of the face/shapes task (see Functional Imaging Metrics) was 117 volumes (∼ 4 min), the explicit emotion regulation task (see Functional Imaging Metrics) was 270 volumes (∼ 9 min), and the resting state was 150 volumes (∼ 5 min). Due to variability in placement by MR technicians the coverage of the functional scans was in general limited to above the cerebellum and below the top aspect of the motor cortex (though this varied slightly between functional sequences). Participants were instructed to lie awake and view a cross hair during resting state.

1.2.3 Emotional reactivity task

The face/shapes task is widely used and has been found to robustly activate the amygdala [6, 7]. Participants were instructed to match either a face cue or a shape cue. A cue was shown on the top center of the screen and they were instructed to respond with an

MR-compatible glove (left or right index finger) by matching to one of two simultaneously presented faces. The facial expressions shown were either angry or fearful. During the shapes, they match a shape to one of simultaneously presented shapes. The shapes task (5 blocks) was interleaved with the faces task (4 blocks) and each block lasted 24 s containing 6 trials (4 s each). Before the beginning of each block participants were instructed visually to "match emotion" or "match form" (2 s). The face images are presented from a set of 12 different images (six per block, three of each sex) and are all derived from a standard set of pictures of facial affect. Stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools, Inc., Pittsburgh).

1.2.4 Explicit emotional regulation task

Participants were shown emotionally neutral or negative images from the standardized International Affective Picture System (IAPS) [8] and were instructed to either "Look" or "Decrease." This task has been described previously [5] and has been used to activate prefrontal cortex (especially the dorsolateral prefrontal cortex) as a means of explicitly regulating limbic reactivity. During the look instruction, participants were to view content naturally. During the decrease instruction, participants were instructed to reappraise negative images to actively alter the elicited emotion. A master level staff member instructed participants on how to reappraise prior to entering the scanner. After each image they were asked to rate how negatively they felt from 1 to 5. The neutral (11 events), negative (15 events), and negative regulate (15 events) conditions were interleaved and each event lasted 6 s. The inter-trial interval was 13 s with no jitter (though they were not locked to a TR). This allowed for modeling of each individual response by allowing for enough time in between each stimulus, but likely resulted in lower power to detect each individual effect. The images are presented from a set of images and stimulus presentation and responses were controlled using E-prime software (Psychology Software Tools, Inc., Pittsburgh).

1.2.5 Structural Processing

All processing was conducted using statistical parametric mapping (SPM12) in MATLAB (MATLAB 2016b, The MathWorks, Natick, 2016). Interpolation was conducted using 4th degree B-spline interpolation, normalized mutual information similarity metric for coregistration between images of different types, and mutual information similarity metric for motion correction unless otherwise stated. The FLAIR was coregistered to the MPRAGE (affine transform). Both images were input into a multi-spectral segmentation [9], which (after bias correction) segmented them into gray, white matter, cerebrospinal fluid, skull, soft-tissue, and air. Due to high white matter hyperintensity burden the number of Gaussians used to identify white matter was set to two (which improves the segmentation). This process generates a deformation field that can be used to normalize other images to a standard anatomic space (Montreal Neurological Institute, MNI) (Ashburner and Friston, 2005). An automatic mask for the intracranial volume was generated by thresholding the intracranial tissues with a probability of 0.1, filling the mask (imfill), and then performing a morphological closing operation (imclose, sphere of one voxel) in MATLAB. This mask (intracranial volume, ICV) was applied to the MPRAGE to remove non-brain tissues (which improves functional-structural coregistration).

1.2.6 BOLD pre-processing

The explicit emotion regulation task and the resting state data were slice time corrected (temporally middle slice was used as reference) prior to performing motion correction. All functional BOLD data was motion corrected (rigid coregistration to the mean), coregistered to the skull-stripped MPRAGE (mean functional image used to calculate affine transformation), normalized to MNI space using the deformation field calculated previously (2 mm isotropic resolution), and smoothed using a Gaussian kernel with FWHM of 8 mm. All images were investigated by human eye to confirm that coregistration and normalization steps were accurate.

Motion was evaluated using ArtRepair toolbox [10]. During the emotional faces reactivity task, participants had low maximum translations [mean = 1.26 mm (std = 1.21)], low root mean squared (RMS) [1.11 mm (0.81)], and low percentage of volumes displaying head jerks above 0.5 mm [6.2% (10.7%)]. During the resting state, participants had low maximum translations [1.27 mm (1.26)], low root mean squared (RMS) [1.04 mm (0.85)], and slightly higher percentage of volumes displaying head jerks above 0.5 mm [10.9% (19.9%)] that were corrected for using wavelet-despiking in later stages. During the explicit emotion regulation task, participants had low maximum translations [1.87 mm (1.91)], low root mean squared (RMS) [1.40 mm (1.08)], but slightly elevated percentage of volumes displaying head jerks above 0.5 mm [9.4% (30.8%)], with a few particularly bad cases that were removed. There were no group differences in any of these motion metrics between remitters and non-remitters between any time points.

For resting state BOLD, spike artifacts were removed using a previously established method that uses wavelets to filter spike artifacts [11]. Five principal components of white matter and cerebrospinal fluid were extracted as well as 6 motion parameters and a vector to model the mean of the time series [12]. Band-pass filtering was conducted by including several regressors that represented cosines with all discrete frequencies except those within the standard expected resting state frequencies (0.008 to 0.15 Hz).

1.2.7 Modeling task activation: emotion reactivity and emotion regulation tasks

Mass-univariate general linear modeling (i.e. each voxel is independently modeled) was performed to model the mean, faces task, shapes task, and six parameters of motion (from motion correction). The canonical hemodynamic response function was used to convolve the faces and shapes tasks to expected hemodynamic responses. A high-pass filter of 1/128 Hz was utilized to account for low frequency noise. An autoregressive [AR(1)] filter was used to account for serial correlations due to aliased biorhythms and unmodelled

activation. The contrast faces minus shapes was used to perform all voxel-wise analyses.

Similarly, the explicit emotion regulation task included similar parameters however it modeled the activation during the neutral and negative viewing tasks as well as the reappraisal task (during viewing of negative images). The contrast of interest was negative reappraise minus negative viewing, which modeled the activation during reappraisal adjusting for activation during the negative viewing task.

1.2.8 Resting state BOLD: eigenvector centrality

Eigenvector centrality was calculated using the fastECM toolbox [13]. Briefly, centrality is a measure of connectedness of a voxel or region. Mean centrality is a related measure that calculates the mean voxel-wise connectivity of a single voxel to all other voxels where a greater centrality would imply that a voxel is more widely connected. FastECM uses singular value decomposition to circumvent the calculation of large correlation matrices.

1.2.9 Response prediction

I used a combination of Principal Component Analysis (PCA), Least Angle Regression, and Logistic Classification to identify differences on the individual level in our fMRI features that could be linked to remission. Using SPM12, each individual's fMRI maps (resting state eigenvector centrality, emotional regulation, and emotional reactivity) were averaged across 116 regions in MNI152 space outlined in the Automatic Anatomical Labeling Atlas [14], resulting in a 348-length feature vector for each individual for a given time point. PCA allows for the estimation of principal component vectors across participants that vary together, thus it is likely that regions that activated similarly were combined into a single vector.

I tested two major fMRI feature vectors. One was the fMRI features (116 regions of activation during emotion reactivity, emotion regulation, or centrality) at baseline, prior to treatment. The other was the change in fMRI features following a single dose (or

placebo), which was defined as the difference between the feature vector after a single dose (or placebo) and baseline. Due to small-number error concerns, I chose this method over a percentage change metric.

All analyses were performed within a ten-fold cross-validation scheme to address over-fitting and multiple comparisons concerns. To avoid biasing our estimates, all data demeaning, dimensionality reduction, feature selection, and hyper-parameter optimization were performed via nested cross-validation loops. To establish bounds on the accuracy of our algorithm, we repeated the cross-validation scheme 30 times, each time redrawing the cross-validation folds.

I used a combination of PCA and Least Angle Regression to selectively reduce the dimensionality of the dataset to components that were relevant to remission [15]. Least Angle Regression has been proposed as a "less greedy" alternative to the popular LASSO [16] algorithm that favors the net contribution by multiple features simultaneously over identifying single features independently [17]. Using the components selected by these two algorithms, a logistic classifier was then trained on these components and used to predict remission on the test fold of the cross-validation scheme. Accuracy was assessed using Receiver Operator Curves (ROC) analysis.

To determine predictors that utilized information from multiple scanning time points or the baseline MADRS score, we averaged the predictors from each individual algorithm to generate an averaged predictor, a concept commonly referred to as an unweighted voting algorithm [18]. This procedure is meant to combine the predictive power of several feature sets without suffering from over-fitting concerns.

1.2.10 Permutation testing

To determine which anatomical regions of the fMRI metric maps leant themselves to accurate and reliable predictions of remission, we utilized permutation testing. More specifically, we randomly shuffled the remitter/non-remitter labels within our dataset and

recomputed the entire cross-validated classification pipeline. Each feature (a single region for a given fMRI feature) was then ranked on its relative importance to the classifier compared to all other regions across all metrics. This ranking procedure was repeated 1000 times and compared to the rankings found when using the "true" remission labels. A region/task pair is significantly associated with remission if its true ranking is within the top 5% of random permutation ranking trials.

To understand the relative contribution of the fMRI metrics to each prediction used in this paper, I also applied an additional permutation test to the feature sets themselves (i.e., instead of permuting remitter/non-remitter labels, the images themselves were permuted). For a given time point, I permuted the features from a single fMRI feature and repeated the cross-validation and training/evaluation procedure. This was repeated 1000 times for each fMRI feature at both time points to assess the individual contribution from each fMRI map.

1.3 Results

Figure 1.2 illustrates the accuracy of the fMRI classification algorithm using the fMRI scans collected at different time points using the Area under Curve (AUC) metric. Within this context, AUC represents the probability that given two participants, one remitter and one non-remitter, that the algorithm of interest will correctly classify the remitter as being more likely to have a positive treatment response. This ranges from 50% for a random-guessing algorithm to 100% for perfect accuracy. I found that utilizing our fMRI procedures and classification algorithm yielded an approximate 15% increase in AUC over that of simply using the MADRS alone. In general, I show that while there is no significant difference in AUC between baseline and change in fMRI features, the two features together significantly improve the overall AUC and adding baseline depression severity further improves it. For comparison, I also show the accuracies when utilizing the MADRS at one or two weeks after the trial start, as well as the composite accuracy of these values when

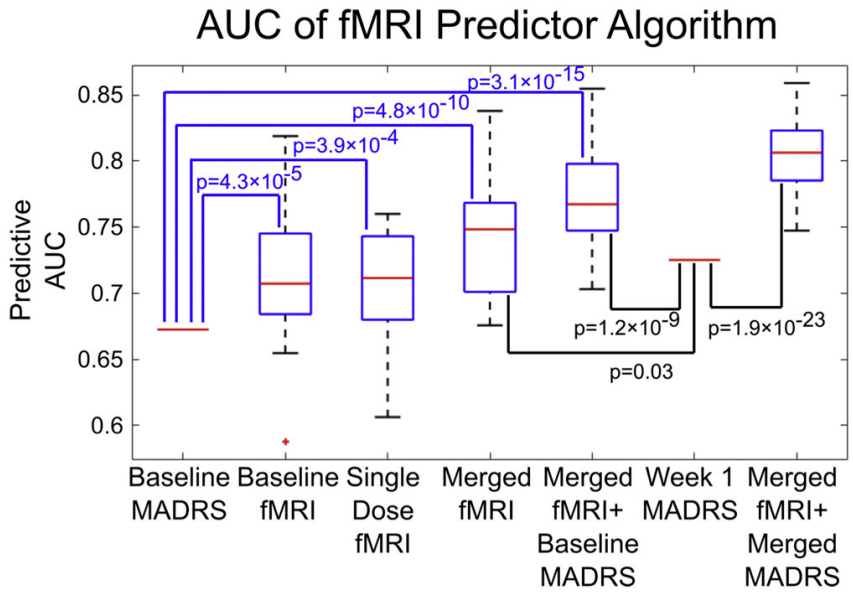

Figure 1.2: The predictive accuracy of remission among 49 subjects was determined using 30 trials of repeating a 10-fold cross-validation scheme and is shown via interquartile range boxplots. The second and third column represent the accuracy of using the classification algorithm on only the functional imaging data (resting state centrality, emotional reactivity task, and emotional regulation task) available at baseline or the change in imaging metric a day after the first dose of venlafaxine. The fourth column represents averaging the predictions from the second and third column, while the fifth column shows the accuracy from averaging the predictions from the first four columns. We find that utilizing functional imaging along with our proposed algorithms improves the predictive power of the MADRS questionnaire by 15% (other demographic variables such as age, sex, education level, and race had no significant predictive power and thus were not included). The last two columns show the accuracy of utilizing the MADRS at one week (change in MADRS was less accurate) and using that value in combination with the fMRI data at baseline and post-first dosage. p-Values were calculated as one-sample t-tests with a null hypothesis that the accuracy of the algorithm was equal to that to the MADRS at baseline or at one week.

used with the fMRI results. I found that our imaging approach significantly outperforms both these values by approximately 7% in AUC. The placebo minus baseline scan did not predict remission [median AUC of 0.56 (IQR, 0.52-0.6)] better than MADRS (p = 2.8e-11). Placebo minus baseline combined with the baseline prediction [median AUC of 0.63 (IQR, 0.68-0.71)] was not significantly better than MADRS (p = 0.688). Thus, the placebo results were excluded from further analysis.

Figure 1.3 shows the region/task pairs that passed permutation significance testing. As

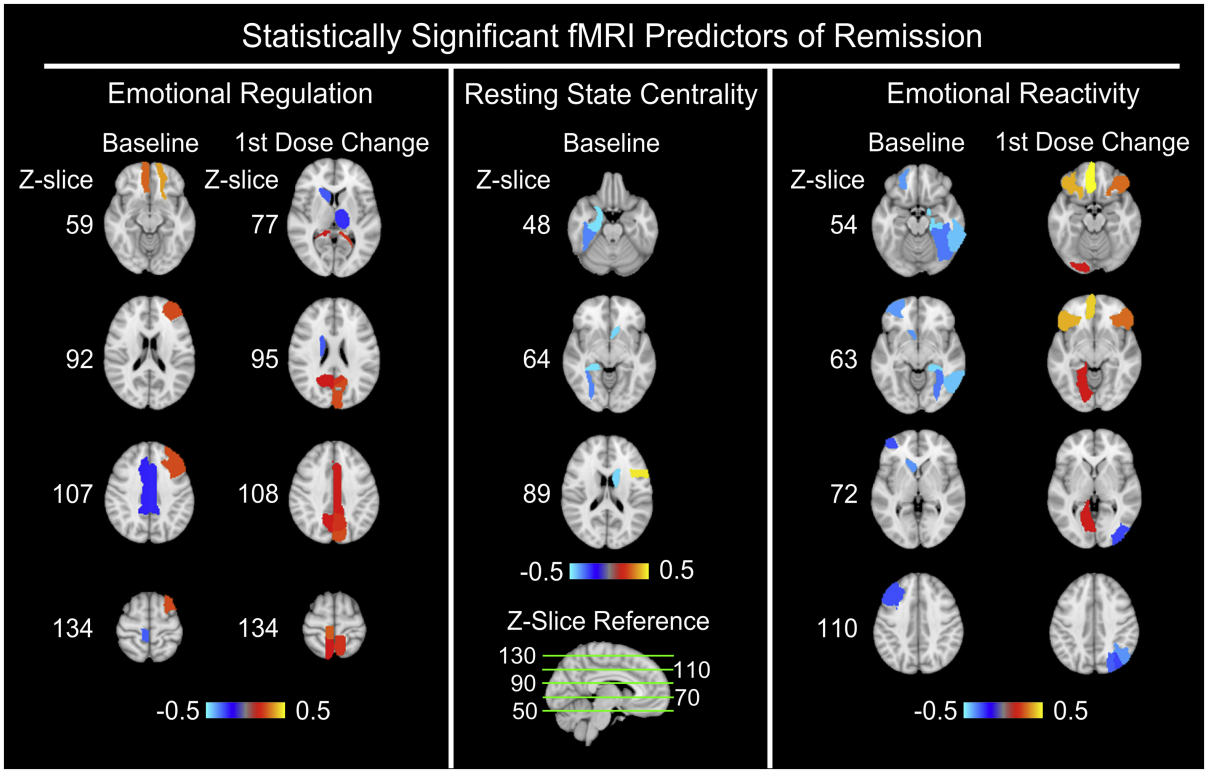

Figure 1.3: Axial slices of the relative importance of region/task pairs that passed statistical permutation significance testing (p = 0.05) are shown above, with the z-coordinates in MNI152 space shown for reference. Here, bright yellow shades indicate that the region/task pair is positively associated with remission (i.e., higher baseline activation or greater increase in activation is predictive of remission), whereas bright blue shades indicate a negative association. As the 1st dose change of resting state centrality only displayed one region that passed permutation testing, maps of that region (left superior temporal gyrus) are not shown. These results were calculated by averaging the predictor importance weights assigned by the classification across all ten folds of cross-validation and over all thirty trials.

only one region from the 1st dose change in the fMRI resting state centrality metric passed permutation testing, maps of that region (left superior temporal gyrus) were not shown. Significant regions included frontal cortex, parahippocampus, hippocampus, caudate, thalamus, medial temporal cortex, middle cingulate, and visual cortex.

Figure 1.4 illustrates the effects of permuting the feature set for a given fMRI metric at a single time point. I found that the largest drop in AUC occurred when permuting the emotional reactivity feature map, indicating the relative utility of this probe in finding metrics that can be used for remission prediction.

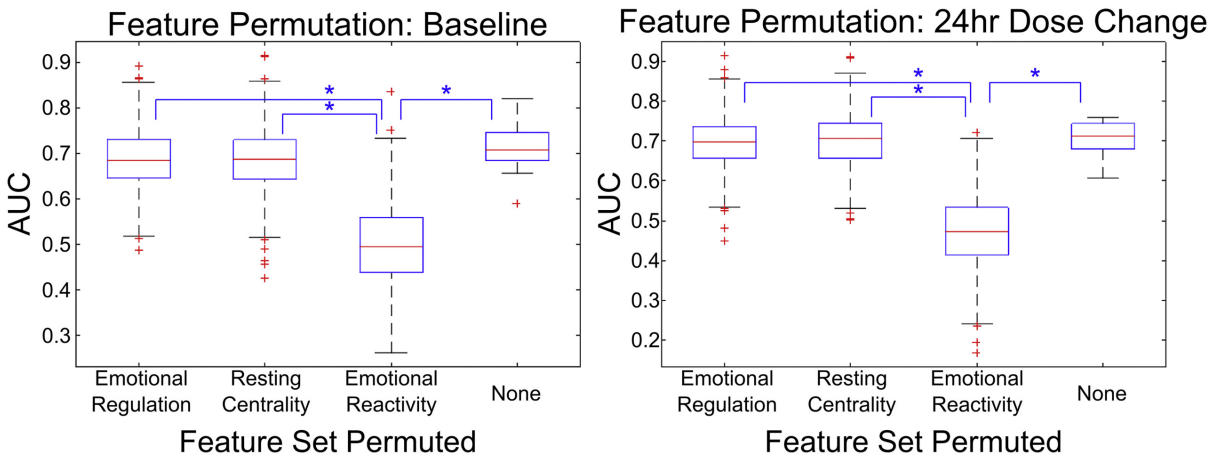

Figure 1.4: The impact to the classification accuracy of the fMRI predictor algorithm when certain feature map sets are permuted between subjects is shown via interquartile boxplots. The first three columns represent the result of shuffling the features of a given fMRI metric map between subjects, while the last column represents the accuracy when no features are permuted. Note that the largest drop in accuracy occurs when the emotional reactivity features are permuted, indicating the utility of using a task to probe specific features of neural activity. Statistical significance was determined via two-sample t-tests. All the pairs marked by asterisks have a p-value bound below 10^{-3}.

1.4 Discussion

In this chapter, I present a novel method of predicting treatment response in late-life-depression by utilizing the functional imaging metrics in response to a single-dose of a pharmacological intervention that demonstrates improvement over using baseline (pre-treatment) clinical and neuroimaging information.

There is a small, but significant past literature in LLD that focuses on pre-treatment activation [5, 19, 20]. These studies have identified hypoactive executive function during emotion reactivity, changes in resting state connectivity, including work on a subset of this study population that found lower pre-treatment centrality in the inferior frontal gyrus (IFG) as well as greater pre-treatment MeFG centrality [4]. One large study identified pre-treatment subtypes of depression, specifically that there exist four major subtypes that have distinct abnormalities in resting state connectivity [21] that also demonstrated a specificity for different treatments. In general, our results and previous studies support the use of pre-treatment fMRI for improving treatment outcomes.

One possible application of these results is facilitating neural target engagement - where we would both find a neural target and engage it to significantly improve depressive symptoms. Given the rising popularity of innovative methods to target neural markers (using interventions such as transcranial magnetic stimulation, TMS, or transcranial direct-current stimulation, tDCS), determining the appropriate neural targets to engage is a field of rising interest [22]. Future studies investigating differential target engagement are needed, as past studies seem to suggest that different therapies result in differential engagement of similar neural targets [23]. This would allow for identification of a pre-treatment neural target and matching with an appropriate antidepressant or therapy to engage or alter that target.

There are several limitations in this study. While the sample size is comparatively large, it is limited from a machine learning perspective. This limitation is especially important in several ways. In particular, past literature has shown that depression is a highly heterogeneous disorder and thus likely contributes to the relatively modest improvement in prediction in our study as well as past work. Further, treatment response itself is highly heterogeneous and it is likely that there are many paths to remission even within a single antidepressant. Finally, there are other factors that may play an important role, for instance individual variability in antidepressant metabolism may contribute to the fMRI response. We employed a 10-fold cross-validation as well as ensuring that any data reduction was done in fold (to avoid bias), to affirm the viability of the algorithm and address over-fitting concerns. These approaches help ensure that the improvement in prediction (above MADRS alone) is stable, and by performing all data reduction in fold we further avoid biasing our model (as this would improve our estimate of the different components within a sample). We also conducted the cross-validation multiple times (redrawing the folds) as it is possible that certain folds are more predictive than others. While these may address some over-fitting, it is not a replacement for independent validation and future studies should include data on larger cohorts and allow

for independent samples for verification.

This study builds on an already existing literature that has identified neural and behavioral subtypes – as our pre-treatment markers predicted remission while also identifying a single dose engagement effect – building on a sparse literature that seems to suggest that the neural activation occurs acutely. Thus, measuring this engagement is likely an important part of improving the overall efficacy of these treatments. Utilizing computational psychiatric approaches will allow for patients to be classified not only by their clinical symptoms, but also a set of neural targets that may need to be engaged. By engaging each target in a systematic manner, we may be able to improve overall response rates for depression treatment.

References

[1] Carmen Andreescu and Charles F Reynolds. Late-life depression: evidence-based treatment and promising new directions for research and clinical practice. *Psychiatric Clinics*, 34(2):335–355, 2011.

[2] Turhan Canli, Rebecca E Cooney, Philippe Goldin, Maulik Shah, Heidi Sivers, Moriah E Thomason, Susan Whitfield-Gabrieli, John DE Gabrieli, and Ian H Gotlib. Amygdala reactivity to emotional faces predicts improvement in major depression. *Neuroreport*, 16(12):1267–1270, 2005.

[3] Greg J Siegle, Cameron S Carter, and Michael E Thase. Use of fmri to predict recovery from unipolar depression with cognitive behavior therapy. *American Journal of Psychiatry*, 163(4):735–738, 2006.

[4] HT Karim, C Andreescu, D Tudorascu, SF Smagula, MA Butters, JF Karp, C Reynolds, and HJ Aizenstein. Intrinsic functional connectivity in late-life depression: trajectories over the course of pharmacotherapy in remitters and non-remitters. *Molecular psychiatry*, 22(3):450–457, 2017.

[5] Alexander Khalaf, Helmet Karim, Olga V Berkout, Carmen Andreescu, Dana Tudorascu, Charles F Reynolds, and Howard Aizenstein. Altered functional magnetic resonance imaging markers of affective processing during treatment of late-life depression. *The American Journal of Geriatric Psychiatry*, 24(10):791–801, 2016.

[6] Ahmad R Hariri, Alessandro Tessitore, Venkata S Mattay, Francesco Fera, and Daniel R Weinberger. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323, 2002.

[7] Ahmad R Hariri, Venkata S Mattay, Alessandro Tessitore, Francesco Fera, and Daniel R Weinberger. Neocortical modulation of the amygdala response to fearful stimuli. *Biological psychiatry*, 53(6):494–501, 2003.

[8] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005.

[9] John Ashburner and Karl J Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005.

[10] P Mazaika, S Whitfield-Gabrieli, Allan Reiss, and G Glover. Artifact repair for fmri data from high motion clinical subjects. *Human Brain Mapping*, 47(58):70238–1, 2007.

[11] Ameera X Patel, Prantik Kundu, Mikail Rubinov, P Simon Jones, Petra E Vértes, Karen D Ersche, John Suckling, and Edward T Bullmore. A wavelet method for modeling and despiking motion artifacts from resting-state fmri time series. *Neuroimage*, 95:287–304, 2014.

[12] Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain connectivity*, 2(3):125–141, 2012.

[13] Gabriele Lohmann, Daniel S Margulies, Annette Horstmann, Burkhard Pleger, Joeran Lepsien, Dirk Goldhahn, Haiko Schloegl, Michael Stumvoll, Arno Villringer, and Robert Turner. Eigenvector centrality mapping for analyzing connectivity patterns in fmri data of the human brain. *PloS one*, 5(4):e10232, 2010.

[14] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[15] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[16] Yves Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998 8*, pages 201–206. Springer, 1998.

[17] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. 2004.

[18] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer, 2000.

[19] Howard J Aizenstein, Meryl A Butters, Minjie Wu, Laura M Mazurkewicz, V Andrew Stenger, Peter J Gianaros, James T Becker, Charles F Reynolds III, and Cameron S Carter. Altered functioning of the executive control circuit in late-life depression: episodic and persistent phenomena. *The American Journal of Geriatric Psychiatry*, 17(1):30–42, 2009.

[20] Stefanie Brassen, Raffael Kalisch, Wolfgang Weber-Fahr, Dieter F Braus, and Christian Büchel. Ventromedial prefrontal cortex processing during emotional evaluation in late-life depression: a longitudinal functional magnetic resonance imaging study. *Biological psychiatry*, 64(4):349–355, 2008.

[21] Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38, 2017.

[22] Mark S George, Ziad Nahas, Monica Molloy, Andrew M Speer, Nicholas C Oliver, Xing-Bao Li, George W Arana, S Craig Risch, and James C Ballenger. A controlled trial of daily left prefrontal cortex tms for treating depression. *Biological psychiatry*, 48(10):962–970, 2000.

[23] Thomas Frodl, Johanna Scheuerecker, Veronika Schoepf, Jennifer Linn, Nikolaos Kout-

souleris, Arun LW Bokde, Harald Hampel, Hans-Jürgen Möller, Hartmut Brückmann, Martin Wiesmann, et al. Different effects of mirtazapine and venlafaxine on brain activation: an open randomized controlled fmri study. *The Journal of clinical psychiatry*, 71(4):4477, 2010.

CHAPTER 2

Deep brain stimulation for Parkinson's disease induces spontaneous cortical hypersynchrony in extended motor and cognitive networks

Parkinson's disease is an interesting neurological disorder as one of few diseases of the brain that we define by its root cause and neural activity: degeneration of dopaminergic neurons in the substantia nigra leading to hypersynchrony of the basal ganglia. It is simultaneously part of a small group of brain-related diseases that we have effective treatments for (especially in context of being a neurodegenerative disorder), one of which is deep brain stimulation (DBS). While DBS's mechanism is still unclear, shedding light on it can serve as a model for how targeted approaches can work in other neurological or psychiatric diseases. Using magnetoencephalography and spectral graph theory, I compared resting-state cortical connectivity between the off and on DBS stimulation states and to healthy controls. I found that turning DBS on increased high beta and gamma band synchrony (26 to 50 Hz) in a cortical circuit spanning the motor, occipitoparietal, middle temporal, and prefrontal cortices. These changes appeared to introduce abnormalities in the brain's functional cortical architecture relative to healthy controls unlike DBS's previously known subcortical normalization of pathological basal ganglia activity. This increased high beta/gamma synchronization may reflect compensatory mechanisms related

to DBS's clinical benefits, as well as undesirable non-motor side effects.

2.1 Introduction

Parkinson's disease is a movement and cognitive disorder characterized by the progressive degeneration of nigrostriatal dopaminergic neurons. While traditionally treated with dopaminergic medications, when pharmaceuticals no longer provide consistent efficacy or lead to severe dyskinesias, high frequency deep brain stimulation (DBS) of the sensorimotor territory of the subthalamic nucleus (STN) or internal globus pallidus (GPi) has been established as the most effective means of managing the symptoms of Parkinson's disease [1–4]. The therapeutic mechanism of action, however, is still elusive and poorly understood, in part due to the difficulty of conducting neuroimaging studies in the presence of DBS stimulator hardware, due to artifacts and potential safety concerns with fMRI [5–7]. This limited knowledge has become a barrier to improving the efficacy of DBS while minimizing side effects [5].

Numerous studies have implicated overactive oscillatory synchrony within the basal ganglia, particularly within the beta band (13–30 Hz), as an important pathological feature of untreated Parkinson's disease [8–11]. Studies examining interregional interactions using both fMRI and intraoperative recordings have demonstrated abnormal basal ganglia-motor functional connectivity in Parkinson's disease [12–14]. Network analyses have shown that brain networks become less organized and less topologically efficient as Parkinson's disease progresses [15]. Beta band hypersynchrony has also been observed in essential tremor, indicating its importance across other movement disorders [16, 17].

Studies comparing neural response when DBS is on to when DBS is off are critical to relate this hypersynchrony to DBS's downstream neural effects and therapeutic benefits. Effective stimulation has been shown to decrease beta band hypersynchrony in the basal ganglia, particularly within the high beta band region (21–30 Hz) [18, 19]. [20] used

electrocorticography recordings in patients with Parkinson's disease to show that STN DBS reduces beta phase-amplitude coupling in the primary motor cortex, in conjunction with reducing motor symptoms. [21] used magnetoencephalography in conjunction with STN recordings 3–6 days after surgery, while DBS leads were still externalized, to demonstrate that acutely after surgery STN DBS modulates connectivity between the basal ganglia and mesial premotor regions in the high beta band range.

How do these results generalize to outside the basal ganglia and motor cortex? [22] used invasive electrophysiology to show that stimulation of the STN could identify a monosynaptic connection with the prefrontal lobe that was associated with stopping-related activity. A meta-analytic study of fMRI and PET studies in [23] showed that both the STN and GPi were coactivated with the inferior frontal gyrus.

A critical question for understanding the mechanism of DBS is how does long range cortical to cortical synchronization differ when stimulation is turned on and do these changes normalize prior Parkinson's-related abnormalities or introduce new circuit dynamics? I investigated how DBS influences functional connectivity across cortical regions not accessible intraoperatively during DBS surgery, utilizing MEG and graph theory analyses. We hypothesized that DBS increases cortical connectivity, similar to dopaminergic replacement therapy [24].

To test this hypothesis, I compared resting-state, whole cortex functional connectivity using MEG in the absence of DBS stimulation (DBS-off) with recordings obtained during clinically effective high frequency stimulation (DBS-on). After artifact removal, I used data-driven analyses to assess network and subnetwork level differences between DBS-on and DBS-off across all frequencies and between all pairs of brain regions (e.g. not restricted to somatomotor networks) in an unbiased manner [6]. I also compared these results using the same methods to age matched healthy control subjects to assess whether differences in functional connectivity in the DBS-off condition compared to DBS-on represented a normalization of functional connectivity. These data driven methods have the disadvantage

of being relatively less sensitive to small differences between conditions and groups, but have the advantage of casting a wide net to catch large effects in a statistically rigorous and unbiased manner that can seed additional future hypothesis testing. These results suggest that turning DBS on increases high beta and gamma band synchrony (26 to 50 Hz) across a broad cortical circuit that includes both motor and non-motor systems. Furthermore, functional connectivity patterns in the DBS-off condition is more similar to age matched controls compared to the DBS-on condition, suggesting that rather than normalization, the increased beta and gamma band synchrony is a result of non-normalizing functional connectivity induced by DBS stimulation.

2.2 Methods

2.2.1 Subjects

DBS subjects were eleven patients with bilateral DBS implants for the treatment of Parkinson's disease, all of whom gave informed consent to participate under STUDY19030378 approved by the University of Pittsburgh Institutional Review Board. All subjects had implants in either the subthalamic nucleus (STN) or globus pallidus internus (GPi). Stimulation parameters are bilateral unless denoted with left (L) and right (R) designations. MDS-Unified Parkinson's Disease Rating Scale (UPDRS) are shown for the on and off medication conditions pre-operatively and while on DBS post-operatively. All subjects were chosen based on clinician and self report to have strong clinical response to stimulation (e.g. based on their charts and self-report, substantial improvement in clinical symptoms were seen when DBS was turned on), but only three had clinical response quantified using continuous and quantitative measures (UPDRS).

34 healthy controls were selected from a larger population on the basis of age and gender matching. All participants gave informed consent to participate under protocols approved by the University of Pittsburgh Institutional Review Board under STUDY19100015.

Healthy controls did not differ in average age (67.8 years with a standard deviation of 5.6 years) compared to the DBS group (66.5 ± 6.3 years, $p = 0.35$). Controls had 21 males, 13 females compared to the 9 males, 2 females in the DBS group ($p = 0.22$).

2.2.2 Data Collection and Preprocessing

Data was collected from 204 gradiometers and 102 magnetometers arranged in orthogonal triplets on an Elekta Neuromag Vectorview MEG system (Elekta Oy, Helsinki, Finland). Data were sampled at 1000 Hz. Electrooculogram and electrocardiogram were concurrently measured to be corrected for during off-line analysis. Head position indicators were used to continuously monitor head position during MEG data acquisition. Signal-space projection (SSP) was performed on MEG data that was subsequently band-pass filtered from 1–70 Hz, notch filtered at 59-61 Hz, down-sampled to 250 Hz via MNE C scripts, then processed via temporal signal-space separation (tSSS) using a previously validated preprocessing pipeline that cleanses DBS artifacts across DBS-on and DBS-off conditions [6]. Signal to noise ratio for the inverse calculation was set at nine per [25] demonstrating that higher ratios yielded more accurate detection of changes in connectivity.

Five minutes of resting-state data was collected when the DBS implant was turned on. The implant was then turned off for a half hour, after which another five minutes of resting-state data was collected while the DBS was still off. Resting-state was collected while subjects had their eyes open and fixated on a centrally presented cross. Five minutes of empty room data was also collected. Resting-state data for the controls were collected using an identical protocol.

2.2.3 Connectivity Analysis

Spontaneous phase locking measures the variability over time of the phase difference between every pairwise cortical location [26]. I calculated phase-locking values (PLVs)

from 1-60 Hz and corrected them using empty room noise as described in [27]. This yielded a 5124 (number of cortical dipoles) x 5124 (number of cortical dipoles) adjacency matrix of pairwise phase locking values between each cortical dipole relative to empty room for each participant at each frequency. To make the data comparable across participants in terms of differential coupling values across frequency bands, I normalized the PLVs with regards to frequency [28]. For each participant, I took the distribution of PLVs over all frequencies and calculated their cumulative distribution function and then scaled all phase locking values to this distribution. Phase locking values were computed in MATLAB using in-house analyses.

2.2.4 Frequency Band Analysis

To identify a frequency band that displayed significantly different connectivity between deep brain stimulation on and off, I utilized nonparametric cluster level statistics [29]. First, I averaged the PLV across all pairs of dipoles resulting in a 60 (frequency) x 1 vector of the "global connectivity" of a subject's entire brain network at a given frequency. I used cluster statistics to identify frequency bands whose overall connectivity changed when DBS was turned on [29]. More specifically, I calculated a paired t-test at each frequency between DBS on and off. Adjacent frequencies with a p-value below 0.05 were clustered together into potentially significant bands. The t-statistic of all frequencies within a given band were summed to give the overall t-statistic of that band. Using permutation trials where the DBS on and off states were swapped randomly with each other, I calculated a null distribution of the largest band t-statistic found each trial over 10000 trials. In the real dataset, we found one frequency band that passed an $\alpha = 0.05$ test as shown in Figure 2.1 where we find one such band. The connectivity matrices for each subject were then averaged over this frequency band to generate a 5124 (cortical dipoles) x 5124 adjacency matrix for each subject. I repeated this protocol except comparing DBS off with health controls. I also repeated this protocol while separating the STN and GPi

groups.

2.2.5 Laplacian Dimensionality Reduction

With only 11 subjects, it is very difficult to precisely identify "which" specific cortical connections are being perturbed by DBS. To generate an initial estimate to inspire future investigations, I used a data-driven dimensionality reduction approach with the caveat that these results should be interpreted as preliminary analyses requiring further investigations in a larger cohort.

I started by averaging each of the 5124 cortical dipoles across the 360 regions defined in the Human Connectome Project (HCP) atlas [30], resulting in two 360 x 360 matrices for each subject for DBS-ON and DBS-OFF. Using the graph Laplacian, I identify a single set of sparse cuts across the graph for all subjects and then identified how the connectivity across these sparse cuts changed when DBS is turned on. This method is based on finding sparse cuts for spectral clustering. In general, measuring connectivity across sparse cuts has been used as a way to summarize salient network features in a dimensionally compact manner and has seen increasing usage in human connectome analysis [31–34].

I start by averaging the connectivity matrices for all subjects during the DBS-off state to obtain an average matrix, $A_{\overline{\text{off}}}$. I calculate its Laplacian matrix as $L_{\overline{\text{off}}} = D_{\overline{\text{off}}} - A_{\overline{\text{off}}}$ with eigendecomposition $L_{\overline{\text{off}}} v^j_{\overline{\text{off}}} = \lambda^j_{\overline{\text{off}}} v^j_{\overline{\text{off}}}$. Each eigenvector/value, $v^j_{\overline{\text{off}}}, \lambda^j_{\overline{\text{off}}}$, represents a set of connections in the network that constitute a single sparse cut with the lowest eigenvalues representing the sparsest cuts [35].

I then calculate the Laplacian of each subject's connectivity matrices when DBS is turned on and off and project them across these eigenvectors to determine their empirical eigenvalues, a marker of their connectivity strength across these sparse cuts as defined in Equation 2.1. This generated two 360×1 vectors for each subject, $\vec{\lambda}_{i,\text{off}}$ and $\vec{\lambda}_{i,\text{on}}$.

$$\lambda^j_{i,\text{off}} = \left(L_{i,\text{off}} v^j_{\overline{\text{off}}}\right)^T v^j_{\overline{\text{off}}} \tag{2.1}$$

I identified a linear weighting of eigenvalues that distinguished between DBS on and off using a feature bagged support-vector-machine tested using leave-one-out cross-validation using Python's sklearn implementation under default settings. The training set was formed as $(x_{tr}, y_{tr}) = (\vec{\lambda}_{i,\text{off}} - \vec{\lambda}_{i,\text{on}}, -1), (\vec{\lambda}_{i,\text{on}} - \vec{\lambda}_{i,\text{off}}, +1)$ for all i representing in-fold subjects. Features that enabled classification between the two label classes were eigenvalues that showed consistent differences between the on and off states across subjects (as consistent as one can get with 11 subjects). I evaluated the statistical significance of the SVM's accuracy using permutation testing with 1000 trials where I randomly exchanged the on and off connectivity matrices for each subject.

To determine the change in connectivity associated with turning on DBS, I took the classification weights assigned to each eigenvalue, w_j, and calculated which connectivity matrix changes were associated with those eigenvalues according to Equation 2.2.

$$\Delta A = -\sum_j v^j_{\overline{\text{off}}} w_j \left(v^j_{\overline{\text{off}}}\right)^T \tag{2.2}$$

2.2.6 Identification of Stimulated and Suppressed Communities

To understand whether the changing connections due to DBS self-organized into a specific circuit (sub-network), I utilized the protocol described in [36]. More specifically, we clustered the change adjacency matrix calculated in Equation 2.2 according to the Arenas, Fernández and Gómez community detection model [37]. The number of clusters was determined according to Newman's modularity [38]. Each cluster was then assigned a C-score which detailed how strong the change in connectivity within that sub-network was relative to how strong it would be if the clusters were chosen randomly. The supposition is that a sub-network is considered more significant if connections within it were changing

greatly relative to the rest of the network [36].

To generate a null distribution of C-scores, we generated a hundred thousand random undirected, weighted graphs that preserved the edge density distribution of the change adjacency matrix calculated in Equation 2.2. We repeated the clustering analyses on these random graphs and selected the highest C-score of the resulting clusters to form our null distribution. For a cluster to be considered statistically significant, its C-score would have to be within the top five percent of this null distribution, which is shown in Figure 2.1.

We also repeated this process on the original resting DBS-off/on and control networks to see whether the identified DBS-activated sub-network was activated significantly prior to DBS and in healthy controls and was simply strengthened by DBS. The permutation process was repeated for each DBS/control group.

2.3 Results

Figure 2.1 shows the average connectivity across all cortical dipoles over all subjects for the DBS-on, off, and healthy control populations. A significant difference between DBS-on and DBS-off was seen in the high beta/gamma band region from 26 to 50 Hz as shown in Figure 2.1A (DBS-on greater than DBS-off, $p < 0.05$, cluster-level correction for multiple frequency comparisons). In contrast, DBS-off did not show significant global differences compared to age-matched controls in any frequency band, suggesting that the increased synchrony observed in DBS-on did not reflect normalization of abnormal functional connectivity. When STN and GPi stimulation groups were separated, no significant difference in any frequency band was detected; a larger sample may be required to determine whether there are more subtle differences between STN and GPi stimulation than can be detected in the present study. To check for confounds, I tested correlation between the average cortical synchrony to age and time after programming and found no significant correlations.

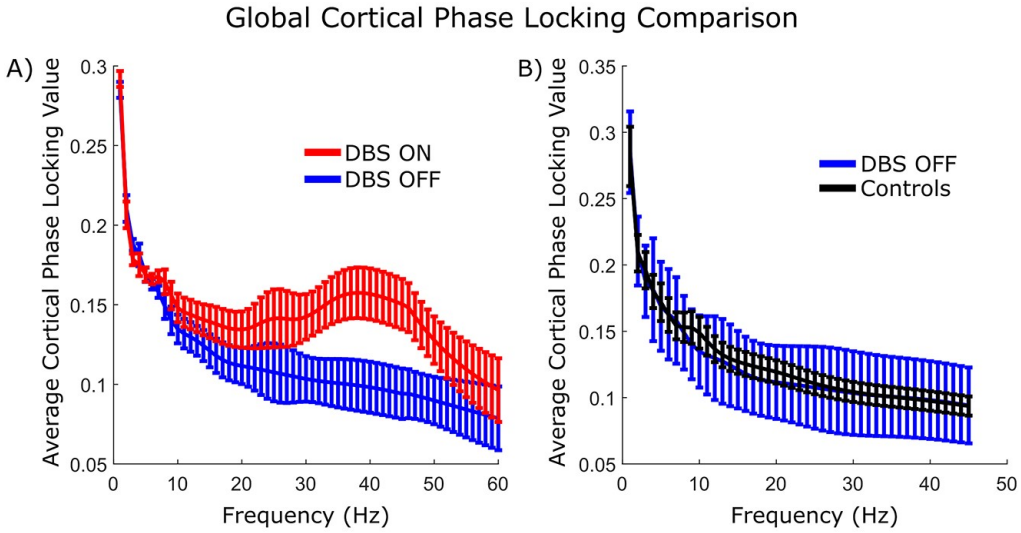

Figure 2.1: A) Spectral signature of global synchrony when deep brain stimulation is turned on and off. Average phase locking between every pair of cortical points with respect to frequency. Significantly increased beta and gamma band synchrony (26-50 Hz) was seen during DBS-on. Error bars indicate paired t-test 95% confidence intervals. B) The spectral signature of healthy controls does not show major deviations compared to the deep brain stimulation off condition. Error bars indicate two sample t-test confidence intervals.

All-to-all connectivity networks averaged across the high beta/gamma band (26-50 Hz) were computed for each subject for both DBS on and off. To identify a weighted group of connections whose average was consistently changing when DBS was turned on, I used a graph theory-based dimensionality reduction approach and a support vector machine whose reliability and significance was assessed via cross-validation. I found that we could identify a pattern of connectivity differences that accurately separated DBS on and off in nine of the eleven subjects (82% leave-one-subject-out cross validated accuracy, $P = 0.0053$ via permutation testing). Both of the GPi implanted patients were correctly classified, reinforcing that using this relatively broad data-driven analysis, GPi and STN stimulation show similar effects. We show which brain regions showed the largest increases in connectivity in Figure 2.2A where we find large effects in the motor cortex bilaterally, frontal cortex, occipitoparietal lobe, and the right middle temporal gyrus.

To quantify the relative similarity of DBS-on, DBS-off, and controls, I first used pattern

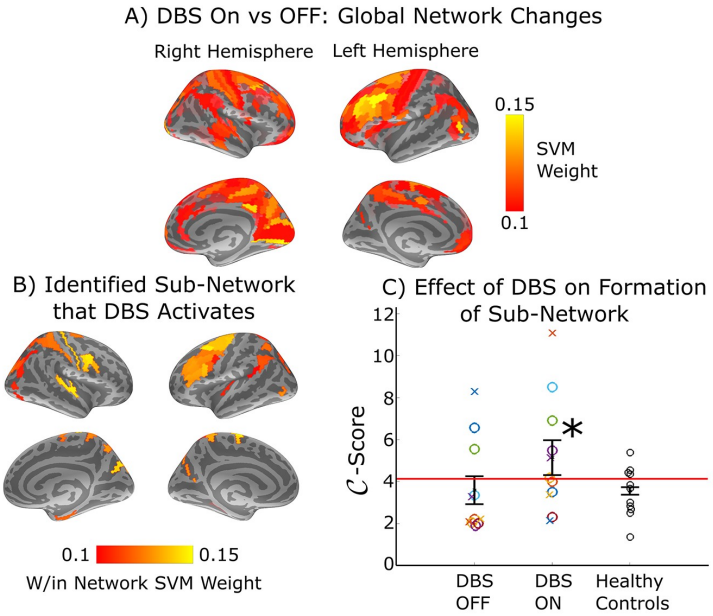

Figure 2.2: Map of high beta/gamma band connectedness. A) Ensemble of connections that were significantly synchronized by deep brain stimulation (DBS) (P = 0.005 permutation testing). Brighter areas indicate larger increases in connectivity with the rest of the cortex when DBS was turned on. B) The connectivity changes from the top figure that forms an inter-connected circuit. A community detection model was used to identify sub-networks whose connectivity within themselves were significantly different across the DBS on and DBS off conditions. Permutation testing revealed one such network, shown here. C) Cluster score of the identified sub-network in the DBS on/off conditions and in healthy controls. The connectivity strength within the sub-network shown in the bottom-left was compared to strength of equal-sized randomly selected sub-networks to assess whether the identified circuit was significantly activated relative to the rest of the cortex. The red line shows the false detection threshold ($\alpha = 0.05$). The results indicate that discovered circuit's activation was not significantly distinguishable from the rest of the cortex in healthy controls and when DBS was turned off but was significantly stronger than background when DBS was turned on (P = 0.048).

classification to train a model to discriminate the connectivity patterns from the DBS-on and DBS-off conditions and used that model to classify the controls. The resting state connectivity patterns of nearly all controls get classified as being more similar to the DBS-off condition than the DBS-on condition (28/34). Similarly, we trained a model to discriminate the DBS-on connectivity pattern from the control connectivity patterns and used that model to classify the DBS-off data, which classified all but one of the DBS-off patterns as controls (10/11). A classifier discriminating between controls and DBS-off had a 48% accuracy rate at discriminating between these two classes as tested via leave-one-out cross-validation. These results show that the connectivity patterns from the DBS-off condition were more like the patterns in controls than in the DBS-on condition.

2.3.1 Identification of Stimulated Subnetworks

In order to identify interconnected neurological circuits that were being activated by deep brain stimulation (subnetworks perturbed by DBS), we utilized the Arenas, Fernández, and Gómez (AFG) community detection model. Using permutation testing, the full cortical connectivity changes shown in Figure 2.2 were clustered into distinct sub-networks [37]. Permutation testing revealed one sub-network that passed statistical significance according to the AFG community detection model, which is illustrated in Figure 2.2B. This sub-network consisted of four major areas of the cortex: the middle/inferior temporal, occipitoparietal, motor, and the prefrontal cortices.

Figure 2.2C shows the cluster score for this circuit when DBS is on and off as well as in the healthy controls. Cluster score indicates how well a given sub-network is interconnected within itself relative to rest of the network using a permutation-generated null distribution illustrated by the red line. The circuit illustrated in Figure Figure 2.2B only emerges as statistically significant when DBS is turned on and is not significant in controls and the DBS off-condition.

2.4 Discussion

We studied the effects of basal ganglia DBS on cortical synchrony in patients with Parkinson's disease and found that DBS causes increased high beta and gamma band corticocortical synchrony (26 to 50 Hz). We show that these changes displace cortical networks relative to age-matched controls instead of normalizing them, with these effects being particularly magnified within an interconnected circuit consisting of the motor, occipitoparietal, temporal, and prefrontal cortices. This circuit does not appear to be significantly more activated than the average cortical resting-state synchrony in healthy controls and when DBS is turned off but emerges when DBS is turned on.

2.4.1 Study Limitations

Several caveats are necessary to consider when interpreting the results of this study. First, we utilized a data-driven approach requiring substantial multiple-comparisons corrections. While this allows us to detect networks that span across non-motor regions that a more targeted approach would not even consider, the tradeoff is that we are only powered to detect very large and straightforward changes. For example, [20] found that DBS normalizes coupling locally in the motor cortex between beta phase and broadband amplitude. By focusing on the motor cortex, such a study can pick up interesting changes that our approach is not powered to detect. In general, a lack of detected differences in any category should not be taken as evidence that those differences do not exist.

The second caveat is that the results rely on a sample size of 11 patients and would benefit from validation in a larger cohort, in particular to replicate the non-motor connectivity changes. Only relatively large effects can be detected reliably with a cohort of this size, thus this study likely misses subtle effects of DBS on cortical networks. The data-driven methods and small sample size likely also explain why few significant differences between DBS-off and controls were found. However, the study was powered to determine a key novel finding that cortical networks in the DBS-off condition were more like functional

networks in controls than in DBS-on. Third, while DBS was able to effectively control symptoms in the patients utilized in this study, metrics involving relative differences in outcomes were not utilized. Additionally, clinically effective stimulation in most of our patients was determined qualitatively rather than quantitatively. Therefore, while the changes in connectivity that we identify can be associated with qualitatitively effective treatment, their association to variability in the degree of individual treatment response would require a more powered study. Similarly, a future study with greater power is required to determine whether aspects of the changes seen in corticocortical connectivity relate to non-motor side effects. And lastly, in order to have sufficient power to detect the effects of DBS, we included all subjects with basal ganglia stimulation given that DBS to both STN and GPi have small, if any, differences motor and cognitive effects [39, 40]. When we did separate the GPi and STN stimulation cohorts, neither group was powered sufficiently to detect global cortical connectivity differences. Thus, these results are not meant to represent specific changes resulting from stimulation in either region but rather changes resulting from clinically effective basal ganglia deep brain stimulation.

2.4.2 DBS Modulates Long-Range Cortical Connectivity Involving the Prefrontal Cortex, Temporal Lobe, Motor Cortex, and Occipitoparietal Regions

Using our network reduction model, we were able to identify a sub-network of increased cortical connectivity involving the prefrontal cortex, temporal lobe, motor cortex, and the occipitoparietal lobe at the 26–50 Hz frequency band.

Prior literature involving the subcortex in Parkinson's, including ones studying the effects of dopaminergic medication, typically highlights low beta band frequencies, which generally fall right below the frequency band we identified [8, 10, 41]. Furthermore, [18] also showed that deep brain stimulation of the basal ganglia predominantly attenuates lower beta band power in that region.

However, when these relationships are expanded to include the cortex, evidence for higher frequencies emerge. [42] demonstrated that increased connectivity between the basal ganglia and premotor areas associated with Parkinson's occurred mostly in the high beta band. [43] also found that dopaminergic medication decreased the number of correlated pairs of scalp EEG pairs mostly at the high beta band (¿20 Hz). [21] showed both properties by demonstrating that DBS decreases basal ganglia power at the low beta band but decreases basal ganglia coherence with the mesial motor cortex in the high beta band. The mechanism of this shift from low beta band synchrony effects subcortically to high beta band synchrony changes in cortical areas may prove an important avenue of future studies, especially in the context of the effects of Parkinson's and its treatments.

Involvement of the lateral prefrontal cortex, somatosensory, motor/premotor, and occipitoparietal areas are supported by diffusion-tensor-imaging (DTI) and probabilistic tractography findings demonstrating structural connectivity between these regions and the basal ganglia [44, 45]. [22] showed evidence of a monosynaptic STN to prefrontal hyperdirect pathway involved in motor control inhibition, lending further credence to an anatomic basis for this network. The involvement of these regions in Parkinson's disease and its treatment are also supported by several functional imaging studies (fMRI and PET) [46, 47]. A recent MEG study supports the involvement of primary and supplementary motor cortices in the effects of DBS [21]. Connectivity between the temporal lobe and the basal ganglia has been validated by a combination of retrograde transneuronal viral studies and PET studies [48, 49]. Interestingly, [50] demonstrated that DBS in the basal ganglia was effective in controlling refractory partial epilepsy in patients with temporal lobe epilepsy.

2.4.3 Effects of DBS Displace Patients with Parkinson's Relative to Healthy Controls

In general, we did not find large differences between the DBS off condition and age-matched controls. We do not believe this means they are absent, on the contrary, a large ensemble of literature would indicate the opposite. As mentioned earlier, our sample was most likely not powered enough to detect these differences using a data-driven approach requiring substantial corrections for multiple comparisons. However, the fact that we did see significant differences when DBS was turned on indicates that in contrast to the reported subcortical effects of stimulation, where synchrony is reduced to resemble states observed in subjects without PD, stimulation's effect cortically appears to be in the opposite direction. A key question for future studies is which of these effects of DBS are associated with therapeutic outcomes, reflecting compensatory mechanisms to overcome Parkinsonian symptoms, versus which drive undesired side-effects.

2.4.4 DBS Activated Circuit Stands Out from Background Synchrony Only when DBS is Turned on

We found that our DBS-activated circuit's synchrony was not significantly different from the rest of the cortex in healthy controls and in patients with Parkinson's when the DBS device was turned off. When the DBS device was turned on, synchrony inside the network increased significantly relative to the rest of the cortex (beyond the overall activation induced by DBS). This increased cortical–cortical high beta/gamma synchrony may be a consequence of the release of pathological basal ganglia hyperinhibition seen in Parkinson's by DBS, leading to the observed network becoming active in DBS-on relative to both DBS-off and controls [51, 52]. There are two major possibilities for this finding. One is that this cortical network is not typically activated at rest but only during specific tasks, possibly higher-order motor control given the involvement of the premotor cortices. However, when DBS is turned on, this circuit is perturbed as a unit, causing it to also be

abnormally activated during resting state. Another is that the magnitude of this circuit's activation, including at rest, is typically small compared to other networks in the cortex, causing it to disappear into the background of other stronger networks. DBS then causes this circuit to become abnormally active. Further explorations into the state of this circuit under using various stimulation parameters and examining how these effects relate to motor and non-motor behavioral changes with DBS could help mediate between these two hypotheses leading to better understanding of the mechanisms of DBS. In particular, it will be important to determine if these changes are compensatory, and related to the magnitude of treatment efficacy, incidental, or related to unwanted DBS side effects [53].

References

[1] Patricia Limousin, Pierre Pollak, Abdelhamid Benazzouz, Dominique Hoffmann, Jean-François Le Bas, JE Perret, AL Benabid, and El Broussolle. Effect on parkinsonian signs and symptoms of bilateral subthalamic nucleus stimulation. *The Lancet*, 345(8942):91–95, 1995.

[2] Günther Deuschl, Carmen Schade-Brittinger, Paul Krack, Jens Volkmann, Helmut Schäfer, Kai Bötzel, Christine Daniels, Angela Deutschländer, Ulrich Dillmann, Wilhelm Eisner, et al. A randomized trial of deep-brain stimulation for parkinson's disease. *New England Journal of Medicine*, 355(9):896–908, 2006.

[3] Alim Louis Benabid, Stephan Chabardes, John Mitrofanis, and Pierre Pollak. Deep brain stimulation of the subthalamic nucleus for the treatment of parkinson's disease. *The Lancet Neurology*, 8(1):67–81, 2009.

[4] WMM Schuepbach, J Rau, K Knudsen, J Volkmann, P Krack, L Timmermann, TD Hälbig, H Hesekamp, SM Navarro, Niklaus Meier, et al. Neurostimulation for parkinson's disease with early motor complications. *New England Journal of Medicine*, 368(7):610–622, 2013.

[5] Ahmad Alhourani, Michael M McDowell, Michael J Randazzo, Thomas A Wozny, Efstathios D Kondylis, Witold J Lipski, Sarah Beck, Jordan F Karp, Avniel S Ghuman, and R Mark Richardson. Network effects of deep brain stimulation. *Journal of neurophysiology*, 114(4):2105–2117, 2015.

[6] Matthew J Boring, Zachary F Jessen, Thomas A Wozny, Michael J Ward, Ashley C Whiteman, R Mark Richardson, and Avniel Singh Ghuman. Quantitatively validating the efficacy of artifact suppression techniques to study the cortical consequences of deep brain stimulation with magnetoencephalography. *Neuroimage*, 199:366–374, 2019.

[7] Vladimir Litvak, Esther Florin, Gertrúd Tamás, Sergiu Groppa, and Muthuraman Muthuraman. Eeg and meg primers for tracking dbs network effects. *Neuroimage*, 224:117447, 2021.

[8] Peter Brown, Antonio Oliviero, Paolo Mazzone, Angelo Insola, Pietro Tonali, and Vincenzo Di Lazzaro. Dopamine dependency of oscillations between subthalamic nucleus and pallidum in parkinson's disease. *Journal of Neuroscience*, 21(3):1033–1038, 2001.

[9] Andrea A Kühn, Andreas Kupsch, Gerd-Helge Schneider, and Peter Brown. Reduction in subthalamic 8–35 hz oscillatory activity correlates with clinical improvement in parkinson's disease. *European Journal of Neuroscience*, 23(7):1956–1960, 2006.

[10] Constance Hammond, Hagai Bergman, and Peter Brown. Pathological synchronization in parkinson's disease: networks, models and treatments. *Trends in neurosciences*, 30(7):357–364, 2007.

[11] Ahmad Alhourani, Anna Korzeniewska, Thomas A Wozny, Witold J Lipski, Efstathios D Kondylis, Avniel S Ghuman, Nathan E Crone, Donald J Crammond, Robert S Turner, and R Mark Richardson. Subthalamic nucleus activity influences sensory and motor cortex during force transduction. *Cerebral cortex*, 30(4):2615–2626, 2020.

[12] Simon Baudrexel, Torsten Witte, Carola Seifried, Frederic von Wegner, Florian Beissner, Johannes C Klein, Helmuth Steinmetz, Ralf Deichmann, Jochen Roeper, and Rüdiger Hilker. Resting state fmri reveals increased subthalamic nucleus–motor cortex connectivity in parkinson's disease. *Neuroimage*, 55(4):1728–1738, 2011.

[13] Coralie De Hemptinne, Elena S Ryapolova-Webb, Ellen L Air, Paul A Garcia, Kai J Miller, Jeffrey G Ojemann, Jill L Ostrem, Nicholas B Galifianakis, and Philip A Starr. Exaggerated phase–amplitude coupling in the primary motor cortex in parkinson disease. *Proceedings of the national academy of sciences*, 110(12):4780–4785, 2013.

[14] Shoichi A Shimamoto, Elena S Ryapolova-Webb, Jill L Ostrem, Nicholas B Galifianakis, Kai J Miller, and Philip A Starr. Subthalamic nucleus neurons are synchronized to primary motor cortex local field potentials in parkinson's disease. *Journal of Neuroscience*, 33(17):7220–7233, 2013.

[15] Kim TE Olde Dubbelink, Arjan Hillebrand, Diederick Stoffers, Jan Berend Deijen, Jos WR Twisk, Cornelis J Stam, and Henk W Berendse. Disrupted brain network topology in parkinson's disease: a longitudinal magnetoencephalography study. *Brain*, 137(1):197–207, 2014.

[16] Efstathios D Kondylis, Michael J Randazzo, Ahmad Alhourani, Witold J Lipski, Thomas A Wozny, Yash Pandya, Avniel S Ghuman, Robert S Turner, Donald J Crammond, and R Mark Richardson. Movement-related dynamics of cortical oscillations in parkinson's disease and essential tremor. *Brain*, 139(8):2211–2223, 2016.

[17] Witold J Lipski, Thomas A Wozny, Ahmad Alhourani, Efstathios D Kondylis, Robert S Turner, Donald J Crammond, and Robert Mark Richardson. Dynamics of human subthalamic neuron phase-locking to motor and sensory cortical oscillations during movement. *Journal of neurophysiology*, 118(3):1472–1487, 2017.

[18] Helen Bronte-Stewart, Crista Barberini, Mandy Miller Koop, Bruce C Hill, Jaimie M Henderson, and Brett Wingeier. The stn beta-band profile in parkinson's disease is stationary and shows prolonged attenuation after deep brain stimulation. *Experimental neurology*, 215(1):20–28, 2009.

[19] Alexandre Eusebio, Wesley Thevathasan, L Doyle Gaynor, Alek Pogosyan, Ed Bye, Thomas Foltynie, Ludvic Zrinzo, Keyoumars Ashkan, Tipu Aziz, and Peter Brown. Deep brain stimulation can suppress pathological synchronisation in parkinsonian patients. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(5):569–573, 2011.

[20] Coralie De Hemptinne, Nicole C Swann, Jill L Ostrem, Elena S Ryapolova-Webb, Marta San Luciano, Nicholas B Galifianakis, and Philip A Starr. Therapeutic deep

brain stimulation reduces cortical phase-amplitude coupling in parkinson's disease. *Nature neuroscience*, 18(5):779–786, 2015.

[21] Ashwini Oswal, Martijn Beudel, Ludvic Zrinzo, Patricia Limousin, Marwan Hariz, Tom Foltynie, Vladimir Litvak, and Peter Brown. Deep brain stimulation modulates synchrony within spatially and spectrally distinct resting state networks in parkinson's disease. *Brain*, 139(5):1482–1496, 2016.

[22] Witney Chen, Coralie de Hemptinne, Andrew M Miller, Michael Leibbrand, Simon J Little, Daniel A Lim, Paul S Larson, and Philip A Starr. Prefrontal-subthalamic hyperdirect pathway modulates movement inhibition in humans. *Neuron*, 106(4):579–588, 2020.

[23] Jordan L Manes, Amy L Parkinson, Charles R Larson, Jeremy D Greenlee, Simon B Eickhoff, Daniel M Corcos, and Donald A Robin. Connectivity of the subthalamic nucleus and globus pallidus pars interna to regions within the speech network: A meta-analytic connectivity study. *Human brain mapping*, 35(7):3499–3516, 2014.

[24] Diederick Stoffers, Johannes LW Bosboom, Erik Ch Wolters, Cornelis J Stam, and Henk W Berendse. Dopaminergic modulation of cortico-cortical functional connectivity in parkinson's disease: an meg study. *Experimental neurology*, 213(1):191–195, 2008.

[25] Ana-Sofía Hincapié, Jan Kujala, Jérémie Mattout, Sebastien Daligault, Claude Delpuech, Domingo Mery, Diego Cosmelli, and Karim Jerbi. Meg connectivity and power detections with minimum norm estimates require different regularization parameters. *Computational intelligence and neuroscience*, 2016, 2016.

[26] Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J Varela. Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208, 1999.

[27] Avniel Singh Ghuman, Jonathan R McDaniel, and Alex Martin. A wavelet-based

method for measuring the oscillatory dynamics of resting-state functional connectivity in meg. *Neuroimage*, 56(1):69–77, 2011.

[28] Winfried Schlee, Thomas Hartmann, Berthold Langguth, and Nathan Weisz. Abnormal resting-state cortical coupling in chronic tinnitus. *BMC neuroscience*, 10:1–11, 2009.

[29] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190, 2007.

[30] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[31] Marco Saerens, Francois Fouss, Luh Yen, and Pierre Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *ECML*, volume 3201, pages 371–383. Springer, 2004.

[32] Ashish Raj, Amy Kuceyeski, and Michael Weiner. A network diffusion model of disease progression in dementia. *Neuron*, 73(6):1204–1215, 2012.

[33] Farras Abdelnour, Henning U Voss, and Ashish Raj. Network diffusion accurately models the relationship between structural and functional brain connectivity networks. *Neuroimage*, 90:335–347, 2014.

[34] Maxwell B Wang, Julia P Owen, Pratik Mukherjee, and Ashish Raj. Brain network eigenmodes provide a robust and compact representation of the structural connectome in health and disease. *PLoS computational biology*, 13(6):e1005550, 2017.

[35] Anand Louis, Prasad Raghavendra, Prasad Tetali, and Santosh Vempala. Many sparse cuts via higher eigenvalues. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 1131–1140, 2012.

[36] Andrea Lancichinetti, Filippo Radicchi, and José J Ramasco. Statistical significance of communities in networks. *Physical Review E*, 81(4):046110, 2010.

[37] Alex Arenas, Alberto Fernandez, and Sergio Gomez. Analysis of the structure of complex networks at different resolution levels. *New journal of physics*, 10(5):053039, 2008.

[38] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[39] Lilei Peng, Jie Fu, Yang Ming, Shan Zeng, Haiping He, and Ligang Chen. The long-term efficacy of stn vs gpi deep brain stimulation for parkinson disease: A meta-analysis. *Medicine*, 97(35), 2018.

[40] Joshua K Wong, James H Cauraugh, Kwo Wei David Ho, Matthew Broderick, Adolfo Ramirez-Zamora, Leonardo Almeida, Aparna Wagle Shukla, Christina A Wilson, Rob MA de Bie, Frances M Weaver, et al. Stn vs. gpi deep brain stimulation for tremor suppression in parkinson disease: a systematic review and meta-analysis. *Parkinsonism & related disorders*, 58:56–62, 2019.

[41] A Priori, G Foffani, A Pesenti, F Tamma, AM Bianchi, M Pellegrini, M Locatelli, KA Moxon, and RM Villani. Rhythm-specific pharmacological modulation of subthalamic activity in parkinson's disease. *Experimental neurology*, 189(2):369–379, 2004.

[42] Vladimir Litvak, Ashwani Jha, Alexandre Eusebio, Robert Oostenveld, Tom Foltynie, Patricia Limousin, Ludvic Zrinzo, Marwan I Hariz, Karl Friston, and Peter Brown. Resting oscillatory cortico-subthalamic connectivity in patients with parkinson's disease. *Brain*, 134(2):359–374, 2011.

[43] Jobi S George, Jon Strunk, Rachel Mak-McCully, Melissa Houser, Howard Poizner, and Adam R Aron. Dopaminergic therapy in parkinson's disease decreases cortical

beta band coherence in the resting state and increases cortical beta band power during executive control. *NeuroImage: Clinical*, 3:261–270, 2013.

[44] Christian Lambert, Ludvic Zrinzo, Zoltan Nagy, Antoine Lutti, Marwan Hariz, Thomas Foltynie, Bogdan Draganski, John Ashburner, and Richard Frackowiak. Confirmation of functional zones within the human subthalamic nucleus: patterns of connectivity and sub-parcellation using diffusion weighted imaging. *Neuroimage*, 60(1):83–94, 2012.

[45] Nora Vanegas-Arroyave, Peter M Lauro, Ling Huang, Mark Hallett, Silvina G Horovitz, Kareem A Zaghloul, and Codrin Lungu. Tractography patterns of subthalamic nucleus deep brain stimulation. *Brain*, 139(4):1200–1210, 2016.

[46] James Rowe, Klaas Enno Stephan, Karl Friston, Richard Frackowiak, Andrew Lees, and Richard Passingham. Attention to action in parkinson's disease: impaired effective connectivity among frontal cortical regions. *Brain*, 125(2):276–289, 2002.

[47] Tao Wu, Liang Wang, Yi Chen, Cheng Zhao, Kuncheng Li, and Piu Chan. Changes of functional connectivity of the motor network in the resting state in parkinson's disease. *Neuroscience letters*, 460(1):6–10, 2009.

[48] Frank A Middleton and Peter L Strick. The temporal lobe is a target of output from the basal ganglia. *Proceedings of the national academy of sciences*, 93(16):8683–8687, 1996.

[49] Ronald B Postuma and Alain Dagher. Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. *Cerebral cortex*, 16(10):1508–1521, 2006.

[50] K Lee, K Jang, and Y Shon. Advances in functional and reparative neurosurgery. 2006.

[51] Arvind Kumar, Stefano Cardanobile, Stefan Rotter, and Ad Aertsen. The role of

inhibition in generating and controlling parkinson's disease oscillations in the basal ganglia. *Frontiers in systems neuroscience*, 5:86, 2011.

[52] Luka Milosevic, Suneil K Kalia, Mojgan Hodaie, Andres M Lozano, Alfonso Fasano, Milos R Popovic, and William D Hutchison. Neuronal inhibition and synaptic plasticity of basal ganglia neurons in parkinson's disease. *Brain*, 141(1):177–190, 2018.

[53] Anna Antosik-Wojcinska, Lukasz Swiecicki, Monika Dominiak, Emilia Soltan, Przemyslaw Bienkowski, and Tomasz Mandat. Impact of stn-dbs on mood, drive, anhedonia and risk of psychiatric side-effects in the population of pd patients. *Journal of the neurological sciences*, 375:342–347, 2017.

CHAPTER 3

Long term brain dynamics form a punctuated equilibrium of stable states interrupted by chaotic-like transitions

Up to this point, I have primarily investigated how the brain changes in response to various therapeutics using short snapshots of neural activity that were collected over a few minutes in each participant. Approaches studying neural activity over this timescale have been a staple of cognitive neuroscience for many decades that have uncovered a large range of knowledge on how our brains interact with our environments and bodies. However, many critical neurocognitive processes, such as performing natural activities and fluctuations of arousal, take place over significantly longer timescales over minutes-to-days in real-world environments.

Here I harnessed the opportunity to study brain dynamics during real-world behavior mostly continuously for between 3-12 days using intracranial multi-electrode recordings in twenty humans. During this time, participants engaged in natural activities, including interacting with friends, family, and staff, watching TV, sleeping, etc., with simultaneous neural and video recordings. We found that brain network dynamics predicted neurocognitive phenomena such as circadian rhythm, arousal, and multiple aspects of behavior (socializing, watching a screen, etc.). The individual functional networks, as well as their

pairwise interactions, possessed simple and stable dynamic properties that were conserved over days. In contrast to single or paired network behavior, the mixture of all functional networks showed patterns of "punctuated equilibrium": periods where networks would remain in stable states that corresponded to behavior and were interrupted by transitory bursts that were difficult to predict, displayed chaotic characteristics, and coincided with behavioral transitions. Brain state statistics displayed power laws characteristic of critical dynamics that are a trait of systems where complexity emerges from simple and stable building blocks. These results indicate that the complex and flexible brain dynamics that underpin real-world behavior are an emergent property of mixtures of individual, stable networks with simple dynamics.

3.1 Introduction

Many important neurocognitive processes take place on the order of minutes to days in dynamic, ever-changing "real" environments. Behaviorally, we transition between tasks like reading and talking to friends over minutes to hours. Neurobiologically, the interaction between someone's brain and body is driven by hormones, sympathetic, and parasympathetic drivers related to processes like arousal and circadian rhythms that fluctuate over a similar timescale [1]. However, the vast majority of what we know about human brain activity is based on studies that record neural signals on the scale of milliseconds-to-seconds while participants perform well-controlled tasks or rest in an artificial neuroimaging environment.

Some studies have analyzed brain state dynamics over minutes in a single sitting or by repeatedly sampling a few minutes per day spread out over days to months in an artificial setting using functional neuroimaging [2–10]. Longitudinal snapshots of a few minutes over days have also been studied in real-world settings, though typically associated with structured tasks or treatment interventions [11, 12]. While much has been learned about human brain network dynamics in these types of studies, we still do not know how

the brain continuously evolves over hours-to-days in real-world settings during natural behavior.

To assess human brain network dynamics in a real-world setting continuously over days, I leveraged chronic intracranial recordings in neurosurgical participants (80-126 electrodes implanted per participant) undergoing treatment for epilepsy (Figure 3.1 and Supplemental Figure S1). Specifically, I examined brain network dynamics from neural recordings in twenty humans for between 75 to 283 hours (near-continuous recordings across approximately 3-12 days). During this time, the participants were confined to the hospital but would freely socialize with friends, family, and staff, interact with digital devices, sleep, watch TV, and perform other volitional natural behaviors while under neural and video monitoring. Using these data, we built all-to-all electrode functional connectivity matrices (partial connectomes) every five seconds across the entire recording session (Figure 3.1), divided these connectivity matrices into data-driven networks, and removed electrodes and activity related to each participant's seizure onset zone and propagation. We then asked, what are the properties of brain network dynamics during continuous, natural behavior over the week?

Notably, the brain must balance competing demands of flexibility that allows people to react to changing cognitive demands while maintaining anatomical stability of networks and systems. Thus, we particularly asked what features of brain network dynamics changed rapidly, what aspects were consistent from hour-to-hour and day-to-day, and how these dynamics related to natural behavior. We investigated these questions over increasing spatial scale: starting with individual areas (functionally defined "parcels"), to the dynamics of covarying sets of parcel activity ("network components" acknowledging that we only have partial brain coverage in individual participants), to patterns of pairwise interactions between different network components, and finally to how the brain forms states and transitions during its week-long trajectory through a space defined by all network components (mixtures of all networks).

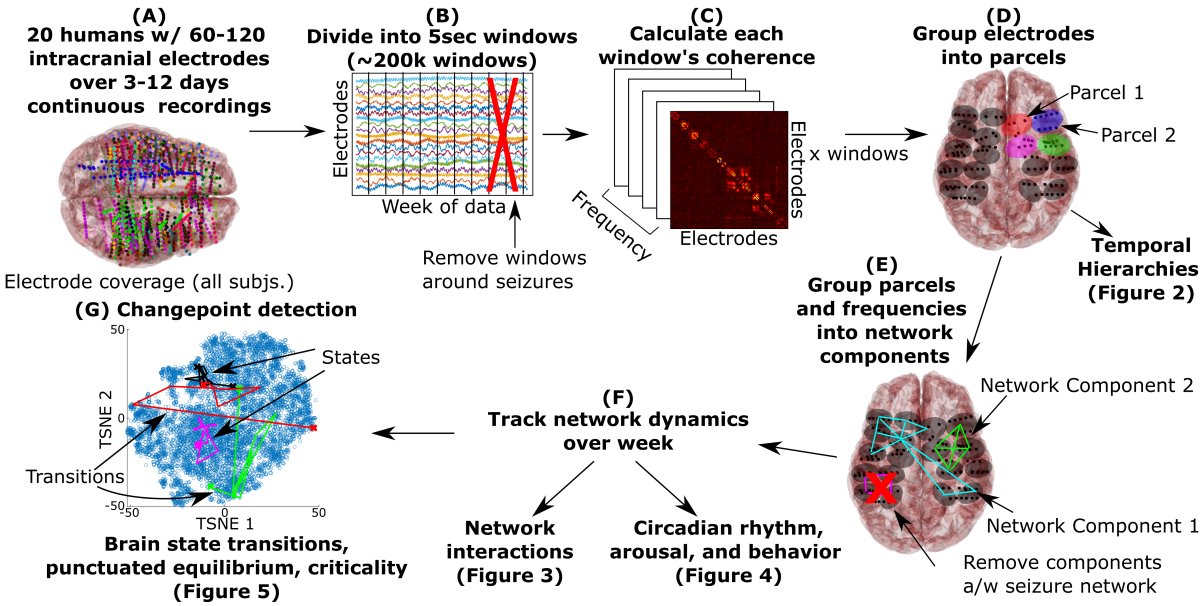

Figure 3.1: (A) We took between 3-12 days of continuous neural recordings from twenty participants and (B) split it into five-second-long non-overlapping windows, removing windows around seizure activity and filtering/regressing out artifacts. (C) We generated a functional connectome for each window as the coherence between all pairs of electrodes and removed additional artifacts by regression and independent component analysis. (D) We grouped electrodes with high coherence to each other over the week into parcels that tended to be anatomically close together (e.g. functional parcellation of cortical recordings). The parcel dynamics fell into temporal hierarchies that followed anatomical trends (Figure 3.2). (E) Parcels and frequencies that covaried were grouped into network components using a robust Principal Components Analysis. (F) The dynamics of these networks showed consistent pair-wise interactions (Figure 3.3) and relations to circadian rhythm, arousal, and behavior (Figure 3.4). (G) Overall brain state dynamics were assessed by finding transition points in the overall mixture of all networks' evolution over time using change point detection. The mixture of all brain networks would fall into stable states punctuated by transitions that were complex, unique, and showed consistent power law distributions (Figures 3.5).

3.2 Main Findings

3.2.1 Functional parcels show temporal consistency over days and reveal anatomic trends

Before studying how the whole measured brain evolves over the course of a week, I started by breaking it down into smaller pieces and studying how those pieces change over the week in isolation. We used a data-driven approach to identify small groups of tightly connected electrodes that made up coherent functional brain parcels (I use the term brain "parcels" rather than brain "areas" because brain areas are traditionally defined based on anatomical landmarks, and these brain parcels are anatomically compact but defined based on similarity/high coherence of the neural activity within a parcel). Our first question was whether these parcels showed stable temporal characteristics over the week: e.g., if a brain parcel fluctuates quickly on one day relative to other parcels, does it do so on other days as well?

After removing an hour before and after ictal (seizure) events as determined by the clinical team, the coherence between all pairs of electrodes in a subject was calculated every five seconds over five frequency bands: theta (θ: 4-8Hz), alpha (α: 8-12Hz), low beta (β_l: 14-20Hz), high beta (β_u: 20-30Hz), and gamma (γ: 30-70Hz). The electrodes were parcellated into tightly connected groups of electrodes (a coherent functional brain "parcel"). The parcel assignments remained very stable throughout the week (Supplementary Figure S2), which allowed us to define one set of parcels for the entire week. Parcels consistently contained electrodes that were anatomically close together, hence our description of this process as "parcellation" of the brain. To remove seizure-related activity from our analyses, I first removed any parcels associated with the subject's seizure onset zone and early propagation. I cleaned the remaining parcels by using linear regression on the activity of the seizure-related parcels to predict the activity on the remaining non-seizure-related parcels and then removed this co-linearity from them. I found that the

coherence within each parcel showed specific dynamical patterns or trajectories that would be repeated over different hours and days of data (Figure 3.2A and Supplementary Figure S3). This dynamical stability can be quantified by how slowly their autocorrelation curves decayed (timescales). found that timescale differences between parcels were preserved over time, indicating that parcels with relatively faster or slower timescales would remain so throughout the week. Specifically, the autocorrelation magnitude at one hour and the timescale of how quickly autocorrelation decayed showed reliable differences between parcels that were conserved over different six-hour time blocks throughout the week (Supplementary Figure S3).

These parcel-to-parcel differences were linked to anatomical trends over all twenty subjects by assigning each parcel to one of five lobes (frontal, temporal, parietal, occipital, and basal ganglia) and one of six canonical fMRI networks ("default mode", "dorsal attention", "salience", "somatomotor", "control", and "visual" as defined in 13) dependent on which lobe/network it had the most overlap with (parcels with no clear overlap were not considered for this analysis). Parcels in the default mode network and basal ganglia consistently showed the longest and slowest timescales across our subjects and parcels of the salience network had the shortest timescales (Figure 3.2B). These findings demonstrate an intrinsic stability in neural dynamics that separates "slow" from "fast" brain parcels over minutes to hours. A temporal hierarchy, typically measured using autocorrelation, has been hypothesized in the brain with transmodal (e.g. default mode network) systems being the slowest, integrating over multiple seconds [2, 3, 13–16]. These findings extend the observation of slow default mode network dynamics to minutes and hours during natural behavior in a real-world setting.

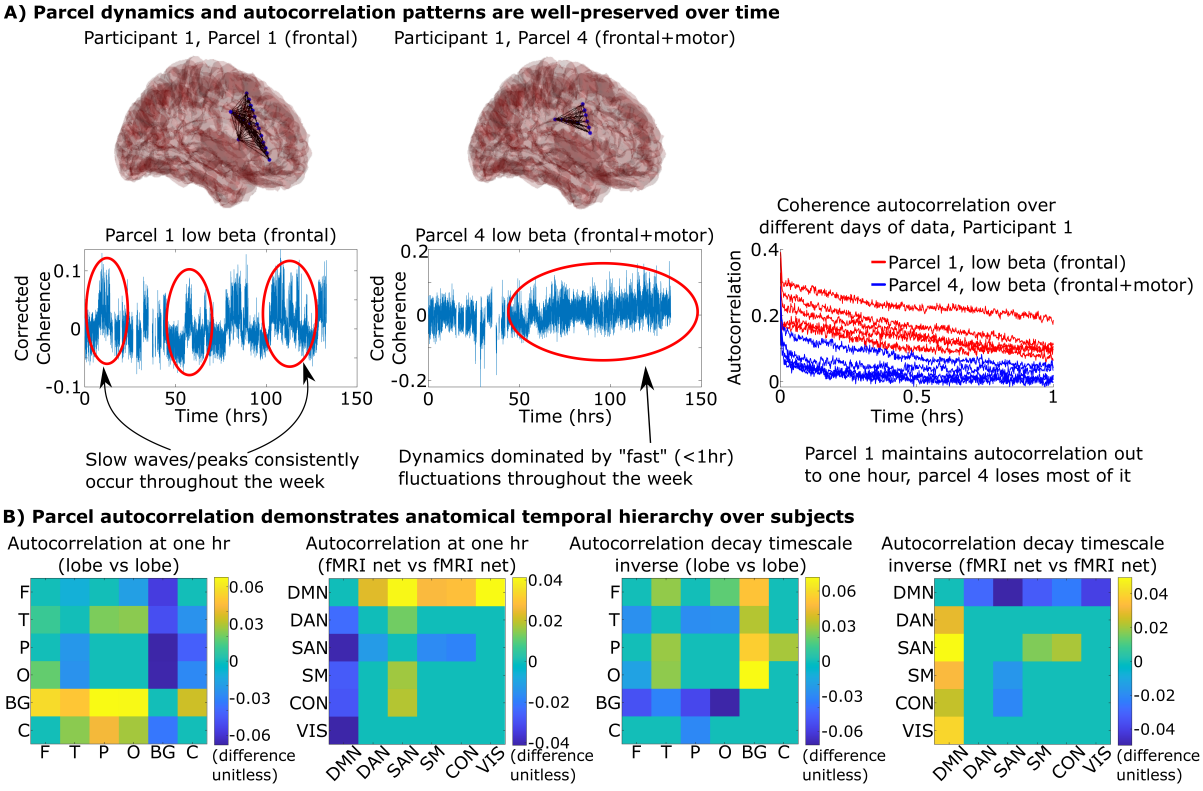

Figure 3.2: A) The coherence within two parcels from a representative participant (left) and the autocorrelations of those regions calculated on each day separately (right). Autocorrelation curves of the same color represent the autocorrelation of that parcel on different days of data. Each parcel's coherence displays a "unique" temporal signature that is conserved over days that is reflected by stability in their autocorrelation curves. Breaks/skips in data represent windows removed due to seizure activity. Comparisons between all parcels for each participant are shown in Supplementary Figure S3. B) Pairwise comparisons between the autocorrelation of parcels (averaged over frequency and participants) that belong to one lobe (F: frontal, T: temporal, P: parietal, O: occipital, BG: basal ganglia, C: cingulate)/fMRI resting network versus another. Cell values indicate the difference in autocorrelation (y axis minus x axis) between two regions or networks found using linear mixed effect models. Autocorrelation decay timescale inverse is an indicator on how sharply the autocorrelation curve decayed with lower values indicating less autocorrelation decay at high timescales. Non-zero cells indicate statistically significant differences post multiple comparisons correction ($p < 0.05$).

3.2.2 Network components form dynamical relationships and joint distributions that are preserved over days

After investigating how parcels of the brain would act on an individual basis, we wanted to understand how they interacted with each other. As many parcels were highly co-linear with each other (Supplementary Figure S17), I decided to group co-linear parcels into networks[1]. After finding that those networks also possessed consistent timescales as individual parcels did (Supplementary Figure S5), I examined how these networks interacted with one another.

More specifically, I used robust principal components analysis to identify parcels and frequencies that covaried with each other, defining each principal component as a "network component" that captured the overall connectome dynamics in a data-driven fashion while reducing the dimensionality of the dataset [17, 18] (I use the term "network components" because the recordings did not have full brain coverage and therefore these covarying parcels are components of brain networks). The activation of each network component during a window was defined as the weighted average of the parcel coherences within a network component (dot product between network activation and principal component weights). These network component activations showed consistent temporal behavior over time as individual functional regions did (Supplemental Figure S5).

I next examined the dynamics of individual networks of parcels and how these networks interacted with one another. I used robust principal components analysis to identify parcels and frequencies that covaried with each other, defining each principal component as a "network component" that captured the overall connectome dynamics in a data-driven fashion while reducing the dimensionality of the dataset (remaining results in this manuscript still hold statistical significance without dimensionality reduction as shown in

[1]To ensure that the rest of the results were robust to this PCA procedure, I examined whether the results held without using PCA for dimensionality reduction. The results in Supplementary Figures S12 to S16 show that our results remain statistically significant, though the effect sizes declined likely because the power of our analysis increases by grouping together covarying parcel activity into functional network components and by removing noisy, low variance PCs.

Supplementary Figures S12 to S16 but some classifier performance unsurprisingly declines) [17, 18]. I included the previously removed seizure-related parcels into the principal components analysis and then removed any network components that overlapped with the seizure-related areas (Supplemental Figure S4) to ensure I removed both the seizure foci and all areas statistically linked to it. The activation of each network component during a window was defined as the parcel coherences projected onto the corresponding principal component. These network component activations showed consistent temporal behavior over time as individual functional regions did (Supplemental Figure S5).

After investigating individual network components, going up one step in spatial scale, we asked whether the activity of pairs of network components could be reliably linked to one another. For each day, we calculated the joint distribution between all possible pairs of network component activations and asked whether this joint distribution was both reliably preserved across the week and indicated significant non-independent and/or non-linear relationships (while principal components will group features with linear relationships together, it will not do the same for non-linearities). More specifically, we calculated the distance between the joint distributions on different days of recordings versus the distance between these distributions and an independent null (more detail in Methods). The joint distribution between brain networks covered characteristic areas in the space that were well preserved over days, indicating that brain networks "dance" with one another in idiosyncratic ways. Some had antagonistic relationships where one network appears to suppress another; some would behave as if one network "gated" the other – inactivity in one network (in other words low coherence within contained parcels) would mandate inactivity in another. For example, the "V" shaped patterns in Figure 3.3A indicate that either both networks would be inactive together, or when one was active, the other would either be positively active or negatively active (but not inactive). All participants possessed several networks that showed such pairwise interactions (Figure 3.3 and Supplemental Figure S6), indicating that not only do individual network components

have consistent dynamics, but they also have consistent pair-wise interactions.

3.2.3 Network components predict both physiology and behavior

One critical question is whether these network dynamics are related to neurophysiology and behavior. I specifically looked at how linear combinations of network components correlated to circadian rhythm (time of day), predicted arousal (heart rate), and classified behavior (video recordings).

I took the first half of the week for each participant and used canonical correlation analysis (CCA) [19] to identify a network mixture that maximized correlation to time of day (CCA was used over regression to allow encoding time of day via phase). I tested this mixture in the second half of the week using permutation testing (out of sample validated correlation) and found that 11 of 20 participants had a network mixture significantly linked with circadian rhythm (Figure 3.4A and Supplementary Figure S7). Notably, six of the nine participants that were not linked to circadian rhythm had notes in their clinical file indicating various sleep disturbances such as nighttime-awakening seizures, intentional sleep deprivation for clinical purposes, or difficulty sleeping suggesting that these participants had disrupted circadian rhythms due to sleep issues during the week.

Seven subjects had sufficiently clean electrocardiogram (EKG) signals that were used to track heart rate. Heart rate is strongly correlated with the degree of arousal and is used here as an approximator [20]. I used L1-regularized [21] regression over the first half of the week to identify a mixture of networks that predicted heart rate and tested it on the remaining half (Figures 3.4B and Supplementary S8). I found that six of the seven subjects had network components that were associated with arousal.

Nine subjects had video recording monitoring throughout the week at a sufficient quality to determine what the subject was doing throughout each day. I randomly selected two days from each subject and labeled times where the subject was watching a digital screen, socializing with another person, or physically interacting and manipulating a held

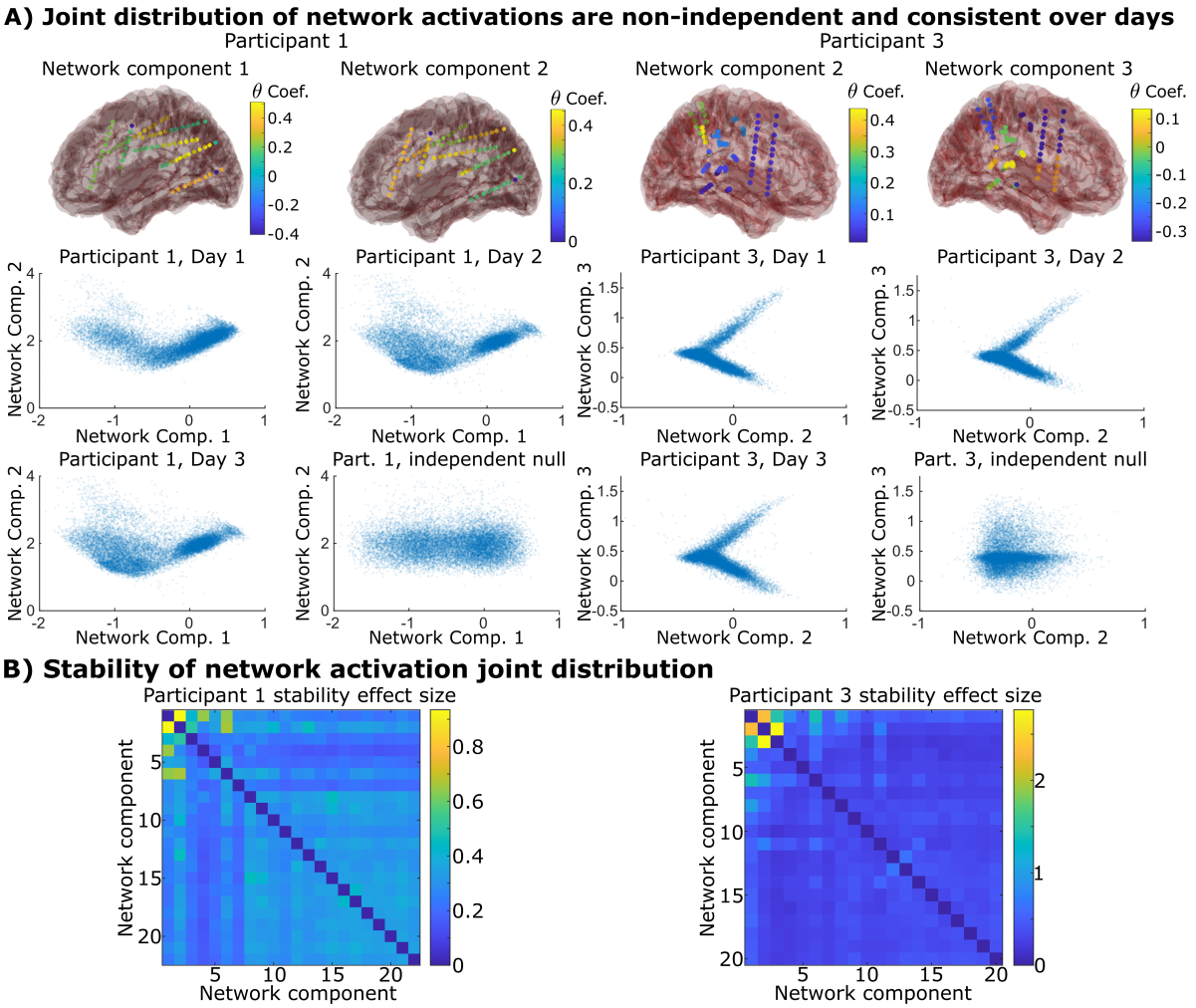

Figure 3.3: (A) A pair of networks/principal components from two subjects that showed non-independent distributions that are conserved over different days of data. The null distribution if the network components were independent of one another are also shown for comparison. B) The distance between the joint distribution of each pair of networks/principal components compared to the null distribution where they are independent of each other. Non-zero effect sizes represent statistically significant distances as determined via permutation testing. We find several pairs that demonstrate such relationships, most notably the lower network components that capture most of the variance in the dataset. Supplementary Figure S6 shows that all twenty subjects possessed several such pairs.

object. These three behavioral labels were not mutually exclusive. I trained L1-regularized logistic classifiers on one day to identify a mixture of network components associated with each behavior and then tested them on the second day. All subjects possessed network components that were associated with behavior, with two subjects shown in Figure 3.4C and all subjects in Supplementary S9.

Taken together, these three tests show that network components predict both physiological metrics and behavior that was replicated on multiple days.

3.2.4 Mixtures of network components form a punctuated equilibrium of stable states that coincide with behavior

After finding that mixtures of network components were linked to behavior and physiology, we wanted to understand how the status of all network components changed throughout the week. Up to this point, I have investigated how individual, or pairs of network components change over time. This is analogous to studying how a hummingbird transitions between hovering and flitting between flowers by only observing their movement in one or two spatial dimensions at a time. Just as how a complete picture of a hummingbird's flight patterns requires a full three-dimensional space, I studied brain trajectories through a high-dimensional neural space, with the dimensions defined by each of the network component's activity.

Figure 3.5A (left) shows the velocity of the brain over one participant's week: how quickly the participant's brain network activations changed every five seconds. This is analogous to the speed of a hummingbird's movement: high when it's flying, low when it's hovering. In this case, hovering means the network activations of someone's brain are remaining relatively steady while high velocity flight indicates that the network activations are changing rapidly. Quantitatively, velocity was calculated by taking the vector of all network activations from one time window and calculating the Euclidean distance between it and the corresponding vector from the next time window. The results indicated

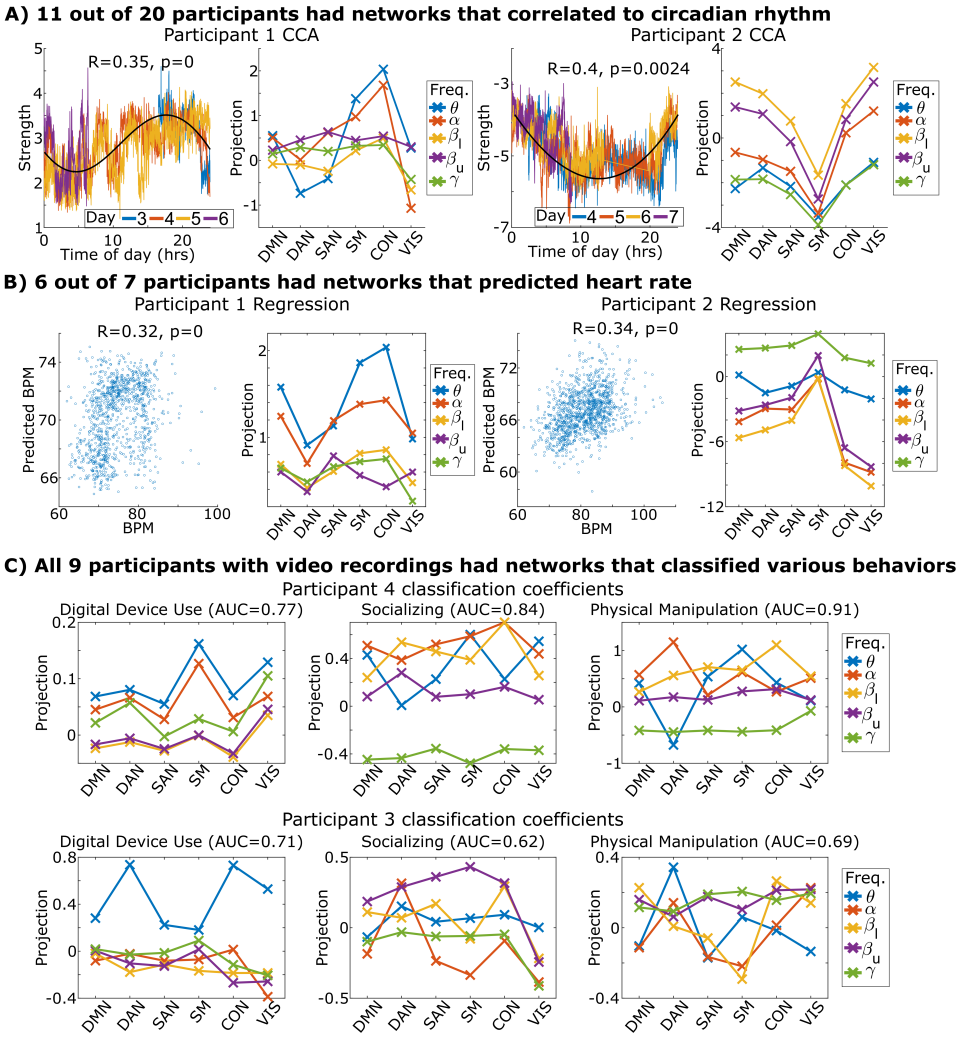

Figure 3.4: (A) We linked a network component mixture to circadian rhythm by training canonical correlation analysis on one half of the week and then testing on the other. The network mixture activations during testing are shown on the left plotted against time with the black line indicating a theoretical circadian rhythm. Skips in data are removals due to seizures or disconnected hardware. The identified mixture's anatomical and frequency coverage are shown projected onto the canonical fMRI networks. B) Network components were linked to heart rate by training linear regressors on one half of the week and testing on the remaining half. Test predictions are plotted against heart rate along with their anatomical and frequency coverage. C) Logistic classifiers identified network components that reliably detected whether the subject was watching a digital screen, socializing with another human, or physically interacting and manipulating a held object. The algorithm was trained over one randomly selected day and tested on another. The classifier's anatomical and frequency coverage is shown on the right. Area under the curve (AUC), a classifier performance metric, is provided in the figure title. All subjects for these analyses are shown in Supplementary Figures S7-S9.

that periods of stability (states), times of low velocity lasting for minutes-to-hours, were interspersed with bursts of dynamic behavior marked by high velocity for periods lasting up to minutes (state transitions).

To quantify that those times of high velocity occurred in "bursts", I tested the time between windows that fell into the top 1% or 10% of speed in each participant (Supplementary Figure S18). I found that windows with high velocity tended to occur temporally adjacent or close to one another (bursts of high velocity) at significantly higher rates than if they occurred randomly via homogeneous Poisson process, just as a hummingbird's high velocity periods occur in bursts. The rest of the analyses in this study examined the nature and statistics of these state transitions, defined as temporally contiguous bursts of high velocity over which the brain is continuously reconfiguring itself [22]. Long periods of stable states interspersed with bursts of high speed transitions is characteristic of "punctuated equilibrium" [23], an observation that many systems and processes in nature, particularly ones that involve adapting to a dynamic environment, do not undergo steady gradual change but rather periods of stability interrupted by rapid bursts of change.

The correspondence between behavioral and brain network transitions was next used to assess the relevance of these brain network state transitions. Specifically, we calculated the median time between a participant's behavioral changepoint (any point one of their three labeled behaviors changed) and the nearest identified neural state transition. We compared this to the expected time if there were no relation between the two using permutation testing. We found that for every participant, the median time was smaller than the expected time using a paired t-test (Figure 3.5C, $p<1e\text{-}4$). This result demonstrates that brain network transitions and behavioral transitions coincide with one another in time.

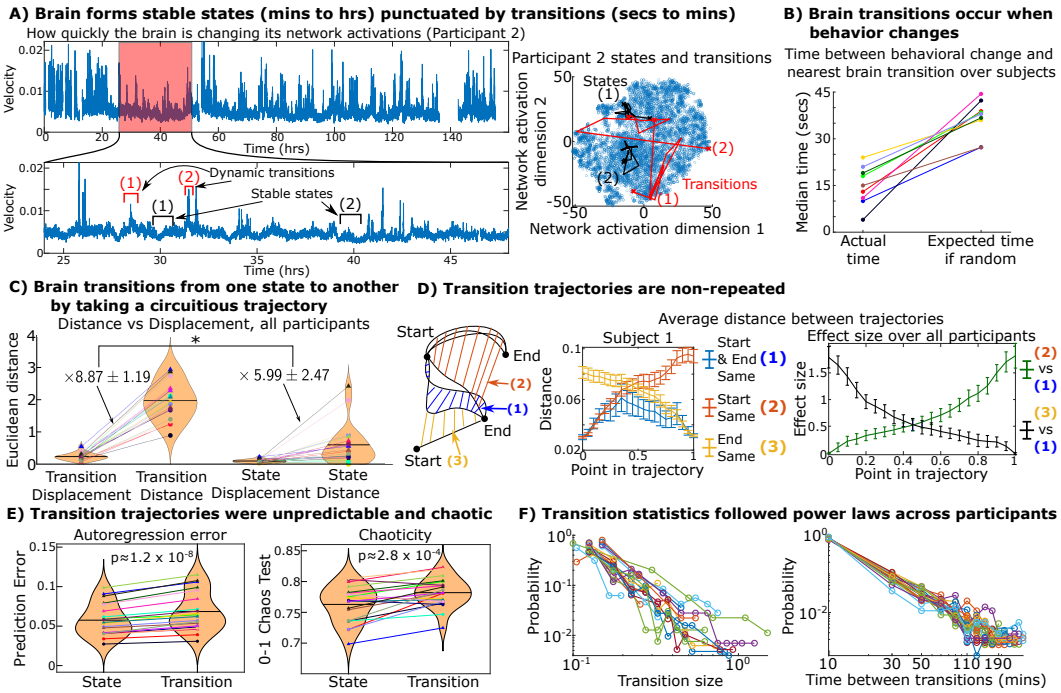

Figure 3.5: A (left) Overall change in network activations between consecutive time windows for one participant (Euclidean distance between the vector of all network activations from one time window to the next). Two segments of states and transitions are marked. A (right) T-distributed stochastic neighbor embedding visualized the week-long time course of network activations. (B) We detected when the brain state was transitioning using change-point detection and asked whether these times corresponded to when the participants' behavior changed. We compared the median time between behavioral change-points and the nearest brain transition to the expected time if random and found that neural and behavioral changepoints occurred together (p=1e-4 by paired t-test). (C) We tested whether brain transitions went directly from one state to another or whether they took indirect, circuitous routes. We plotted the average total distance (sum of velocity) traveled during transitions and stable states for all twenty participants versus their net displacement (distance between start and end state). Net distances for all participants were several times larger than the net displacement, indicating that transition trajectories were lengthy and indirect. The ratio between distance and displacement were higher for transitions than states (paired t-test). D (top) Average distance between pairs of transition trajectories as a function of what proportion of the trajectory was complete. Trajectory pairs are grouped into three categories: transitions with similar starting and ending points (1, blue) vs similar starting points only (2, red) vs similar ending points only (3, yellow). These distances are shown for one participant in the middle with all participants are shown in Supplementary Figure S10. The effect size of the differences between these distances over all participants is shown on the right. (E) We used autoregression to demonstrate that the predictability of brain dynamics decreases during transitions. We used a 0-1 chaos test to demonstrate that the chaoticity of brain dynamics rises during transitions. (F) The distribution of the size (net displacement) of each transition and the time between them is shown for all participants in log-log form. Both distributions formed power laws (linear on log-log axes) consistently across participants.

3.2.5 Brain network transitions are circuitous, unpredictable, and chaotic

After we found that stable networks and pairwise interactions could lead to the formation of neural states that lasted for minutes to hours and were associated with behavior, we asked how the brain transitions from one state to another. Specifically, does the brain take relatively straight trajectories when transitioning between states, as a hummingbird does when it moves between one hovering location to another, or are brain trajectories more circuitous. We visualized the transitory bursts of high velocity and stable states in Figure 3.5A (left) onto a t-distributed stochastic neighbor embedding (TSNE) representation of the week-long data in Figure 3.5A (right). TSNE is a data visualization technique that shows a two-dimensional representation of the data that preserves the distance between points plotted; thus, points that are close together are ones that have similar brain network activation vectors [24]. Visually, these transitions appeared to take very indirect trajectories. Instead of the brain transitioning directly from one stable state to another, it would appear to "wander around" and explore several possible intermediate states of various network activations or deactivations before stabilizing into the destination.

We quantitatively tested this by comparing the total "distance" traveled by the brain during a transition trajectory (the sum of the distance traversed during each step of the trajectory during a transition) compared to the net "displacement" (the distance between the starting and end states). For a straight-line trajectory (a hummingbird flying directly from one flower to another), the distance equals the displacement. Transition trajectories were defined as periods of high velocity surrounding detected change points (more details in Methods). All quantitative analyses were done in the original network activation space, TSNE was only used for visualization. Transitions across all participants showed total distances several times larger than the net displacement on average, indicating they were taking indirect routes between states (Figure 3.5C). Notably, the ratio of the distance to displacement was larger during transitions than during stable states, indicating that

between-state trajectories were more circuitous than within state trajectories.

These results demonstrate that transition trajectories are indirect, but indirect routes can still be consistent each time (e.g., when the brain transitions from stable states A to B does the path taken remain the same each time?). To assess their consistency, we compared transitions with the same start and end point both to transitions with the same start point but different end points and to transitions with different start points but same end point (Figure 3.5D and Supplementary Figure S10). If transition trajectories were repeated over the course of the week, the distance between pairs of transitions that started and ended in similar states (repeated transitions) would remain small compared to the distance between pairs of transitions that started in the same state but ended up in differing ones. On the contrary, the results indicated that transitions with the same end state were no more similar than transitions with different end states until 40-50% completion (Figure 3.5D). When they separated, the Cohen's d effect size of this divergence remained below 1 (less than one standard deviation apart) until 3/4 of the transition was complete. Thus, even if two trajectories had the same start and end point, the trajectories were typically very different from each other given that the distance between them was comparable to the distance between trajectories with different ending points for much of their journey.

Supporting the idea of diverse and hard to predict state transitions, autoregressive prediction error increased during transitions compared to within state periods (Figure 3.5D bottom). Autoregressive predictors were trained on half the week and tested on the remaining half, demonstrating lower error at predicting "within state" relative to "between state" movement. In addition, transitions exhibited increased chaoticity compared to within state dynamics, measured using the 0-1 chaos test [25]. Taken together, these results indicate that when the brain transitions from one stable state to another, not only would it rapidly explore a large set of intermediate states before settling down into the destination, but that these intermediate states seemed to be chosen in a disorganized,

chaotic-like manner.

3.2.6 Power laws indicate a consistent set of forces governing these transitions across subjects

I sought to reconcile the picture of simple neural networks with reliable pairwise interactions against the complex and chaotic-like transitory bursts that periodically spread throughout the brain. Many hypotheses investigating how complex neural processes can emerge from simple systems are distinguished by the distribution of various metrics summarizing salient features of their dynamics. Up to this point, I have primarily investigated the average value or variance of metrics on the brain's long-term dynamics such as timescales, mean chaoticity, or the predictability of neurocognitively interesting variables. Here I investigated the distribution of two simple metrics of neural state transitions: how large they are and how frequently they occur.

Figure 3.5F shows the distribution of the net displacement of these transitions and the time between transitions. These distributions visually followed power laws (linear on log-log plots), indicating that while transitions at first glance appeared to be disorganized and chaotic-like, there were overlying patterns governing the transition statistics that remained consistent from participant to participant. This overall dichotomy is illustrated in Figure 3.6.

I quantitatively tested this finding using likelihood ratio and Kolmogorov-Smirnov tests. I tested the likelihood that each subject's distribution came from a power law distribution vs exponential or log-normal distributions, finding that power law distributions were the most likely fit in the transition size distribution of 16 of 20 subjects ($p = 0.03$) and the most likely fit in all 20 time between transition distributions ($p = 4.8 \times 10^{-5}$). I used Kolmogorov-Smirnov tests to see if I could reject a null hypothesis that each subject's distribution plausibly came from a power law distribution, finding that I failed to reject in 17/20 subjects' transition size distribution and in 20/20 subjects' time between transition

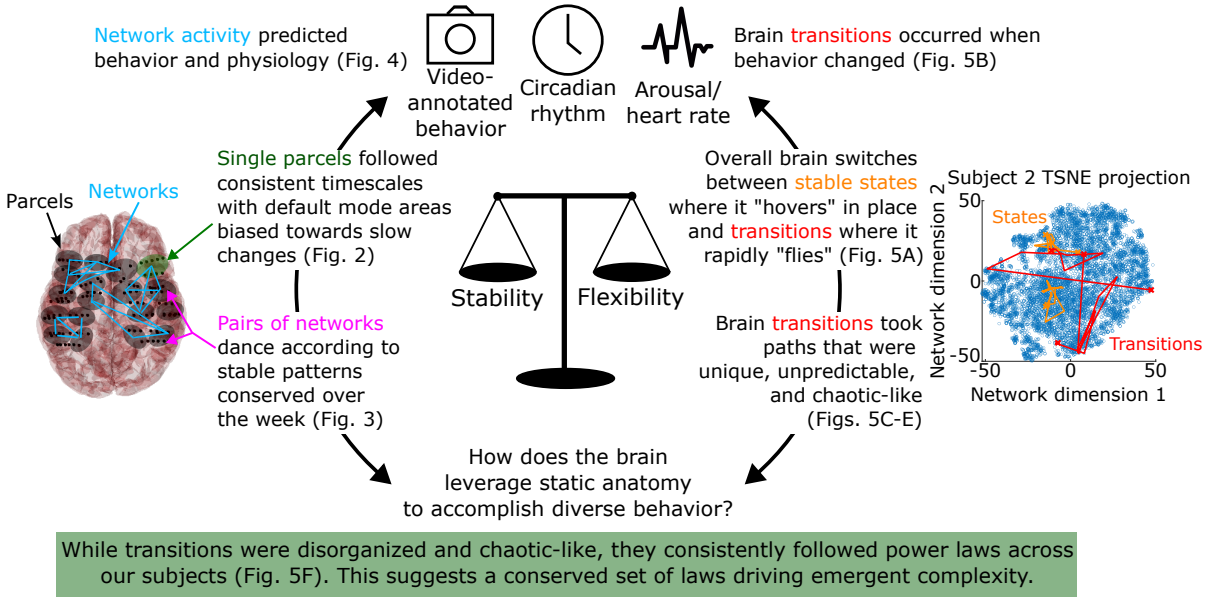

Figure 3.6: Summary of results presented in this manuscript and of the overall dichotomy between a brain made of simple, stable "parts" that is still capable of generating complex group-level behavior as it reacts to an ever-changing environment over long time periods.

distributions. Together, these tests support that the brain network state transitions follow power law distributions (Figure 3.5F) that occur when someone's behavior also changes (Figure 3.5B), linking power laws of the brain to real-world behavior and cognition. These results are consistent with the "critical brain hypothesis" which claims complex group-level behavior can emerge from simple and stable local activity in systems perched between order and disorder [26]

3.3 Discussion

The results of this study demonstrate key properties of how the brain continuously evolves over long timescales by recording roughly a week of near-continuous intracranial recordings in each of twenty human participants. The results indicate a dichotomy between relatively "local" brain dynamics that remained remarkably stable over time whereas overall brain trajectories and state transitions were varied and chaotic.

Individual parcels and networks had characteristic time constants and trajectories

that were well preserved across multiple days (Figure 3.2). Pairs of parcels and networks displayed characteristic, non-random "dances" that also remained stable over time (Figure 3.3). In contrast, the global mixture of brain networks displayed stable states punctuated by chaotic transitions between them (Figure 3.5). When the brain entered a stable state, the balance of its networks would remain relatively consistent for periods lasting from minutes to hours. We found that by using these networks, we could predict somebody's behavior (such as were they interacting with a screen or talking to a friend) and their physiological status (circadian rhythm and arousal; Figure 3.4). When somebody's behavior changes, potentially reflecting environmental or cognitive changes such as deciding to switch from reading a book to watching television or a friend walking into their room and beginning a conversation, their brain state would also change [27]. These transitions did not occur by the brain traveling from one state directly to the next but rather by circuitous, difficult to predict, and chaotic-like routes where the brain would explore many intermediate stages before settling down into a new stable state (Figure 3.5). While these transitions were difficult to predict and remarkably varied, their overall statistics (size and frequency) consistently followed power law distributions across participants (see Figure 3.6 for a summary).

Stable brain states interrupted by chaotic-like transitions are akin to punctuated equilibrium in evolutionary biology. Evolutionary history does not only show steady and gradual development but also alternates between periods of stability and transient bursts of rapid change [28]. These transitory periods are relatively disorganized: in evolution, these bursts involve phylogenetic "explosions" that generate multiple species or variants that quickly undergo environmental selection [23]. This also generalizes to large human organizations and political systems [29]. In most successful businesses, innovation efforts typically come in waves where an organization will explore several possible opportunities before settling on a much smaller number to develop [30] Indeed, some studies suggest that punctuated equilibrium is a hallmark of relatively efficient group decision-making

[31]. In our participants, their long-term neural dynamics would typically explore a large space before settling into stable behaviorally associated states.

Punctuated equilibrium may be how our brains deal with the chaotic and unpredictable real-world environment. Just as Darwinian evolution does not know the true end-optimal genetic state for the current ecosystem, our brains do not intrinsically know what will happen an hour in the future. Generating unpredictable and chaotic exploratory trajectories may be a key strategy of how our brains react to changing environments.

How can the brain quickly generate chaotic-like trajectories from simple and stable networks that are conserved over the week? While these trajectories were remarkably varied and difficult to predict, their overall statistics did consistently follow power laws across our subjects. Power law distributions have been a popular area of investigation in many fields across nature due to their "fat-tailed" divergence from the "typical" normal/log-normal or Poisson distributions that govern most systems. The "standard" distribution used to describe "when" events occur is the Poisson distribution: exponentially distributed time between events. Compared to rapidly decaying exponential functions, power laws have an increased probability of long times between events, aka long neural states. The same holds true for the size of an object: most natural systems form normal or log-normal distributions through the Central Limit Theorem, compared to which power laws showcase an increased probability of large cascades of dynamism that avalanche throughout the system.

All of this indicates that while these trajectories appeared at first glance to be almost "pseudo-random", consistently across our subjects, there was a consistent set of laws or forces driving their distributions whose effects we could identify.

What these laws are is an interesting question. One intriguing hypothesis would be self-organized criticality (SOC) from statistical mechanics which argues that systems made up of many small actors with simple interactions can manifest complex group-level

behavior by oscillating around a "critical point" – transition points between phases of matter or system behavior [32]. For example, while freezing and melting will spread simply from molecule to molecule to lead to ice cubes or simple liquids, a snowflake that is oscillated between melting and freezing dynamics records the cumulative history of all these oscillations which leads to their unique shapes. SOC argues that in systems that achieve complexity in this manner, their dynamics form power law distributions which have since been identified in a variety of fields ranging from how tectonic plates interact to form earthquakes, solar flares, and COVID-19 outbreaks [33–35].

SOC may engender several benefits to neural systems. Punctuated equilibrium, which is a necessary but not sufficient feature of critical systems, is seen in optimal group behavior and decision-making [31]. Modeling has demonstrated that being at a critical point optimizes the information processing and propagation capacities of a system [36]. Furthermore, being at a critical point yields a balance between the competing demands of flexibility to adapt to changes in the environment while maintaining stable representations of learned and predictable behavior [37].

One major caveat of this interpretation however is that while SOC is a compelling theory that could conceivably explain many of our findings, there are alternative ways to generate power laws such as long-range statistical processes, successive fractionation, and combinations of exponentials [38, 39]. In general, the dynamics I found could stem from a variety of causes ranging from neuroanatomy, endogenous physiological variation, environmental variation, behavioral trends, the behaviors of others, and a host of other factors.

Ultimately, the primary purpose of this work is to demonstrate that we can identify properties of how the brain changes over minutes-to-days in the ever changing real-world during natural behavior. Some of these properties remain consistent within a single subject over time and some remain consistent between different subjects. Many of these patterns are also meaningfully connected to someone's physiological and cognitive status.

Studying brain dynamics at this scale enables the study of cognitive and physiological processes inaccessible on shorter timescales. Someone's attention, mood, and arousal oftentimes fluctuate on the order of hours-to-days. Physiological changes, such as dynamics related to circadian rhythms, hormones, and gene expression do the same. Recent technological advances providing the ability to record both neural activity [40] and physiological biomarkers [41] in an animal's home environment can provide a fine-grained view into the cell-to-circuit neural behavior underlying cognitive and physiological fluctuations over hours-to-days. Clinically, many neuropathological states evolve and fluctuate over hours-to-days-to-years. In humans, chronic and continuous neural recordings that are performed as standard of care for certain patient populations (including fully natural and deployable recordings in patients with certain deep-brain stimulation systems [42]) can provide the opportunity to study real-world neural behavior on this timescale, both to understand basic neural behavior (as in this study) and to better understand their pathology. Stable and deployable wearable technologies for non-invasive neural recordings in real-world setting are also starting to be developed [43]. I hope this chapter demonstrates that these advances can lead to a new field of neuroscience dedicated to studying how the brain slowly changes in the chaotic real world.

3.4 Methods and Supplement

3.4.1 Subjects

Twenty participants (nine males, 11 females; mean age 40 years with a standard deviation of 12 years) had intracranial surface or depth electrodes implanted for the treatment of pharmacologically intractable epilepsy. All participants gave informed consent to participate under research protocols approved by the University of Pittsburgh Institutional Review Board. Depth electrodes were produced by Ad-Tech Medical and PMT and were 0.86 and 0.8 mm in diameter, respectively. Grid electrodes were produced

by PMT and were 4 mm in diameter. Sixteen participants had depth electrodes only, three had grids only, one had a combination of both.

3.4.2 Analysis overview

In summary, we collected and preprocessed lengthy intracranial recordings which we divided into five second windows (Figure 3.1A-B), calculated the functional connectome of each window via all-to-all electrode coherence (Figure 3.1C), grouped electrodes into tightly connected parcels which we analyzed (Figure 3.1D), grouped parcels and frequencies into functional network components using Principal Components Analysis (Figure 3.1E), and then studied the overall mixture of all functional networks (e.g. all principal components, Figure 3.1F). Artifacts were removed at multiple points in the analysis. Specifically, a comb filter was applied to remove line noise, an hour before, during, and after all seizures were removed to eliminate ictal and peri-ictal activity, spatial regression was used to remove local and global artifacts (such as motion, respiratory, and cardiac artifacts that tend to be similar in neighboring electrodes), ICA was used to remove large spike artifacts that sometimes occur due to disturbing the cables or connections, and epileptogenic areas and activity that correlated with the activity in these regions was removed to eliminate interictal activity or other pathological activity.

3.4.3 Intracranial EEG data collection

Twenty patients (nine males, 11 females; mean age 40 years with a standard deviation of 12 years) had intracranial surface or depth electrodes implanted for the treatment of pharmacologically intractable epilepsy (Figure 3.1A). All subjects gave informed consent to participate under research protocols approved by the University of Pittsburgh Institutional Review Board. Depth electrodes were produced by Ad-Tech Medical and PMT and were 0.86 and 0.8 mm in diameter, respectively. Grid electrodes were produced by PMT and were 4 mm in diameter. Sixteen subjects had depth electrodes only, three had grids only,

one had a combination of both.

Electrodes were localized via postoperative MRI or CT scans coregistered to the preoperative MRI using Brainstorm [44]. Surface electrodes were projected to the nearest point on the preoperative cortical surface automatically parcellated via Freesurfer to correct for brainshift [45, 46]. Electrode coordinates were then coregistered via surface-based transformations to the fsaverage template using Freesurfer cortical reconstructions.

Intracranial electroencephalography data was collected using the Natus system at 1kHz. Notch filters at 60/120/180Hz were applied with a subsequent bandpass filter from 0.2 to 115Hz. The spatial autocorrelation between an electrode and all electrodes within 2cm of it was then measured and regressed out to eliminate local and global artifacts, including motion and current spread due to volume conduction. Segments of time around all seizures, electrographic or clinical, were removed starting an hour before and an hour afterwards before calculating coherence (Figure 3.1B). A board-certified neurologist identified the seizure network in all but two patients, with those two patients having no recorded seizures during their stay in the hospital.

The data was then separated into five second windows (Figure 3.1B) with coherence computed over each window between all pairs of electrodes over five frequency bands: theta (4-8Hz), alpha (8-12Hz), low beta (14-20Hz), high beta (20-30Hz), and gamma (30-70Hz). Using scipy's coherence function under default settings (version 1.9.3), we generated coherence matrices between all pairs of electrodes at 1Hz, 2Hz, ...,70Hz which were then averaged according to the five frequency bands. In summary, this generated five connectome structures every five seconds (Figure 3.1C).

Independent component analysis was then applied using sklearn's FastICA implementation under default settings as of version 1.2.1, and components were visually inspected for any artifacts which were then removed. Our criteria for removal were independent components that possessed time course activations that were clearly non-neurological

(such as step-functions or near dirac deltas).

3.4.4 Parcellation (Figure 3.1D)

For each subject, we parcellated their electrodes into groups of tightly coherent electrodes (Figure 3.1D). We utilized the Leiden algorithm to identify a single regional atlas that optimized graph modularity over the entire week-long period across all five frequency bands46. Modularity (Equation 1) was calculated separately over each network from every five-second window with the Leiden algorithm optimizing the average modularity across all windows and frequencies. This generated on average 10-15 parcels for each patient.

For each subject, we parcellated their electrodes into groups of tightly coherent electrodes (Figure 3.1D). We utilized the Leiden algorithm to identify a single regional atlas that optimized graph modularity over the entire week-long period across all five frequency bands [47]. Modularity was calculated separately over each network from every five-second window with the Leiden algorithm optimizing the average modularity across all windows and frequencies as defined in Equation 3.1. $A_{i,j}^{b,t}$ refers to the weighted connectivity (coherence) between electrodes i and j at time window t and frequency band b. $k_i^{b,t}$ is the degree of electrode i and $m^{b,t}$ is the sum total of all connections at that time and frequency. $\delta(c_i, c_j)$ is an identity function as to whether electrodes i and j are in the same parcel. This process generated on average 10-15 parcels for each subject.

$$\text{modularity} = \sum_{b \in \{\theta,\alpha,\beta_l,\beta_u,\gamma\}} \sum_t \sum_{i,j} \left[A_{i,j}^{b,t} - \frac{k_i^{b,t} k_j^{b,t}}{2m^{b,t}} \right] \delta(c_i, c_j) \tag{3.1}$$

To assess the stability of which electrodes would be grouped into which parcels, we separated the data into six-hour non-overlapping segments (between 18-80 segments per subject) and found the optimal community structure for each chunk. We quantified the similarity between each segment's parcel definitions using the Rand Index [48] (percentage

of electrode pairs that were parcellated equivalently under the two parcel definitions) which almost universally returned values greater than 0.9 as illustrated in Supplementary Figure S2, indicating that the overall parcellation was well-preserved over time, motivating our decision to use the same parcel structure over the entire work for interpretability.

3.4.5 Autocorrelation stability (Figure 3.2)

We tested whether the autocorrelation of each parcel's coherence would show consistent patterns of "fastness" or "slowness" throughout the week (Figure 3.2). We split the week-long time course for each subject into six-hour non-overlapping segments. After removing parcels associated with the seizure network, we then took the average coherence between electrodes within a single parcel for a single frequency band and then calculated its autocorrelation up to one hour. We fit this autocorrelation curve to a power law ($\text{autocorrelation}(t) = \text{AC}_1 \times \text{time}^{-\text{AC}_2}$) to generate two timescale parameters: AC_1 (autocorrelation strength) and AC_2 (autocorrelation steepness) which described the autocorrelation of a single parcel at a single frequency at a single time segment. For a given frequency band, we took the timescale parameters across all parcels and time segments and grouped the parameters by which parcel they were measured in. We used Kruskal-Wallis ANOVA tests to show that in almost all subjects and frequency bands, there were statistically significant differences between the group means, mostly with high effect sizes ($\eta > 0.12$, Supplementary Figure S3).

We tested whether parcels from different anatomical regions tended to have reliable differences in their autocorrelation across subjects using linear mixed effect models48. We assigned each parcel to a lobe and canonical fMRI network based on its largest overlap. Parcels with less than 60% of their electrodes belonging to the same anatomical group were excluded for this analysis. For each parcel, we calculated the autocorrelation of its average intra-parcel coherence for a given frequency over the entire week out to one hour and calculated AC_1 and AC_2 as described above. We then averaged both parameters

across all frequency bands.

We then chose a single pair of lobes or fMRI networks (such as frontal vs temporal) and selected all the parcels across our subjects that fell into one of those two anatomical groups. We used MATLAB's fitlme (linear mixed effect model) to model each parcels autocorrelation parameters with both the subject and the anatomical group as fixed-effects, allowing us to determine whether one anatomical group had a reliably higher autocorrelation parameter than the other. We repeated this for all possible pairs of lobes/fMRI networks and used Bonferroni multiple comparisons correction to identify pairs with significant differences (Figure 3.2B).

3.4.6 Robust principal components analysis (Figure 3.1E)

We found that many parcels tended to be highly colinear with each other as shown in Figure S17. To reduce dimensionality, we used a modified PCA protocol to take advantage of this finding. While the results in this manuscript still hold statistical significance without this dimensionality reduction as shown in Supplementary Figures S12 to S16, parcel pair-wise interactions are less interpretable due to a high degree of covariance between them, and some classifier performance degrades due to the added noise.

We grouped parcels and frequencies that tended to covary together using random sample consensus PCA (RANSAC-PCA) on the parcel coherences (Figure 3.1E). By taking the average intra-parcel coherence during each time window and frequency, we formed a (number of parcels x 5 frequency bands) by (number of time windows) 2D matrix which we then reduced to a (number of components) by (number of time windows) matrix using a modified PCA protocol. This identifies parcels and frequencies that tend to strongly covary together that we could easily interpret as a single network component feature that captures cross-frequency relationships while also reducing the dimensionality of the original dataset to simplify further analyses.

Our modified PCA protocol uses random sample consensus to avoid PCA's susceptibility

to noisy outliers by attempting to exclude outliers by taking multiple small subsamples of the data and selecting one with the fewest number of outliers to train the model [17, 49]. We generated 1000 subsamples where in each subsample, we selected six 30-minute segments of data from each day. Outliers were defined by calculating the Mahalanobis distance between each time window's feature vector and each subsample's distribution. In each subject, we found that these distances would take on clear bifurcations between relatively small distances and short "spikes" of extremely high distances (more than three standard deviations) away from the mean that typically lasted for a few minutes. We manually drew a cutoff for each subject that was approximately half the average Mahalanobis distance of these spikes. For each subsample, we calculated the number of outliers within the subsample, and calculated PCA over the subsample with the fewest outliers. We utilized enough PCs to capture 90% of the variance in the dataset, generally resulting in 12-24 networks/PCs per subject.

The network component activation of a principal component was defined as the projection of the parcel coherences onto the principal component weights. We repeated the same autocorrelation stability analysis described above on each network component's activation (Supplemental Figure S5).

3.4.7 Seizure network removal

When analyzing parcel dynamics (Figure 3.2), we excluded all parcels with electrodes part of the seizure onset zone and early propagation as defined by a board-certified neurologist. For network component dynamics (Figure 3.3 onwards), we first re-added these seizure-related areas before grouping parcels and frequencies into network components through robust principal component analysis. We then removed any network components that were associated with the seizure network before analyzing their dynamics. More specifically, we calculated the dot product similarity between the absolute value of a principal component vector (normalized to a magnitude of one) and a binary vector that

marked all electrodes that were part of the seizure network (also normalized to one). The similarity between these two vectors indicated how anatomically similar the driving factors of a principal component and the seizure network were to each other. A null distribution for this similarity was formed by randomly permuting the principal component vectors, and all principal component vectors that showed statistically significant similarity to the seizure network ($p < 0.05$) were removed from all further analyses.

3.4.8 Network components show non-independent relationships that are well-preserved over days (Figure 3.3)

We tested whether network components had reliable interactions with each other by examining their joint distribution stability. For each pair of network components in a subject, we divided the total range of each component's activation into 1000 discrete bins to generate a 1000 x 1000 grid covering the space of both components' activation. We calculated the empirical joint distribution of the network component pair over these bins for each day separately. We calculated the Bhattacharyaa distance [50] between each day's distribution to each other as well as the distance to the expected joint distribution if each network were independent but possessed the same marginal distribution (which we denote the expected independent distribution). The "effect size" metric shown in Figure 3.3 is the average distance between the real distributions on different days divided by the distance to the expected independent distribution. Using permutation testing (10k trials), we established a null distribution for effect size if the networks were in fact independent by randomly drawing however many days of samples we had from the expected independent distribution and calculating the effect size over these draws. Statistical significance was then corrected for multiple comparisons across all possible pairs of joint network distributions using Bonferroni correction.

3.4.9 Network component activation is tied to circadian rhythm, heart rate, and behavior (Figure 3.4)

We tested whether we could identify combinations of network components that were associated with neurophysiologically relevant markers. More specifically, we looked at circadian rhythm, heart rate, and behavior.

Canonical correlation analysis (CCA) was used to identify a mixture of network components that matched a circadian sinusoid with a period of 24 hours. The circadian sinusoid was defined as $a_1 cos(t/24\text{hrs}) + a_2 sin(t/24\text{hrs})$ where a_1 and a_2 are constants learned by CCA. CCA simultaneously tried to find a linear combination/weighting of network component activations to fit to this sinusoid.

The model was trained over the first half of the week and then tested on the second half through Pearson correlation (out of sample validation of correlation). The Pearson R of the fit on the test dataset was calculated and then compared to a null distribution of R that was formed via permutation tests that temporally shifted each day's network component activity forward or backwards by a uniform random number ranging from 0-24 hours. This preserves the autocorrelation of the neural signals while eliminating any consistent circadian-like pattern across days.

Heart rate was assessed using collected EKG signals that were processed using heartpy [51]. The instantaneous heart rate for any window was the average heart rate for a 30-second period centered on the window. L1-regularized regression was trained on the first half of the week to identify a mixture of networks that predicted heart rate using sklearn's implementation (out of sample validation of regression). Hyper parameterization was optimized on the training set using ten-fold cross-validation. The quality of the fit was assessed on the remaining half via Pearson correlation with a null distribution formed using the same permutation tests used for circadian rhythm to preserve both the autocorrelational properties of the heart rate and neural signals.

Video and audio recordings of the subjects from two separate angles were used to assess the subject's behavior over two randomly selected days via manual annotation. Digital device usage was defined as looking at any digital screen, such as a smartphone or laptop. Socialization was defined as verbal communication with another human being, or in one case a canine companion, either in person or over the phone. Physical manipulation was defined as actively grasping and interacting with any physical object or person. These three behaviors were not mutually exclusive with one another.

Windows with the desired behavior were manually annotated on two separate days of data. L1-regularized logistic classification was used to identify a mixture of network components that classified each behavior independently using one day for training and the other for testing using sklearn's implementation (out of sample validation of classification). Hyper parameterization was optimized on the training set using ten-fold cross-validation. The area-under-curve of the receiver-operator-curve of each network's ability to classify the desired behavior was calculated.

3.4.10 Transitions in the overall brain state fall into a punctuated equilibrium (Figure 3.5)

We examined how the overall brain state (the status of all recorded network components in the brain) would change over time by dividing the week into "transitions", periods when the brain was rapidly reconfiguring itself, and "states", periods of time where the brain's functional connectome appeared to be relatively stable.

In Figure 3.5A and Supplementary Figure S18, we provide evidence that the brain falls into states and transitions by examining the "velocity" of the brain. Velocity was defined as how much the brain's state changed between one five-second window and the next. More specifically we took the vector of all network activations of each window (the parcel coherences projected into the network PCA space) and calculated the Euclidean distance between the network activation vector of one window and the next. Speed is technically a

more appropriate term than velocity from a kinematics perspective. However, we avoided the term speed due to its connotation with cognitive processing speed which is unrelated to this analysis.

We calculated the distribution of the time between windows that fell into the top 1% of velocity and compared that distribution to Poisson distributions with $\lambda = 0.01$. The Poisson distribution captures what the expected time between high-speed windows would be in a memoryless process (non-autocorrelated speed). By examining whether windows with high speed tended to cluster next to each other temporally, we show that there are specific periods of time when the brain is quickly changing and times when the brain is relatively static. We also tested this between windows falling within the top 10% of velocity ($\lambda = 0.1$) and found the same result.

In Figure 3.5B, we evaluated whether state transitions were linked to behavioral changepoints by marking out times when the participant's behavior changed in any of the three categories. We calculated the median time between this changepoint and the nearest state transition (zero if the changepoint occurred during a transition). We calculated an expected value for this metric assuming no relationship by temporally shifting the behavioral changepoints forward or backwards by a uniform random number ranging from 0-24 hours, calculating the median time difference, and then averaging over 1000 trials. We tested whether the real time difference between behavior changes and state changes was consistently smaller than the expected time difference using a non-parametric t-test across participants (paired t-test).

Next, we analyzed properties of these transitions (Figures 3.5C-E). We defined transitions in two different ways to ensure our conclusions were not overly method dependent. The transitions described in the main text figures were identified using binary segmentation change point detection on the overall brain state. The transition trajectory was defined as the period around a change point that possessed above-average velocity (the Euclidean distance between the vector of all network activations from one time point to another). If

the trajectories associated with two neighboring change points overlapped, the change points were "merged" into one.

We replicated Figures 3.5C/D with a different trajectory definition in Supplementary Figure S11. We calculated the velocity over the week and took the moving average of it over a 30 second window and defined trajectories as periods of time when this smoothed speed reached the top 20% quantile over the week. We replicated Figure 3.5E in Supplementary Figure S11 where we found the main relationships highlighted in these figures still held true.

In Figure 3.5C, we examined the "distance" and "displacement" of these transitions. During a transition, we calculated the speed between each window. The distance was the sum of those speeds. The displacement was the distance between the first window of the transition and the last. This comparison allowed us to determine how direct the transitions were because in a "straight line" transition, distance would equal displacement, but if distance is much greater than displacement, then the transition is more circuitous.

In Figure 3.5D (top), we examined the distance between the paths traversed by different transition trajectories. We calculated the distance between the start points of all trajectories in a participant and the distance between their end points. Two trajectories were considered to have the same starting or ending point if the distance between the points fell into the bottom 10% of trajectory pairs. We grouped trajectories into three groups: trajectories with the same starting and ending point, trajectories with the same starting point only, and trajectories with the same ending point only. We calculated the average distance between trajectories that fell into each group as a function of how much of the trajectory has been completed. More specifically, we used linear interpolation to determine what was the brain state $5\%, 10\%, 15\%, 20\%, \ldots, 95\%$ of the way into each trajectory. We calculated the distance between brain states of the same percentage in each of the three groups. Figure 5D shows the distribution of these distances for a single participant along with the effect size of the difference between these distances across all

participants. For effect size, for each subject individually, we calculated the Cohen's d between the distances between trajectories that start and end the same to the distances between trajectories that start the same but end differently or trajectories that end the same but start differently. This measures the number of standard deviations that separate the distributions in the trajectory categories at different points along the trajectory. We then calculated the standard error of these Cohen's d across all 20 subjects. The average Cohen's d and these standard errors are shown in Figure 3.5D.

In Figure 3.5E, we studied whether these transitions influenced the autoregressive prediction error and chaoticity of the brain dynamics. Our autoregressive model was vector autoregression with the number of previous time-steps being selected using Bayesian Information Criterion. These models were trained and evaluated using cross-validation by "holding out" one day of data and training on the remainder of the week. The predictive error was the average mean-squared error on the held-out day during either transition or state dynamics. Chaoticity was defined using the 0-1 chaos test using the protocol described in [25] and was calculated over non-overlapping ten minute segments. In summary, we calculated the chaoticity of each network component independently over each time segment. The chaoticity of the overall neural dynamics for a given time segment was defined as the median chaoticity of all network components. Segments with a transition were compared to segments without one.

The 0-1 chaos test is described in Equations 3.2-3.5. Define $\phi(n)$ as the network component activation of interest at time window n for a given ten-minute segment. This is used to "drive" the dynamical system described in Equations 3.2 and 3.3 where c is a randomly chosen "resonance" parameter between 0 and π that remains constant during a single "iteration" of this process. $M(n)$ is evaluated up to an n of $N/10$ where N is the number of time windows in the ten-minute segment. K_c is estimated by fitting a straight line between the numerator and denominator of Equation 2.4 and represents the chaoticity of a single iteration. c is redrawn 1000 times and the median K_c is defined as

the chaoticity of the network component over the ten-minute segment.

$$p(n+1) = p(n) + \phi(n)\cos cn \tag{3.2}$$

$$q(n+1) = q(n) + \phi(n)\sin cn \tag{3.3}$$

$$M(n) = \lim_{n\to\infty} \frac{1}{N} \sum_{j=1}^{N} \left([p(j+n) - p(j)]^2 + [q(j+n) - q(j)]^2 \right) \tag{3.4}$$

$$K_c = \lim_{n\to\infty} \frac{\log M(n)}{\log n} \tag{3.5}$$

In Figure 3.5E, we analyzed the distribution of the transition size which we defined as the net displacement of a transition and the time-between transitions. We fit power law exponents to these distributions using MATLAB's nlinfit function with power laws defined as $a_1 \times \text{frequency}^{-a_2}$ where a_1 and a_2 are learned.

We tested whether these distributions came from power law distributions using two methods from [52]. First, we used Kolmogorov-Smirnov (KS) tests to test whether we failed to reject the null hypothesis that the distributions plausibly came from power laws. We fit power law distributions to each subject's transition size and time between transitions distributions separately and calculated the KS distance between the experimental distributions and their theoretical power law ones. We formed a null distribution on these distances by drawing 1000 random samples from the theoretical power law distribution, fitting a power law distribution to those samples, and then calculating the KS distance between the sampled distribution and the fitted one. If these distances were consistently lower than the distance between the real distribution and its estimated power law one, then we reject the null and conclude that the distribution did not come from a power law. We found that 17/20 subjects had transition size distributions that plausibly came from power law distributions ($p > 0.05$), and 20/20 subjects had time-between transition distributions that plausibly came from power law distributions.

We then used likelihood comparison tests to see whether the transition size and time-between distributions were more likely to have come from power law, exponential, or log-normal distributions. We calculated the log-likelihood that each subject's distributions came from each of the three categories. We used a Wilcoxon signed-rank test to test whether the log-likelihood of power law distributions were higher than exponential and log-normal distributions across subjects. We found that power law distributions were more likely than exponential (p=0.007 for transition size and p=4.8e-5 for time-between) and more likely than log-normal (p=0.03 for transition size and p=4.8e-5 for time-between).

3.4.11 Supplemental Figures

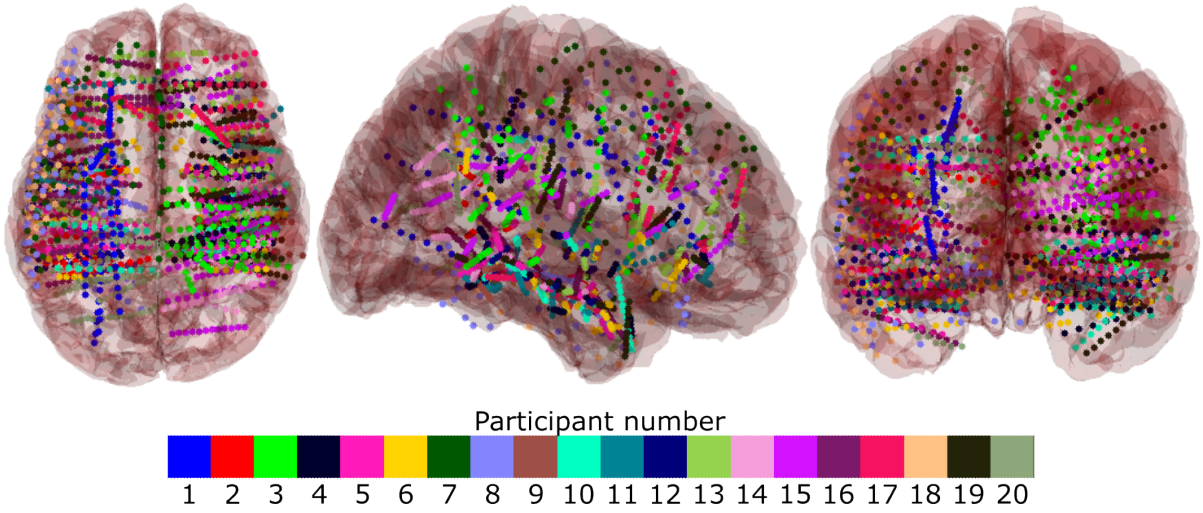

Figure S1: The location of each subject's electrodes in MNI coordinate space. Electrodes with the same color come from the same participant.

In Figure 3.2A, we showed an example from two parcels that consistently had different timescales throughout the week: e.g. a parcel with a slow timescale on one day would remain slow on other days. In Figure S3 we quantitatively tested this across all parcels for all subjects by taking each subject's weeklong period and dividing it into six-hour non-overlapping windows and asking if some parcels had consistently faster or slower timescales across windows using ANOVA tests.

We replicated the results shown in Figure 3.2A and Figure S3 using the timescales of

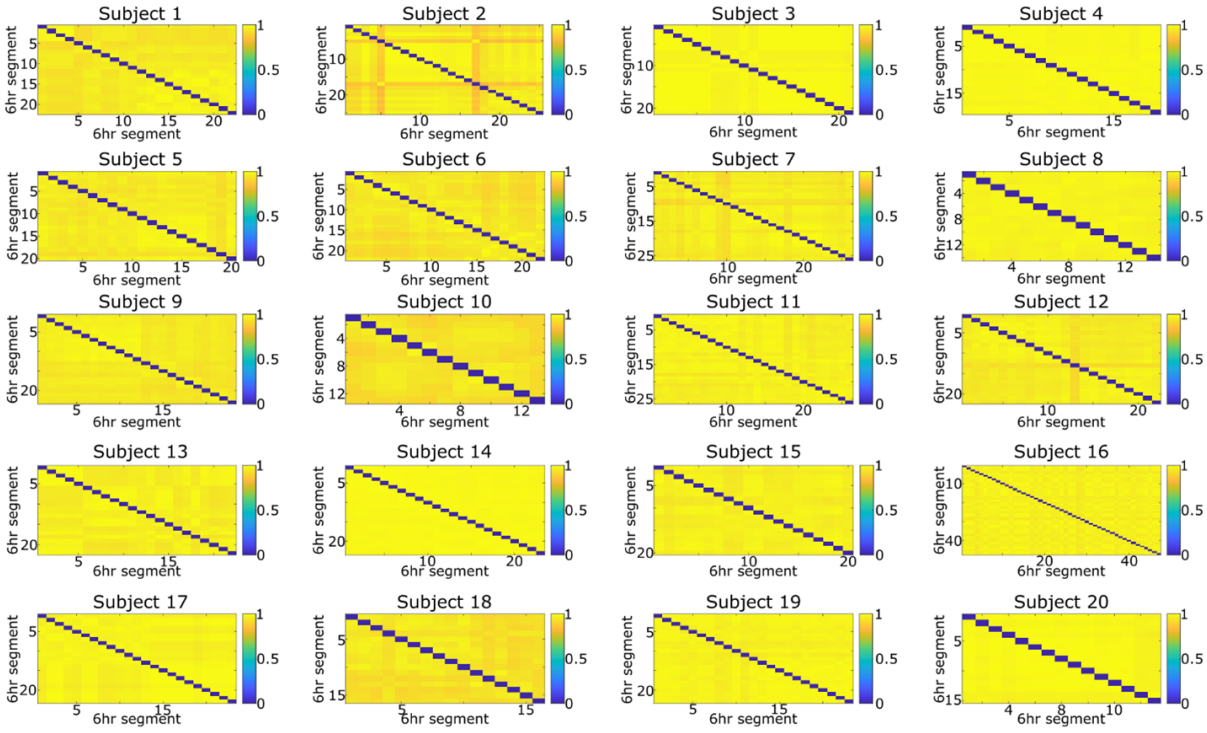

Figure S2: Parcellation stayed steady over time. We divided each subject's data into non-overlapping six hour blocks and computed the optimal parcellation for each block. The Rand Index, which represents the proportional overlap between two parcellations of electrodes, is shown between all six hour blocks for each subject. All subjects generally show over 90% similarity between blocks, indicating that the parcellation schemes discovered are stable over time.

network components rather than individual parcels in Figure S5.

Figures S7 to S9 show the results in Figure 3.4 for all subjects.

We replicated the findings of Figures 3.2-3.5 without using PCA to group parcels into network components to show the robustness of our results to changes in our methods in Figures S12 to S16. In other words, instead of seeing how the (network component activation x 1) vector would change from window to window over the week we tested how the coherence of each parcel at each frequency band (parcels over frequencies x 1) vector would change over the week. We found that the main results still held true, however, some classifier performance unsurprisingly degrades without the dimensionality reduction provided by PCA.

We primarily show the results of using PCA in the manuscript since many parcels

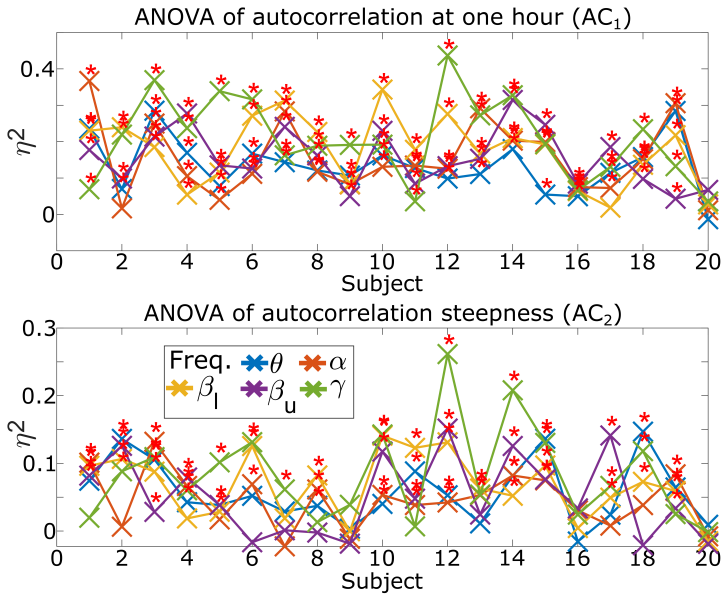

Figure S3: We assessed whether parcels had consistently different timescales using nonparametric ANOVA tests. For each subject individually, we divided their weeklong time-course into six-hour non-overlapping blocks. We calculated the autocorrelation of each parcel's coherence at a given frequency band (θ: theta, α: alpha, β_l: low beta, β_u: high beta, γ: gamma) across all blocks. We then tested whether the parcels from a single subject and frequency band had different autocorrelations from each other over these blocks using a Kruskal-Wallis one-way ANOVA test. Each group in the ANOVA test was the autocorrelations of a single parcel across all blocks, and we tested for whether there were differences in the group means. The effect size of the ANOVA test is shown above with asterisks marking statistically significant differences (p¡0.05). η^2 effect size indicates the percentage of variance in autocorrelation that is explained by which region autocorrelation was measured in.

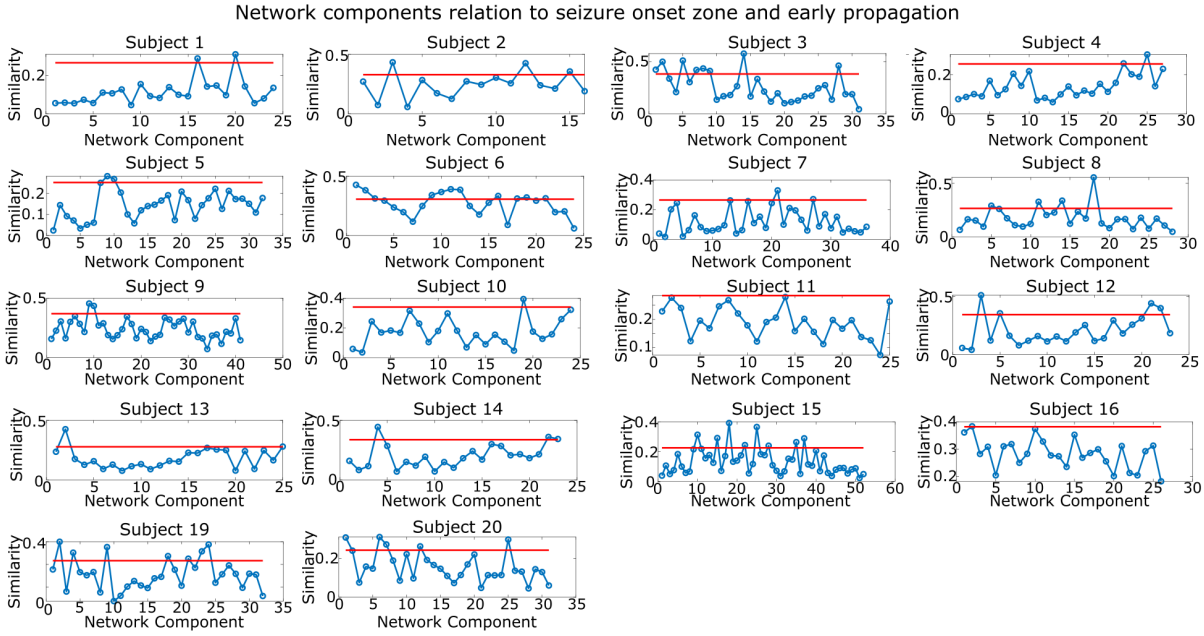

Figure S4: To ensure we remove all network components related to the subject's seizure related areas, we calculated the similarity between each subject's network component and their seizure zones and removed any network components showing above-chance similarity. Seizure zones were defined as any electrodes marked as part of the seizure onset zone or early propagation. Similarity was defined as the dot product between the absolute value of each subject's network component and their seizure zone and is shown above for all 20 subjects. Subjects 17 and 18 did not have any determined seizure network. A null distribution for the dot product similarity generated by randomly permuting each network is shown with the red line to denote statistical significance threshold (p=0.05). All network components with significant similarity to seizure related regions were removed for all analyses on network components (Figure 3.3 and onwards).

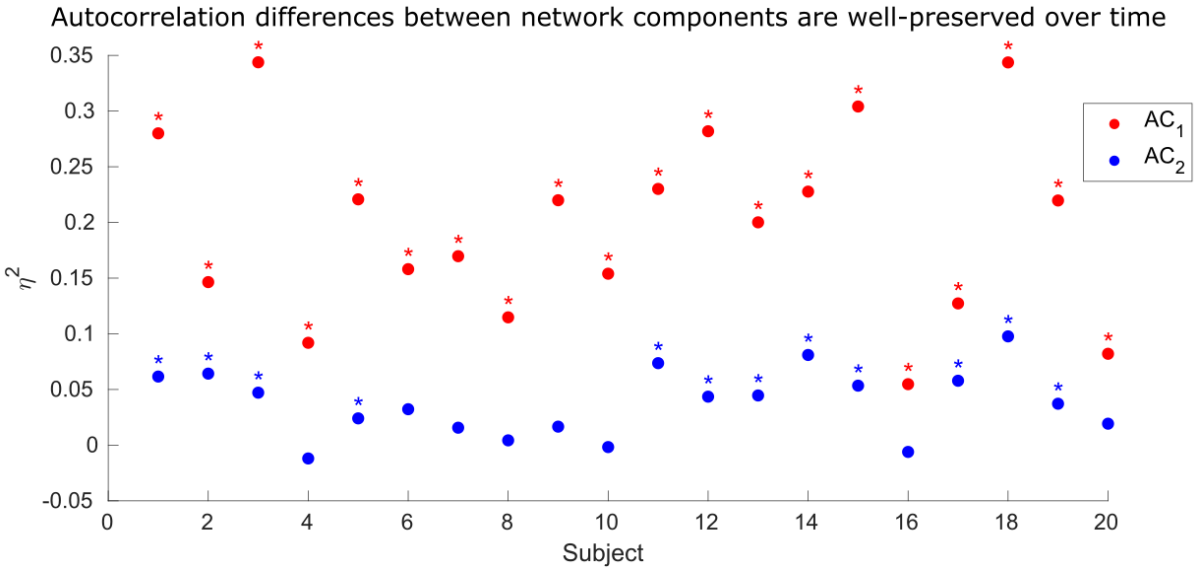

Figure S5: (A) To quantify the stability of a network's timescale in the same way we did in Figure S3, we used a nonparametric ANOVA test to see if different six hour blocks of a network had reliably different autocorrelation from other networks. Asterisks mark subjects that had a statistically significant difference in the autocorrelation across their components in either autocorrelation at one hour (AC_1, blue) or autocorrelation curve steepness (AC_2, red).

tended to be highly colinear with each other as shown in Figure S17. Additionally, since we measure each parcel's coherence across multiple frequency bands, PCA allows us to group parcels and frequencies with strong co-linearities into network components that reduce the dimensionality and noise within our dataset. This also coincides with the neuroscience definition of network: group of parcels/regions that tend to covary together.

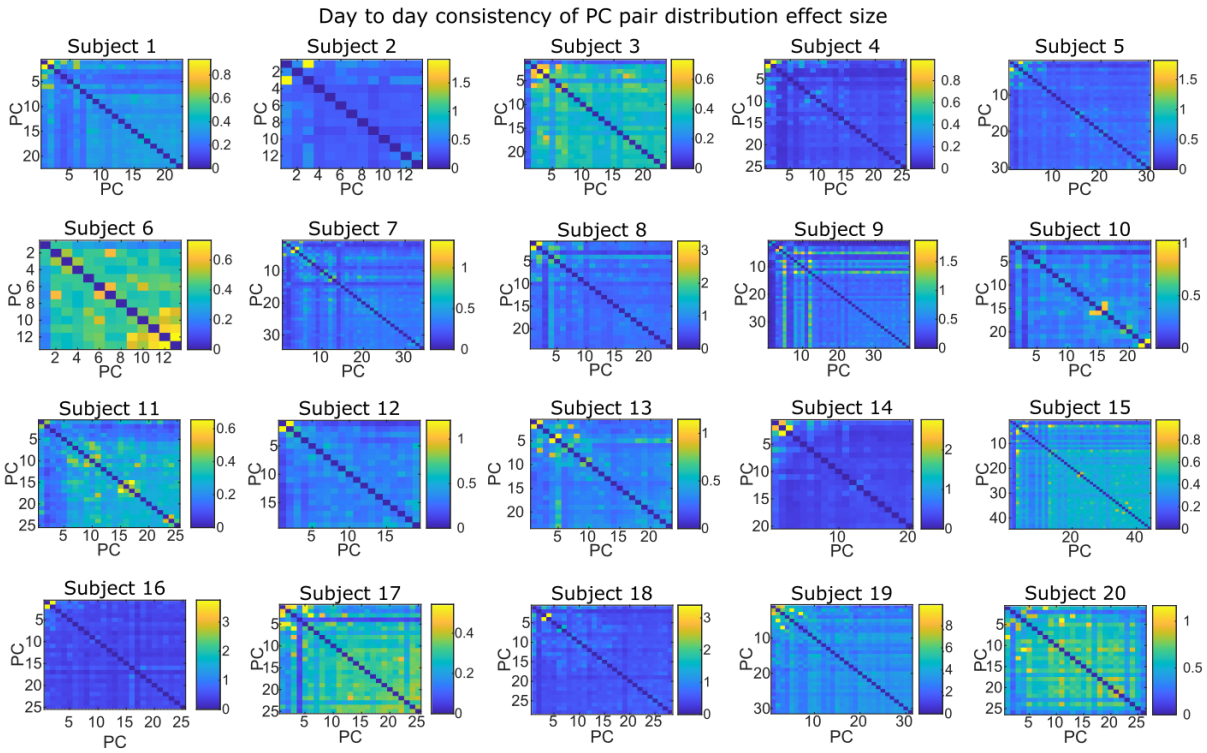

Figure S6: This figure shows the results in Figure 3.3B for all twenty subjects. The effect size of the distance between the joint distribution of each pair of networks/principal components compared to if they were independent. Non-zero effect sizes represent statistically significant distances determined via permutation testing.

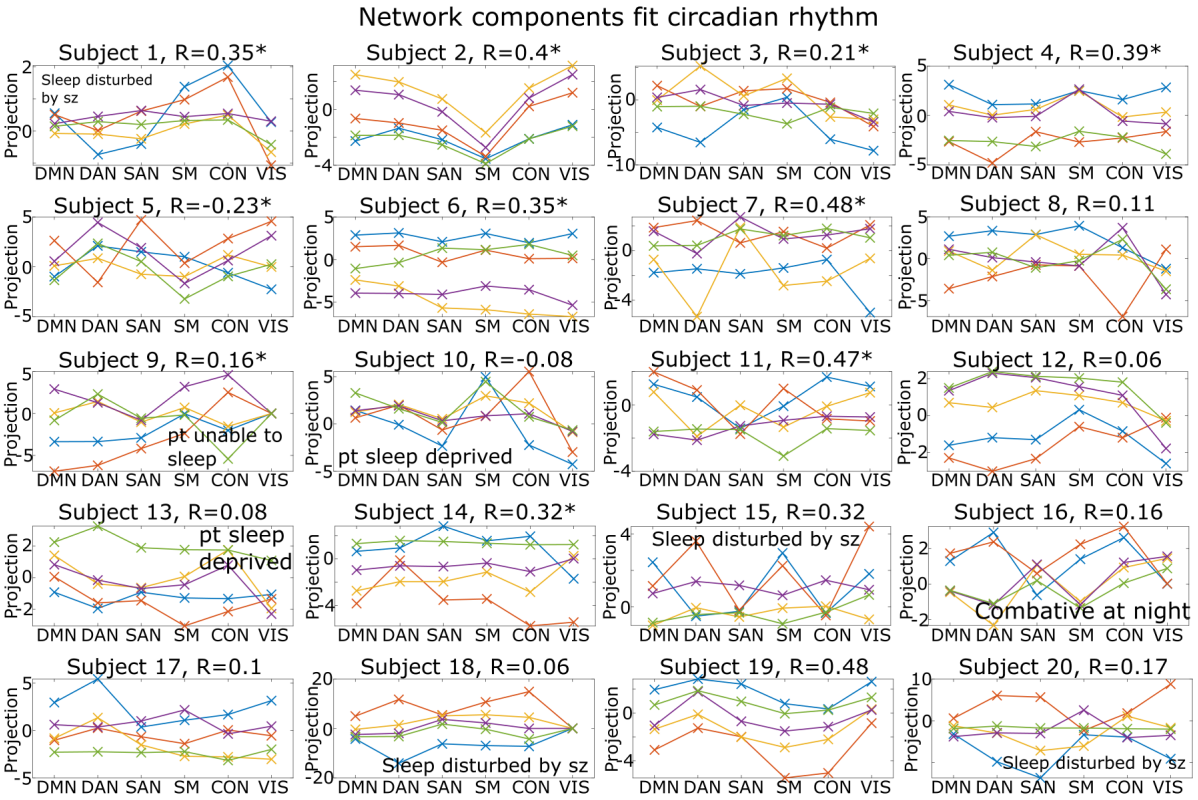

Figure S7: This figure shows the results in Figure 3.4A for all twenty subjects. The anatomical and frequency coverage of the mixture of network components associated with circadian rhythm in each subject are shown above. The correlation between the circadian sinusoid and the mixture's activation is shown above each plot (R). Asterisks indicate statistically significant correlations.

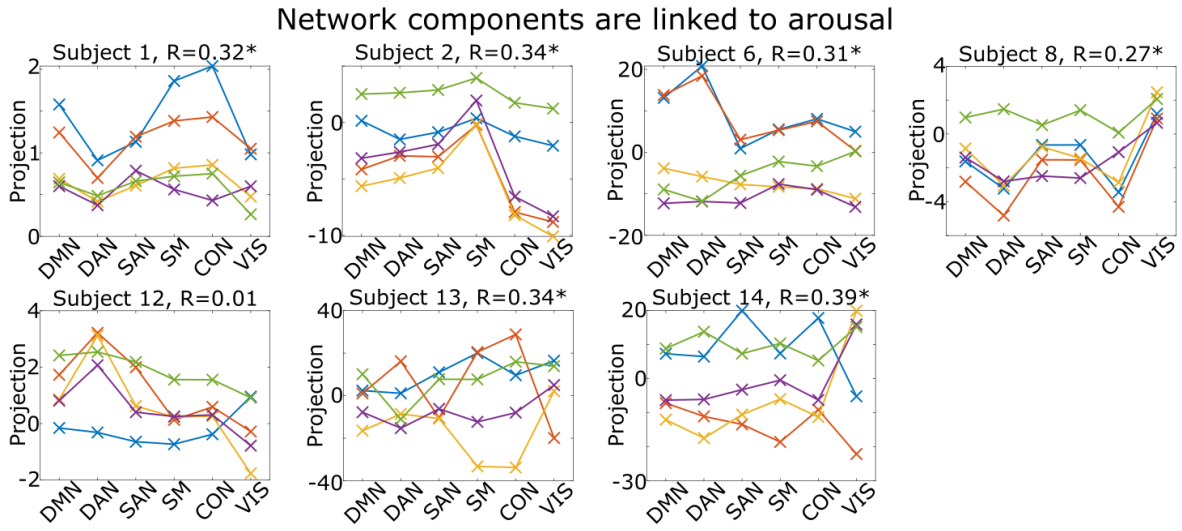

Figure S8: This figure shows the results in Figure 3.4B for all subjects with recorded EKG. The anatomical and frequency coverage of the mixture of networks associated with heart rate (which is used as a proxy for arousal) in each participant are shown above. The correlation between the heart rate and the mixture's activation is shown above each plot (R). Asterisks indicate statistically significant correlations.

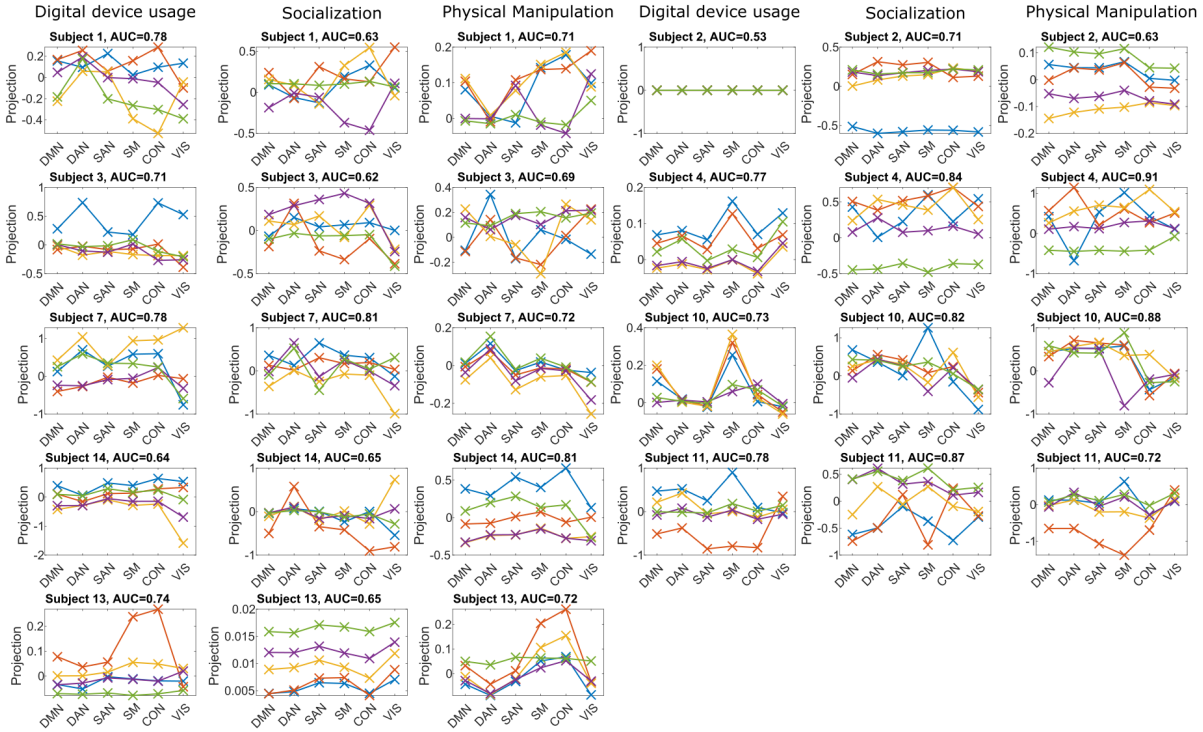

Figure S9: This figure shows the results in Figure 3.4C for all subjects with annotated video recordings. The anatomical and frequency coverage of the mixture of networks associated with each behavior in each participant are shown above. The area-under-curve classification accuracy of the mixture's activation is shown above each plot (AUC).

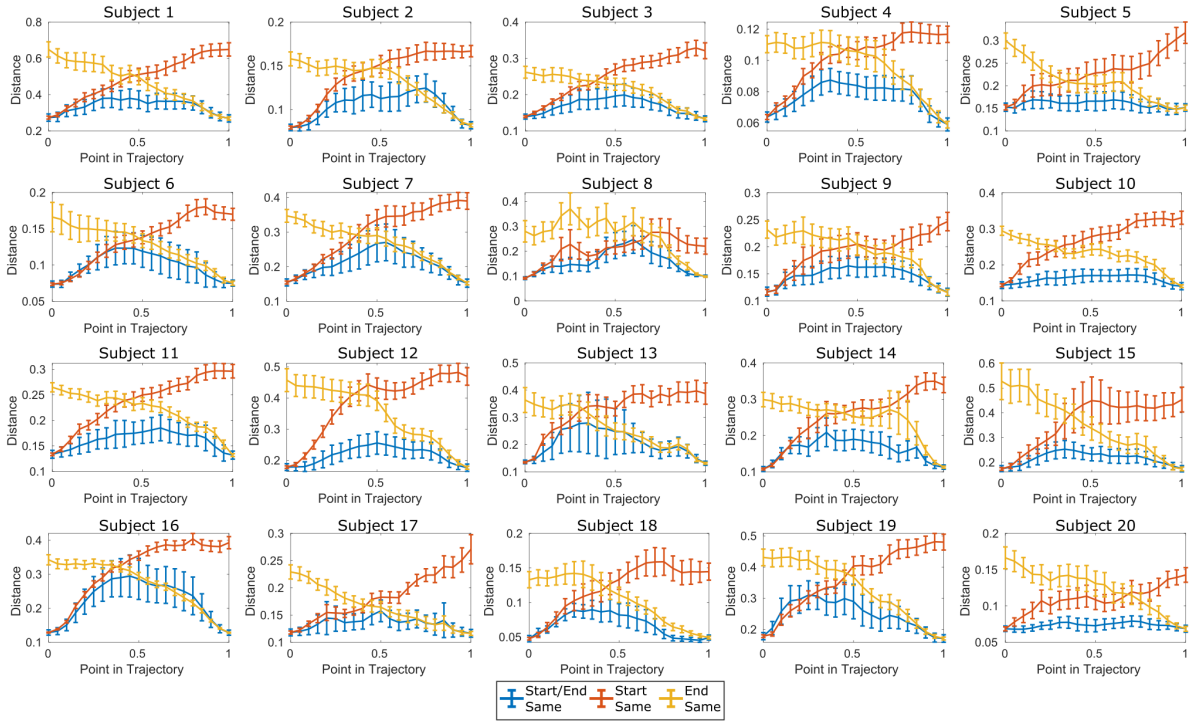

Figure S10: This figure shows the results in Figure 3.5D for all twenty subjects. Average distance between pairs of transition trajectories as a function of what proportion of the trajectory was complete. Trajectory pairs are grouped into three categories: transitions with similar starting and ending points (1, blue) vs similar starting but different ending points (2, red) vs different starting but similar ending points (3, yellow).

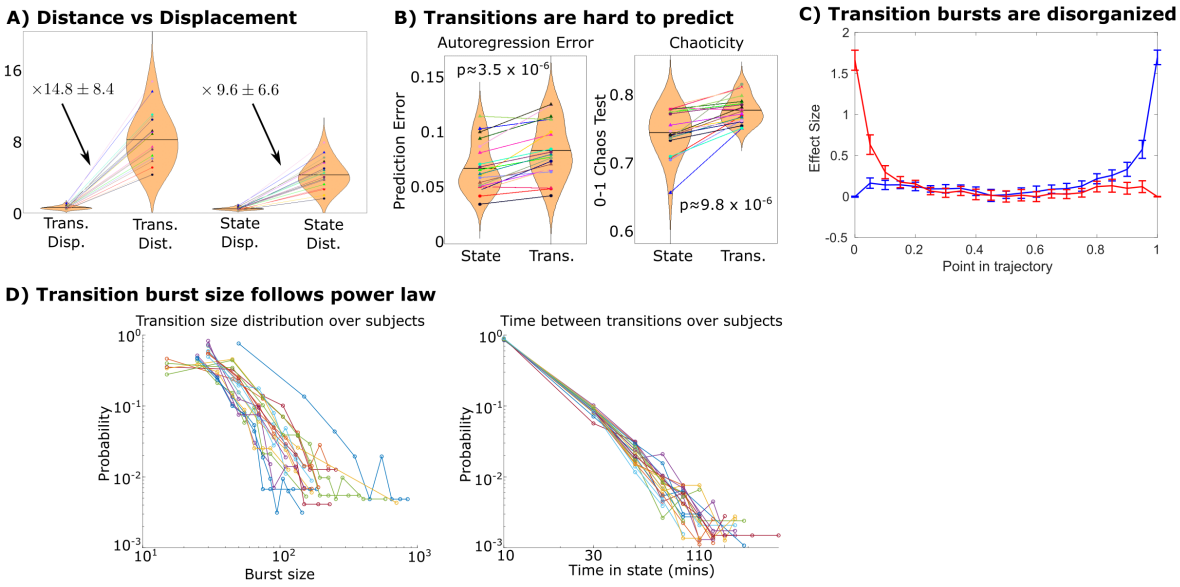

Figure S11: We replicated the results shown in Figure 3.5 using a different method of defining what constituted a trajectory (identifying continuous periods with high speed rather than change point detection) to further demonstrate the robustness of our results to analysis choices. D) We used log-likelihood tests to ask if transition size and frequency distributions were more likely to follow power-law vs exponential or log-normal distributions on a subject-by-subject basis. We used Wilcoxon signed ranked tests on the difference between power law likelihood vs alternative distributions across subjects to see if power law distributions had consistently higher likelihood. Transition size distributions were more likely to follow power law distributions over both exponential (p=0.02) and log-normal (p=0.001) by log-likelihood comparison tests across subjects. The time between transitions were also more likely to be best fit by power laws (p<1e-4 in both cases).

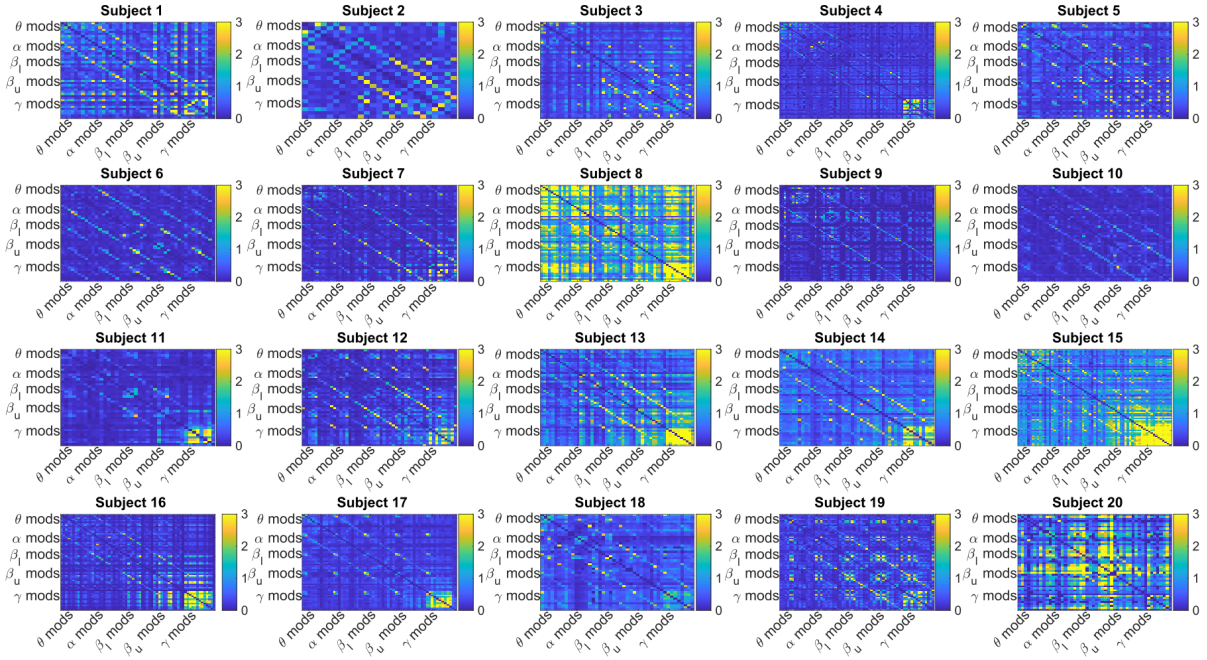

Figure S12: We replicated the results in Figure S6 using the original parcel coherences without any dimensionality reduction. The effect size of the distance between the joint distribution of each pair of parcels over the five frequency bands is compared to an independent null. Non-zero effect sizes represent statistically significant distances determined via permutation testing. Parcels are ordered from left to right/top to down by all the parcels for a single frequency band and then the same parcels in identical order for the next frequency band and so on.

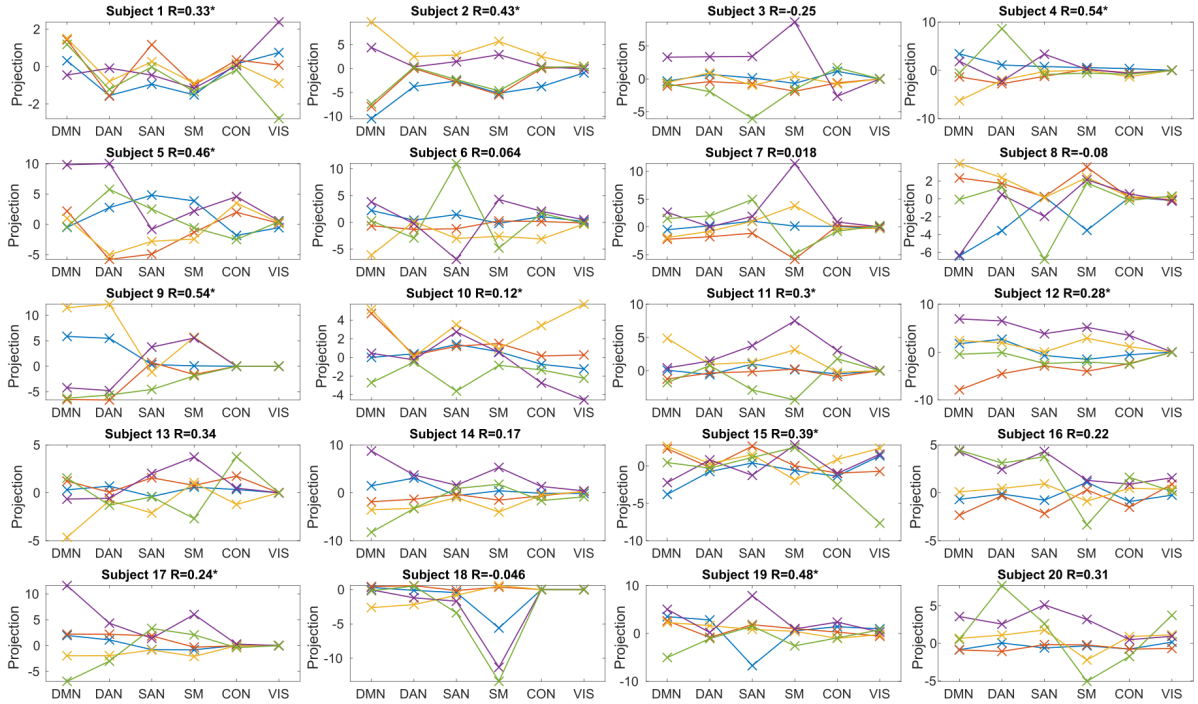

Figure S13: We replicated the findings in Figure S7 using the original parcel coherences without any dimensionality reduction. The anatomical and frequency coverage of the mixture of parcels associated with circadian rhythm in each participant are shown above. The correlation between the circadian sinusoid and the mixture's activation is shown above each plot (R). Asterisks indicate statistically significant correlations.

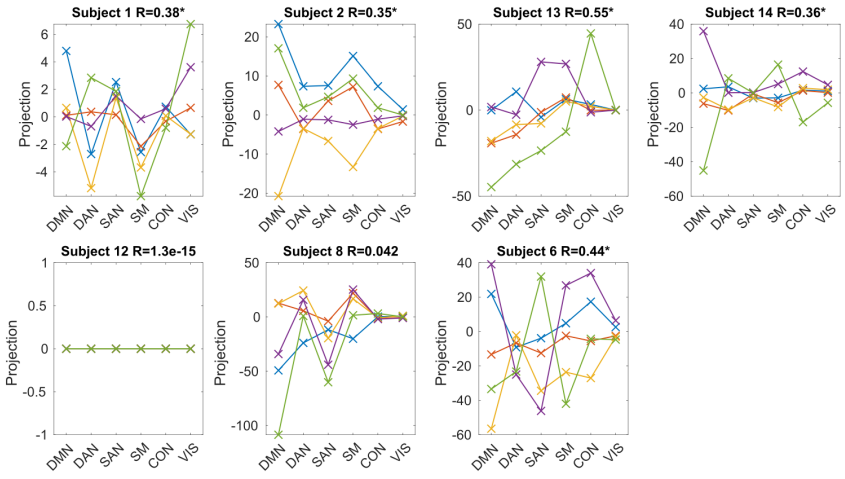

Figure S14: We replicated the findings in Figure S8 using the original parcel coherences without any dimensionality reduction. The anatomical and frequency coverage of the mixture of parcels associated with heart rate (which is used as a proxy for arousal) in each participant are shown above. The correlation between the heart rate and the mixture's activation is shown above each plot (R). Asterisks indicate statistically significant correlations.

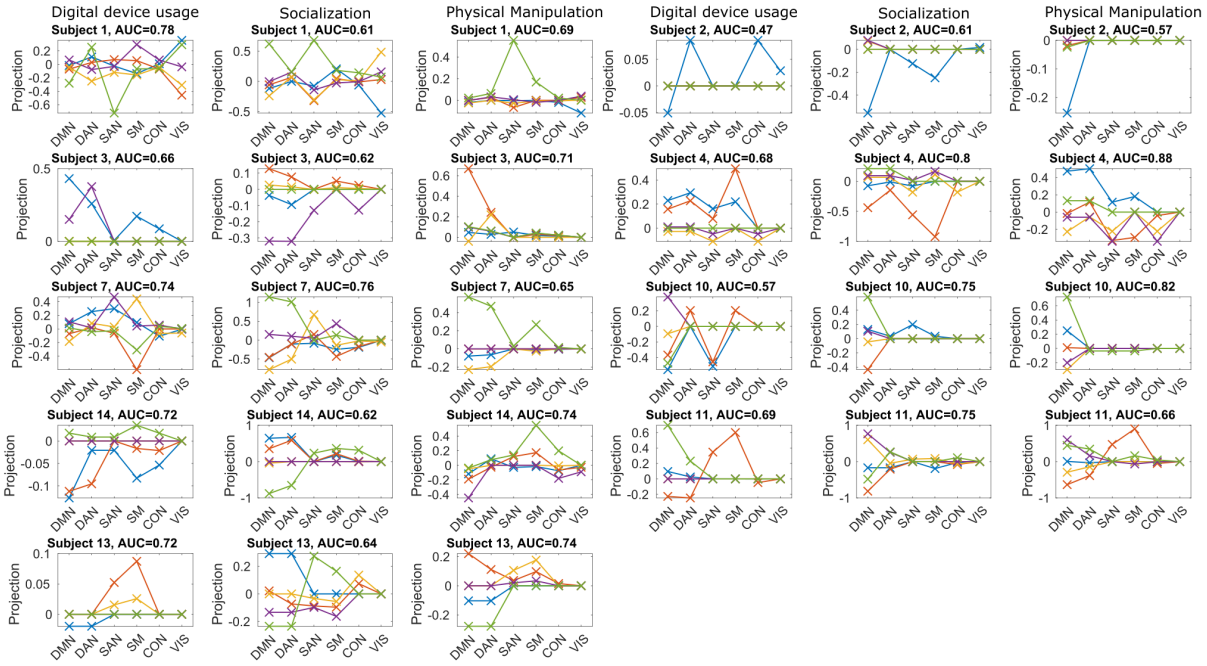

Figure S15: We replicated the findings in Figure S9 using the original parcel coherences without any dimensionality reduction. The anatomical and frequency coverage of the mixture of parcels associated with each behavior in each participant are shown above. The area-under-curve classification accuracy of the mixture's activation is shown above each plot (AUC).

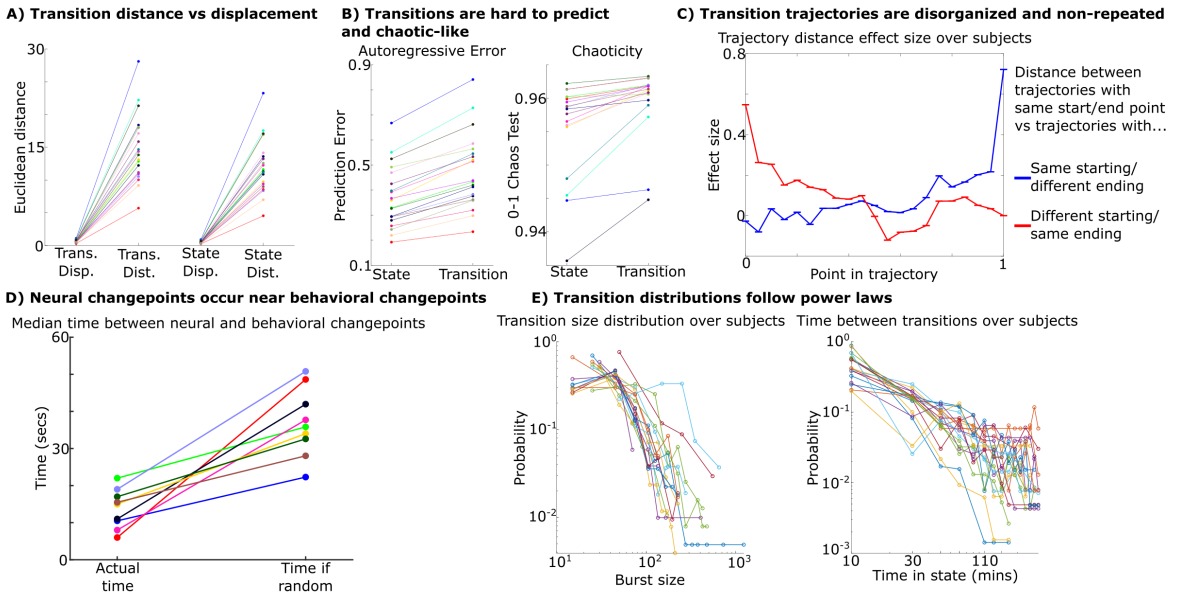

Figure S16: We replicated the findings in Figures 3.5 using the original parcel coherences without any dimensionality reduction. A(replication of Figure 3.5B) The total distance traveled during transitions and stable states on average for all twenty subjects, versus the net displacement (distance from start and end state). B(replication of Figure 3.5E) We used vector autoregression in a leave-one-day-out cross-validation test to predict the future evolution of brain states. We found that autoregression error significantly increases during transitions, indicating that transitions were difficult to predict. By using a 0-1 chaos test on non-overlapping ten-minute segments of data, we demonstrated that brain signals also demonstrated elements consistent with high chaoticity during transitions. C(replication of Figure 3.5D) The effect size of the difference in distance between trajectories that start/end in similar brain states vs trajectories that only start in similar states (blue) or those that only end in similar states (red). Error bars indicate confidence bounds over all 20 participants. D(replication of Figure 3.5B) The median time between behavioral changepoints and the nearest neural transition point is compared to the expected time if neural transitions occurred randomly. E(replication of Figure 3.5F) The size (net displacement) of each transition and the time between them is shown over all participants in log-log form. We used Wilcoxon signed ranked tests on the difference between power law likelihood vs alternative distributions across participants to see if power law distributions had consistently higher likelihood. Transition size distributions were more likely to follow power law distributions over both exponential (p=0.020) and log-normal (p=0.006) by log-likelihood comparison tests across subjects. The time between transitions were also more likely to be best fit by power laws (p=5e-4 for exponential and p=0.007 for log-normal).

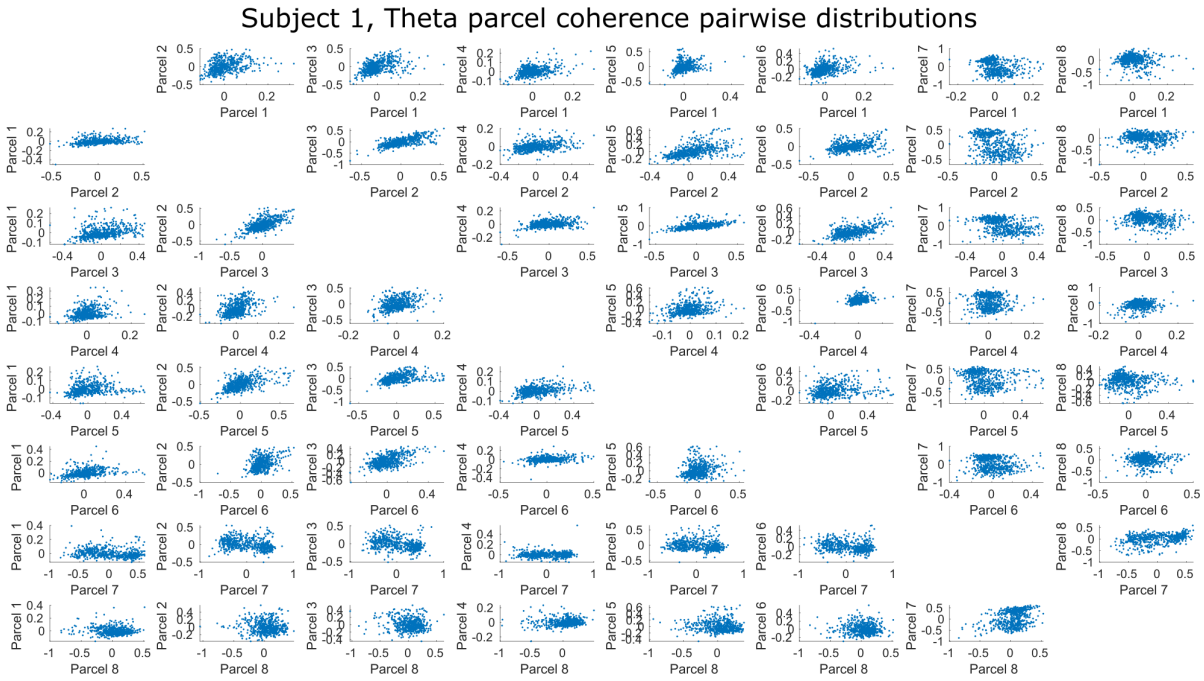

Figure S17: The pairwise distribution between the theta coherence of all non-seizure related parcels for a single subject.

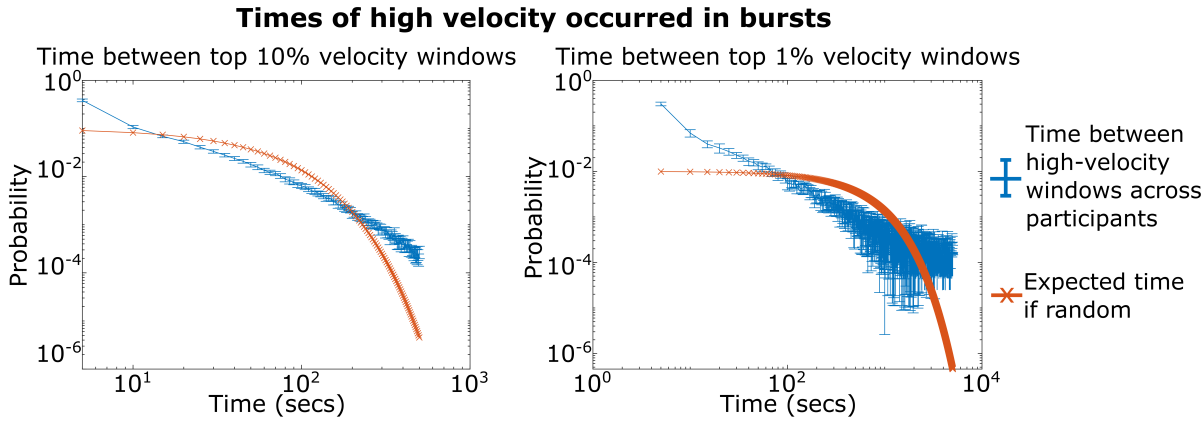

Figure S18: The average time between windows with the top 1% or 10% of velocity across all participants is shown in blue vs the expected time if windows of high velocity occurred via homogenous Poisson process ($\lambda = 0.01, 0.1$). Error bars show 95% confidence intervals across participants.

References

[1] Yumi Matsushita, Yasunori Takata, Ryoichi Kawamura, Misaki Takakado, Toshimi Hadate, and Haruhiko Osawa. The fluctuation in sympathetic nerve activity around wake-up time was positively associated with not only morning but also daily glycemic variability in subjects with type 2 diabetes. *Diabetes Research and Clinical Practice*, 152:1–8, 2019. Publisher: Elsevier.

[2] Christopher J Honey, Thomas Thesen, Tobias H Donner, Lauren J Silbert, Chad E Carlson, Orrin Devinsky, Werner K Doyle, Nava Rubin, David J Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434, 2012. Publisher: Elsevier.

[3] Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications*, 7(1):1–13, 2016. Publisher: Nature Publishing Group.

[4] Russell A Poldrack, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L Boyd, and others. Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6(1):1–15, 2015. Publisher: Nature Publishing Group.

[5] Evan M Gordon, Timothy O Laumann, Adrian W Gilmore, Dillan J Newbold, Deanna J Greene, Jeffrey J Berg, Mario Ortega, Catherine Hoyt-Drazen, Caterina Gratton, Haoxin Sun, and others. Precision functional mapping of individual human brains. *Neuron*, 95(4):791–807, 2017. Publisher: Elsevier.

[6] Laura H Schulte, Mareike M Menz, Jan Haaker, and Arne May. The migraineur's brain networks: Continuous resting state fMRI over 30 days. *Cephalalgia*, 40(14):1614–1621, 2020. Publisher: SAGE Publications Sage UK: London, England.

[7] Diego Vidaurre, Stephen M Smith, and Mark W Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, 2017. Publisher: National Acad Sciences.

[8] Danielle S Bassett, Nicholas F Wymbs, Mason A Porter, Peter J Mucha, Jean M Carlson, and Scott T Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011. Publisher: National Acad Sciences.

[9] Justin C Williams, Robert L Rennaker, and Daryl R Kipke. Long-term neural recording characteristics of wire microelectrode arrays implanted in cerebral cortex. *Brain Research Protocols*, 4(3):303–313, 1999. Publisher: Elsevier.

[10] Dillan J Newbold, Timothy O Laumann, Catherine R Hoyt, Jacqueline M Hampton, David F Montez, Ryan V Raut, Mario Ortega, Anish Mitra, Ashley N Nielsen, and Derek B Miller. Plasticity and spontaneous activity pulses in disused human brain circuits. *Neuron*, 107(3):580–589, 2020. Publisher: Elsevier.

[11] Katherine W Scangos, Ghassan S Makhoul, Leo P Sugrue, Edward F Chang, and Andrew D Krystal. State-dependent responses to intracranial brain stimulation in a patient with depression. *Nature medicine*, 27(2):229–231, 2021. Publisher: Nature Publishing Group.

[12] Sameer A Sheth, Kelly R Bijanki, Brian Metzger, Anusha Allawala, Victoria Pirtle, Joshua A Adkinson, John Myers, Raissa K Mathura, Denise Oswalt, and Evangelia Tsolaki. Deep brain stimulation for depression informed by intracranial recordings. *Biological Psychiatry*, 92(3):246–251, 2022. Publisher: Elsevier.

[13] John D Murray, Alberto Bernacchia, David J Freedman, Ranulfo Romo, Jonathan D Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, and others. A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience*, 17(12):1661–1663, 2014. Publisher: Nature Publishing Group.

[14] Uri Hasson, Janice Chen, and Christopher J Honey. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19(6):304–313, 2015. Publisher: Elsevier.

[15] Mehrshad Golesorkhi, Javier Gomez-Pilar, Federico Zilio, Nareg Berberian, Annemarie Wolff, Mustapha CE Yagoub, and Georg Northoff. The brain and its time: intrinsic neural timescales are key for input processing. *Communications biology*, 4(1):1–16, 2021. Publisher: Nature Publishing Group.

[16] Annemarie Wolff, Nareg Berberian, Mehrshad Golesorkhi, Javier Gomez-Pilar, Federico Zilio, and Georg Northoff. Intrinsic neural timescales: temporal integration and segregation. *Trends in cognitive sciences*, 2022. Publisher: Elsevier.

[17] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. Publisher: ACM New York, NY, USA.

[18] Mark A Kramer, Uri T Eden, Kyle Q Lepage, Eric D Kolaczyk, Matt T Bianchi, and Sydney S Cash. Emergence of persistent networks in long-term intracranial EEG recordings. *Journal of Neuroscience*, 31(44):15757–15767, 2011. Publisher: Soc Neuroscience.

[19] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. Publisher: MIT Press.

[20] Ali Azarbarzin, Michele Ostrowski, Patrick Hanly, and Magdy Younes. Relationship between arousal intensity and heart rate response to arousal. *Sleep*, 37(4):645–653, 2014. Publisher: Oxford University Press.

[21] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. Publisher: Taylor & Francis.

[22] Eun-Kyeong Kim and Hang-Hyun Jo. Measuring burstiness for finite event sequences. *Physical Review E*, 94(3):032311, 2016. Publisher: APS.

[23] Connie JG Gersick. Revolutionary change theories: A multilevel exploration of the punctuated equilibrium paradigm. *Academy of management review*, 16(1):10–36, 1991. Publisher: Academy of Management Briarcliff Manor, NY 10510.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[25] Georg A Gottwald and Ian Melbourne. The 0-1 test for chaos: A review. *Chaos detection and predictability*, pages 221–247, 2016. Publisher: Springer.

[26] Luca Cocchi, Leonardo L Gollo, Andrew Zalesky, and Michael Breakspear. Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, 158:132–152, 2017. Publisher: Elsevier.

[27] Sabrina A Jones, Jacob H Barfield, V Kindler Norman, and Woodrow L Shew. Scale-free behavioral dynamics directly linked with scale-free cortical dynamics. *Elife*, 12:e79950, 2023.

[28] Stephen Jay Gould and Niles Eldredge. Punctuated equilibria: an alternative to phyletic gradualism. *Models in paleobiology*, 1972:82–115, 1972. Publisher: San Francisco.

[29] Frank R Baumgartner and Bryan D Jones. *Agendas and instability in American politics*. University of Chicago Press, 1993.

[30] Ram Mudambi and Tim Swift. Proactive R&D management and firm growth: A punctuated equilibrium model. *Research Policy*, 40(3):429–440, 2011. Publisher: Elsevier.

[31] Per Bak and Stefan Boettcher. Self-organized criticality and punctuated equilibria. *Physica D: Nonlinear Phenomena*, 107(2-4):143–150, 1997. Publisher: Elsevier.

[32] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Physical review letters*, 59(4):381, 1987. Publisher: APS.

[33] Dimitrije Marković and Claudius Gros. Power laws and self-organized criticality in theory and nature. *Physics Reports*, 536(2):41–74, 2014. Publisher: Elsevier.

[34] Markus J Aschwanden and Manuel Güdel. Self-organized criticality in stellar flares. *The Astrophysical Journal*, 910(1):41, 2021. Publisher: IOP Publishing.

[35] Y Contoyiannis, SG Stavrinides, MP Hanias, M Kampitakis, P Papadopoulos, R Picos, SM Potirakis, and EK Kosmidis. Criticality in epidemic spread: An application in the case of COVID19 infected population. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(4):043109, 2021. Publisher: AIP Publishing LLC.

[36] Woodrow L Shew and Dietmar Plenz. The functional benefits of criticality in the cortex. *The neuroscientist*, 19(1):88–100, 2013. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

[37] Jordan O'Byrne and Karim Jerbi. How critical is brain criticality? *Trends in Neurosciences*, 2022. Publisher: Elsevier.

[38] Klaus Linkenkaer-Hansen, Vadim V Nikouline, J Matias Palva, and Risto J Ilmoniemi. Long-range temporal correlations and scaling behavior in human brain oscillations. *Journal of Neuroscience*, 21(4):1370–1377, 2001. Publisher: Soc Neuroscience.

[39] John M Beggs and Nicholas Timme. Being critical of criticality in the brain. *Frontiers in physiology*, 3:163, 2012. Publisher: Frontiers Research Foundation.

[40] Jongyup Lim, Jungho Lee, Eunseong Moon, Michael Barrow, Gabriele Atzeni, Joseph G Letner, Joseph T Costello, Samuel R Nason, Paras R Patel, and Yi Sun. A light-tolerant wireless neural recording IC for motor prediction with near-infrared-based power and data telemetry. *IEEE Journal of Solid-State Circuits*, 57(4):1061–1074, 2022. Publisher: IEEE.

[41] S Authier, Paul Haefner, S Fournier, E Troncy, and LB Moon. Combined cardiopulmonary assessments with implantable telemetry device in conscious freely moving cynomolgus monkeys. *Journal of pharmacological and toxicological methods*, 62(1):6–11, 2010. Publisher: Elsevier.

[42] Uros Topalovic, Zahra M Aghajan, Diane Villaroman, Sonja Hiller, Leonardo Christov-Moore, Tyler J Wishard, Matthias Stangl, Nicholas R Hasulak, Cory S Inman, and Tony A Fields. Wireless programmable recording and stimulation of deep brain activity in freely moving humans. *Neuron*, 108(2):322–334, 2020. Publisher: Elsevier.

[43] Paola Pinti, Ilias Tachtsidis, Antonia Hamilton, Joy Hirsch, Clarisse Aichelburg, Sam Gilbert, and Paul W Burgess. The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1464(1):5–29, 2020. Publisher: Wiley Online Library.

[44] François Tadel, Sylvain Baillet, John C Mosher, Dimitrios Pantazis, and Richard M Leahy. Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational intelligence and neuroscience*, 2011, 2011. Publisher: Hindawi.

[45] Anders M Dale, Bruce Fischl, and Martin I Sereno. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999. Publisher: Elsevier.

[46] Dora Hermes, Kai J Miller, Herke Jan Noordmans, Mariska J Vansteensel, and Nick F Ramsey. Automated electrocorticographic electrode localization on individually rendered brain surfaces. *Journal of neuroscience methods*, 185(2):293–298, 2010. Publisher: Elsevier.

[47] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019. Publisher: Nature Publishing Group.

[48] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal*

of the American Statistical association, 66(336):846–850, 1971. Publisher: Taylor & Francis.

[49] Daniel Pimentel-Alarcón and Robert Nowak. Random consensus robust PCA. In *Artificial Intelligence and Statistics*, pages 344–352. PMLR, 2017.

[50] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946. Publisher: JSTOR.

[51] Paul van Gent, Haneen Farah, Nicole Nes, and Bart van Arem. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In *Proceedings of the 6th HUMANIST Conference*, pages 173–178, 2018.

[52] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. Publisher: SIAM.

CHAPTER 4

Default mode network emerges as a homeostatic-like attractor over weeklong recordings of the human brain during natural behavior and wakeful rest

In the previous chapter, I asked the question "what are some of the largest and most basic laws that seem to govern how the brain slowly changes over time?". While this is a logical first-pass approach to understanding what overall patterns surround the brain's long term dynamics, continuous brain activity recorded over a week offers the opportunity to analyze the brain's dynamics over hundreds of thousands of trials of neural activity at a finer level. Here, I investigated the primary drivers of the brain's dynamics using Koopman operators.

4.1 Main Findings

Studying the brain as a dynamical system has a long history that arguably started from early electroencephalography studies on how our brain propagates excitatory and inhibitory activity in a way that allows it to remain responsive and ever-changing without letting its dynamics become overwhelmed with either [1, 2]. At the heart of it are models that attempt to capture and explain the temporal evolution of the brain's activity in a

similar manner to how we think about the movement of celestial bodies in the solar system [3]. Comets or planets are drawn by gravity to the central attractor of our solar system, the sun, and through this force can achieve complex, dynamical orbits and fluctuations.

This approach has had tremendous impact on our investigations of many human physiological systems. We understand how physiological markers such as our blood pressure, heart rate, and blood sugar oscillate throughout each day by observing how they tend to be drawn towards a central state. Just as how we can launch a rocket from our planet that will temporarily achieve great heights before plummeting back to the earth's surface, someone's blood sugar will spike after consuming a meal before slowly returning back to baseline over hours due to the force of the body's insulin production. Compared to what is typically studied in cognitive neuroscience, most of these processes evolve on very slow timescales. If we receive a sharp, stressful shock that causes our heart rate to double, it can take minutes for our body to calm itself back down.

What are the primary tendencies governing the brain's dynamics over similarly long timescales? Here I used the dataset described in the previous chapter to learn dynamical systems models to investigate this question. Dynamical systems theory analyzes physical and biological systems by their behavior surrounding "critical/equilibrium points", places where the first derivative of the system over time is zero [4]. A canonical example would be a ball being kicked up and down a valley of hills. When the ball is at the bottom of a hill, it can be kicked away from its current position, but given enough time, it will be pulled back to the bottom (attractors). When it is at the top of the hill and is kicked away, it will do the opposite (repellers). While finding these hills and valleys is very straightforward when studying linear systems using eigendecomposition, we have observed in the previous chapter that many networks in the brain appear to be interacting in a nonlinear fashion.

To solve this, I turned to an idea in ergodic theory known as Koopman operators [5]. Koopman operators find a non-linear embedding of neural activity onto a manifold

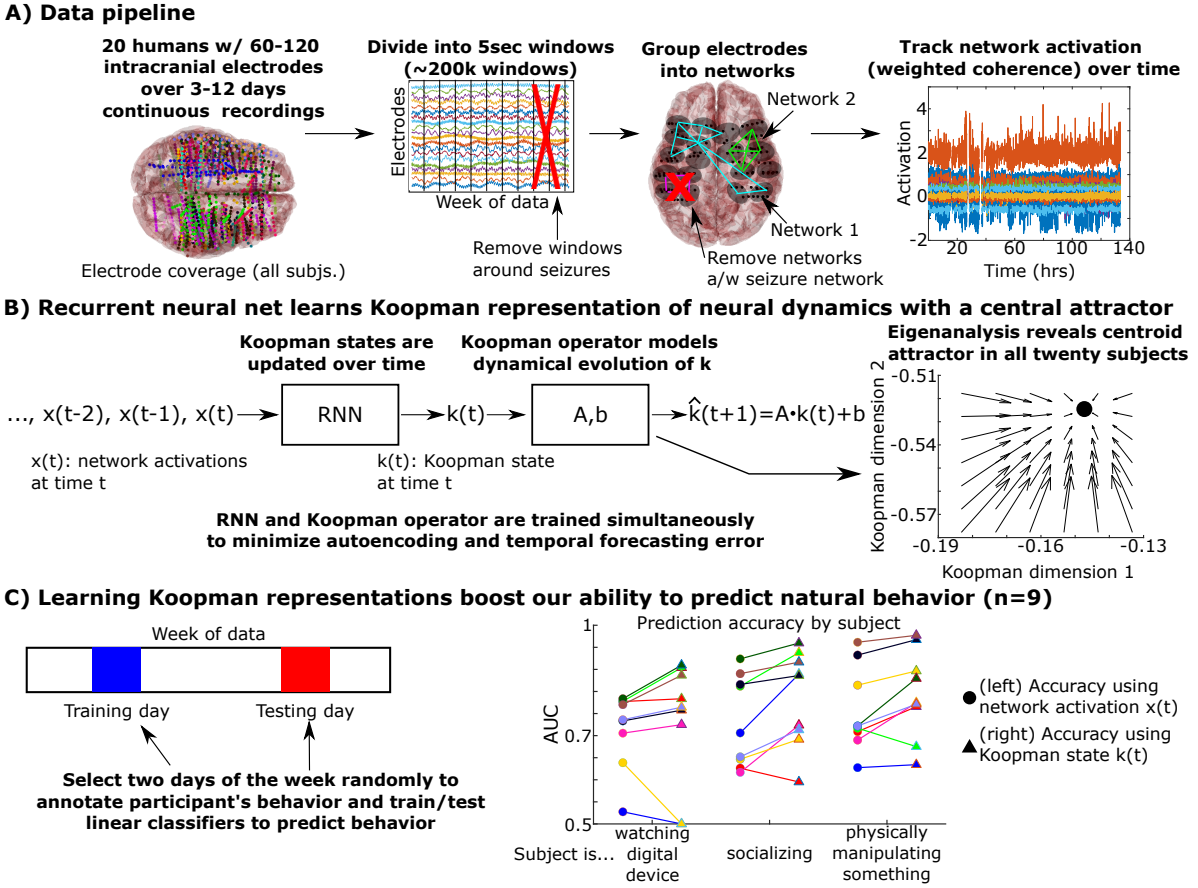

Figure 1: (A) I took between 3-12 days of mostly continuous neural recordings from twenty subjects and split it into five-second-long non-overlapping windows, removing windows around seizure activity. I grouped electrodes into data-driven networks where the activation of each network was defined as the weighted coherence of the contained electrodes according to a robust PCA protocol. Networks showing high similarity to the seizure onset zone and early propagation regions were removed. (B) I learned a Koopman representation of the brain's dynamical state by using a recurrent neural network to project the original network activations into a higher-dimensional "Koopman space" where the trajectories of the brain in this space could be captured by linear laws. (C) I found that trajectories in Koopman space more accurately predicted natural behavior than the original network activations by selecting two days of the week, annotating the subjects' behavior during those days, and training/testing linear classifiers to predict behavior. Paired t-test on the average difference in AUC between the two models showed statistical significance (p=0.006).

where its dynamics can be easily interpreted by conventional dynamical systems analyses (eigendecomposition) [6]. I combined this idea with Kalman filters [7] by training recurrent neural networks to find this non-linear embedding by using the ∼200k five-second windows of data collected for each participant. The resulting network, which is generated for each participant separately, takes a participant's time course of network activations and calculates an overall summary state (a numerical vector) for each five second window that describes the salient features of the brain up to that time (Figure 1B). This summary state, which we call "Koopman states", is trained to capture two properties. First, does the Koopman state accurately capture the most recent network activations. Second, does the Koopman state encode the necessary information to predict its own temporal evolution using linear models.

Our first question was whether these Koopman states accurately captured "useful" neurocognitive information. I chose nine subjects that had video recording monitoring at a sufficient quality to determine what the subject was doing throughout at least two days of the week. I randomly selected two days out of those with sufficient recording quality from each subject and labeled times where the subject was watching a digital screen, socializing with another human, or physically interacting with a held object. These three behavioral labels were not mutually exclusive, resulting in a three by one binary label vector for each time window. I trained L1-regularized linear classifiers on one day and then tested them on the second day. The input to these models were either the network activation features described in Chapter 3 or the learned Koopman states. For this analysis, the testing day was excluded when training the Koopman model. The resulting accuracies are shown in Figure 1C where learning these Koopman states increased the algorithm's capability to predict natural behavior (average AUC across all three behaviors increased by $\sim 14\%$, p=0.0062 by paired t-test), indicating that they contained neurocognitively useful information that could be accurately decoded using linear methods.

Our second question was what were the overall properties of the Koopman state

dynamics? Linear dynamical systems are primarily studied around what systems tend to do relative to their "critical/equilibrium points", places where the first derivative of the system over time is zero. Critical points are broadly either attractors (system tends to gravitate towards moving towards these points), repellers (points the system tends to move away from), or saddle points (combination of the two) with all three being identified in a variety of neuroscientific contexts [8–11]. Using eigendecomposition on the Koopman operator, I found that over all twenty subjects, their dynamics were captured by a single central attractor (only one equilibrium point and all eigenvalues were below one). This attractor is visualized in Figure 2A where I plot the Koopman state trajectories associated with behavior along with this central attractor in a subspace derived from the linear classifiers to maximize behavioral separation. Qualitatively, I describe these behavioral trajectories as "hourglasses" where different behaviors formed separate quadrants in the top of the hourglass, sleep formed the bottom of the hourglass, and periods where the participant is awake but not doing any of the three annotated behaviors formed the middle funnel of the hourglass with the attractor state.

I quantitatively verified aspects of this finding in Figure 2C by asking whether times with no active behavior tended to be closer to the central attractor state than times with active behavior. All quantitative analyses were done in the full Koopman representation space, not the subspace used for visualization in Figure 2A. For each subject, I calculated the average distance between each subject's Koopman state and the central attractor as a function of whether they were awake and doing one of the three behaviors, awake and not doing any of them, or asleep. Using paired t-tests I found that times of active behavior tended to depart further away from the attractor state compared to times where the subject was awake but not doing any of the three behaviors ($p=0.03$). The comparison between times of active behavior versus sleep was non-significant ($p=0.08$).

Finally, I asked whether there was a consistent neural state at this central attractor. In other words, when the brain was at the bottom of the valley, what was it doing? Using

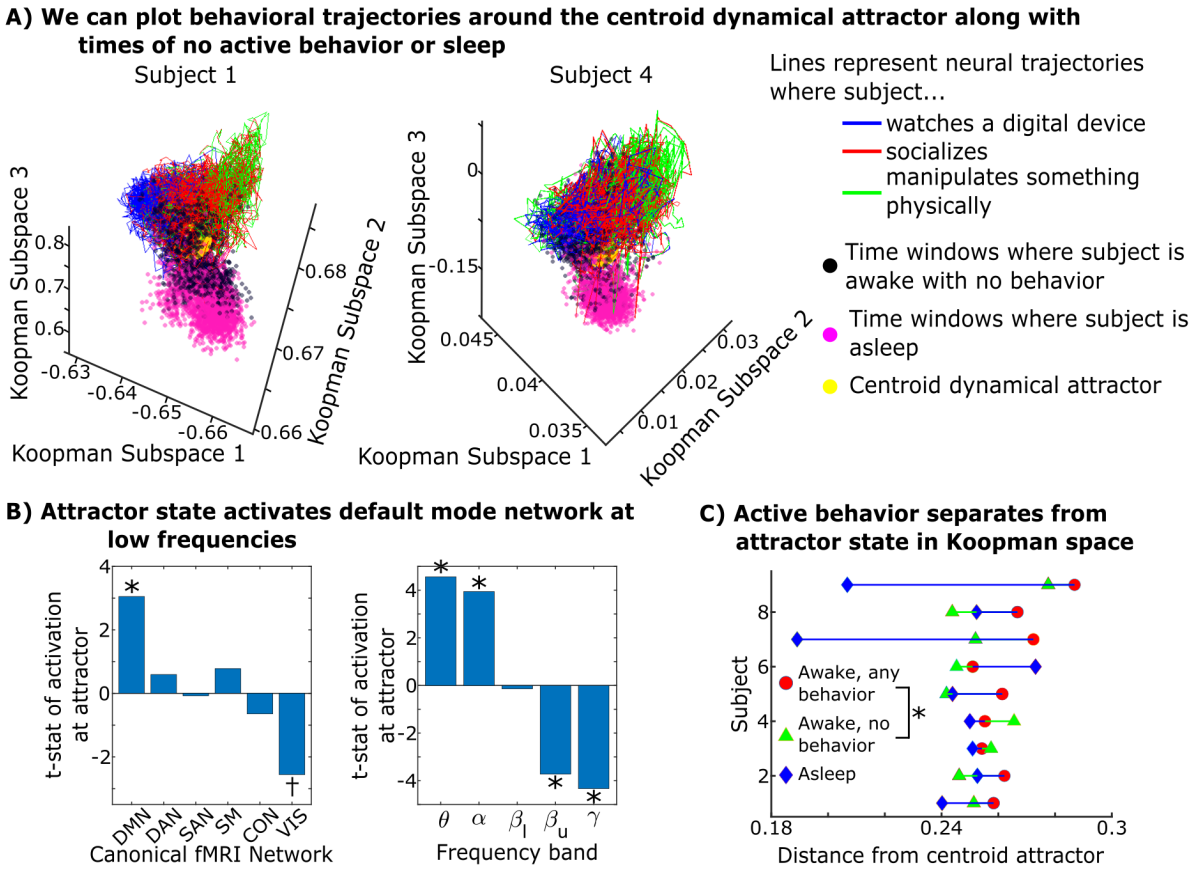

Figure 2: (A) Using eigendecomposition, I found that the linear operator modeling the brain's temporal evolution in Koopman space (Figure 1B) was dominated by a central attractor in all twenty subjects. I plotted trajectories associated with different behaviors in two subjects relative to the dynamical attractor state in a Koopman subspace derived from the linear classification coefficients in Figure 1C along with points in this subspace where the subject is either asleep or awake with no behavior. (B) I calculated which canonical fMRI networks and frequencies tended to be activated at the attractor state relative to the mean activation across each subject's functional connectome. The t-statistics of that activation/inactivation are shown with asterisks marking significant findings ($p<0.05$) post multiple comparisons correction. The dagger marks a t-statistic that is significant independently ($p=0.02$) but not significant post multiple comparisons correction. (C) I found that times where the participants were doing one of the three active behaviors tended to depart further away from the central dynamical attractor (DMN) state relative to times when the patient was awake but not doing any of the three behaviors ($p=0.03$).

a decoder model (which is trained as part of the Koopman model to assess autoencoding error), I asked which brain networks tended to be activated at this central Koopman attractor state. I then projected these networks onto the six canonical fMRI networks (default mode, salience, dorsal attention, control, somatomotor, and visual networks as defined in [12]) in Figure 2B. I found that the default mode network was consistently activated at low frequencies at the attractor state ($p<0.01$). I also found that the visual network trended towards being deactivated at the attractor state ($p=0.02$), but this did not pass multiple comparisons correction.

In summary, this central attractor state appears quite similar to what is seen during resting state functional MRI. According to long-standing fMRI studies, when our brains are awake but not performing an externally driven task, the brain activates a set of regions called the "default mode network" which has commonly been denoted the "resting state" of the human brain [13]. When our brains are instructed to perform a task, such as recognizing faces or words in an n-back experiment, this default mode network is deactivated while task-specific ones are activated. Since then, resting state brain activity has turned into the standard state assessed when studying how the brain's activity changes during various diseases ranging from dementia, mood disorders, traumatic brain injury, and more [14–16].

But whether this resting state is a uniquely important state of the brain has been oft debated. [17] argues that our brains do not simply "rest" when we are told to sit in a functional MRI scanner without specific instructions other than to not sleep or move, but rather engages in a mind wandering and internally focused state that is arguably more active than many explicitly driven tasks set by experimenters. Under this interpretation, resting state brain activity is still an interesting query of brain function, albeit one that should be renamed to "inward mindfulness", but not one that is uniquely more central or significant than any other state of brain activity [18].

Critically, since the Koopman states and their corresponding attractor point is calcu-

lated purely using unlabeled neural data with no information about behavior or rest, the default mode network emerging as a central attractor point is completely independent of any artificially derived experimental definition of resting state. All behavioral labels were merely used to investigate the significance of this attractor point after we found its location. I did not define this model around the concept of resting state: I used a model designed to determine what were the primary dynamical tendencies of the brain's evolution over long time periods and attraction to the default mode network during wakeful rest was the outcome.

While this finding suggests that the default mode network fulfills a centrally important role in the brain's dynamics, this finding is also compatible with the shift of the default mode network from its original definition as a uniquely important "resting state" to being a supporter of inwardly thoughtful cognitive actions that do not involve immediate perceptual information. At this central attractor state, we found a contrast of default mode network activation and a trend of visual network deactivation. [19] argues that the default mode network is anatomically isolated from the sensory periphery and centrally connected to the output of processing streams in the brain (such as the ventral visual stream) that are responsible for taking sensory input and converting them into abstract information. This allows it to focus on tasks such as mentally processing memories and experiences that had occurred in the recent past [20], understanding abstract concepts [21], and internally directed attention [22]; all tasks that have been linked to the default mode network but also ones that do not neatly fit into a conventional definition of "resting state brain".

In these analyses, while I found that the central attractor state was associated with "wakeful rest", I defined wakeful rest as anytime our participants were awake but not interfacing with a digital screen, physically manipulating some object, or socially interacting with somebody else. In practice, this meant that from the perspective of an external camera, they were not doing anything other than being introspective. Taking time to

reflect on and internally wander through our varied experiences is an intrinsic part of how humans navigate our chaotic, ever-changing environments. Studying the brain over very long time periods in a real-world environment is a unique way to investigate that experience.

Why would the brain possess a central tendency of moving towards an attractor state? In most physiological systems, there is a common word for this: homeostasis. The idea that having a resting blood pressure of 120/80 mmHg, a resting heart rate of $60 - 80$ bpm, and a fasting blood sugar of $70 - 100$ mmol/L is in some way optimal for physiological well-being and maintenance. That while it is possible for the body to temporarily leave these homeostatic ranges to achieve some task, once that task is over, mechanisms in the body take over to return systems back to homeostasis. Is an introspective state commonly associated with mind-wandering and mindfulness the cognitive equivalent of this? Is there an evolutionary reason why our brains still use massive amounts of metabolic energy to be inwardly introspective when we are not actively dealing with an immediate task in our environment instead of simply entering a "shut-off/stand-by" state [23]? Is the specific state of default mode network activation and visual network suppression important to fulfilling whatever that task is?

Highly-controlled, short neuroimaging experiments have demonstrated that resting state default mode activity is perturbed with various diseases such as Alzheimer's disease, traumatic brain injuries, and post-traumatic stress disorder [14–16]. Under this interpretation, these perturbations may be akin to elevated baseline blood pressure following kidney damage: a shift in homeostatic equilibrium from an optimal, desired state to new, faulty state that can cause long-term health detriment directly stemming from its existence.

One caveat of our approach is the use of a first-order dynamical systems model of the brain's Koopman state. No model is ever fully accurate. However, since our model was able to capture and predict natural behavior more accurately than using the raw features of brain activity (Figure 1C), I believe this model does capture neurocognitively useful

information. More broadly, this approach demonstrates a way to leverage unannotated neural data to better learn how to predict and understand labeled data. We faced the challenge that manually annotating video recordings for what behavior a participant was exhibiting is time-intensive, which is why we only annotated two of the on average seven days of data we had for each participant. This does not mean that we had to ignore the remaining five days of data. By learning statistical representations of the data over those days in a self-supervised fashion, we can boost the performance of our classifiers over labeled data. In general, due to the high cost of attaining curated labeled data, this approach has seen widespread success in AI applications ranging from computer vision [24], voice recognition [25], and natural language models such as GPT-3 [26, 27], one that I now extend to analyzing neural recordings in natural environments. I hope this demonstrates the intrinsic value of unlabeled neural data, even in the absence of any plans to directly query it.

The default mode network may not truly be a "resting state" of the brain, but these results indicate that it still fills a unique role in the brain's dynamics that may indicate a form of neural cognitive homeostasis.

4.2 Methods

For the analyses described in this chapter, I started with the same network activations over a week of recordings from twenty human subjects described and used in Chapter 3. I then used a recurrent neural network to learn a high-dimensional representation of these network activations where the brain's dynamics could be captured using linear laws. This approach falls under the class of Koopman operators, a set of approaches popularized in nonlinear control theory.

Broadly, a Koopman operator starts by taking a set of state variables observed from a dynamical system, $x(t) \in \mathbb{R}^{d_1 \times 1}$ a vector describing the overall state of a system at time

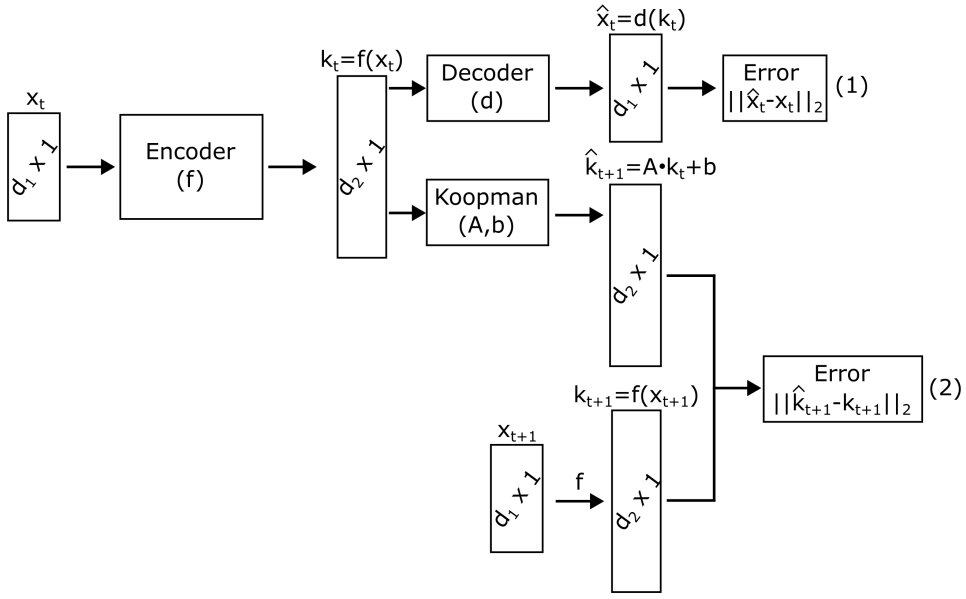

Figure 3: The Koopman model described in Figure 1B. x_t represents the network activations at time t. k_t is the output of the encoding model as is the Koopman state representation at time t. Here the encoding model is a recurrent neural network implemented via LSTM, the decoding model is a feedforward neural network, and the Koopman operator (A, b) is a first-order discrete differential equation.

t, and uses a non-linear transform, f to map this vector onto a point in a new space, $k(t) = f(x(t)) \in \mathbb{R}^{d_2 \times 1}$. Here the dimensionality of this new space (Koopman space), d_2, is chosen by the user. This mapping is chosen such that the dynamics of how $k(t)$ evolves into $k(t+1)$ can be modeled well using a linear operator that can be easily interpreted be conventional dynamical systems analyses[5].

In the past, the non-linear transform (f) was chosen according to an underlying intuition about the system, but it in recent years, it has been shown we can learn it instead [28, 29] Here we learn f using a recurrent neural network with the overall model shown in Figure 3.

To describe a single forward-pass through this model, I start with $x_t \in [-c_1, c_2]^{d \times 1}$, the network activation vector at time t (the intraparcel coherences projected onto the robust principal components found in Chapter 3) for a single subject. This vector is bounded between two constant values since coherence is bounded from $[0, 1]$ and the

network activation is simply a linear projection of these coherences through PCA. d is the number of networks found in that subject as described in Chapter 3.

This vector is first fed into the encoder, f, which is a long short-term memory (LSTM) unit [30, 31]. This unit maintains/updates an internal cell state $c_t \in \mathbb{R}^{l\times 1}$ which it uses to generate an output $k_t \in (-1,1)^{l\times 1}$ (Koopman state) where l is the dimensionality of the Koopman state which we choose later on. The LSTM updates are formalized in Equations 4.1-4.6 where σ_g is the sigmoid function, σ_c is the hyperbolic tangent function. The W's, U's, and b's are learnable parameters. $\odot$ is the Hadamard product. $f, i, o \in (0,1)^{l\times 1}$ are forget, input, and output gates respectively.

$$f_t = \sigma_g(W_f x_t + U_f k_{t-1} + b_f) \tag{4.1}$$

$$i_t = \sigma_g(W_i x_t + U_i k_{t-1} + b_i) \tag{4.2}$$

$$o_t = \sigma_g(W_o x_t + U_o k_{t-1} + b_o) \tag{4.3}$$

$$\tilde{c}_t = \sigma_c(W_c x_t + U_c k_{t-1} + b_c) \tag{4.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4.5}$$

$$k_t = o_t \odot \sigma_h(c_t) \tag{4.6}$$

In practice, this was implemented as an LSTM in tensorflow using default settings (tf.keras.layers.LSTM layer, kernel initializer glorot uniform, bias initializer zero, tanh activation, sigmoid recurrent activation)[31].

I tested the output of the LSTM (Koopman state k_t) to have dimensionalities ranging from $l = 0.5d, d, 5d, 10d, 20d$. For the main results section, I presented everything with an $l = 10d$. This was the hyperparameter I first tried with the model due to wanting a higher dimensionality than d to function as a nonlinear kernel but not being prohibitively high enough to prevent a future behavioral classification task. Other dimensionalities did

not significantly change the results in general as shown in Figures 4 and 5.

k_t was passed to two models during training. The first was a linear autoregression model with learnable parameters A, b that attempted to predict the next time step's Koopman state as $\hat{k}_{t+1} = Ak_t + b$. The autoregressive error of this is $||\hat{k}_{t+1} - k_{t+1}||_2^2$. In other words, how well does the information encoded in k_t predict its own temporal evolution using linear methods?

I also pass k_t to a decoding model (d), a feedforward neural network that takes k_t and attempts to predict x_t. This ensures k_t still retains information about the most recent network activations. This loss function is defined as $||d(x_t) - x_t||_2^2$. The decoding model was three layers deep with each layer containing as many units as the Koopman state dimensionality. The weights and bias term of each layer were all learnable. This was implemented as a feedforward neural network in tensorflow using ReLU activation functions and default settings (tf.keras.layers.Dense layer, kernel initializer glorot uniform, bias initializer zero)[31].

I trained all models (f, d, A, b) simultaneously according to these two loss functions. A single training step is described in Algorithm 1.

Algorithm 1 Training step

Feed the first half hour of $x_1, x_2, \ldots, x_{30\text{min}}$ into the LSTM encoder f to initialize its internal state
Define $t = 30\text{min} + 1$ where x_t is the time window right after the initializing half hour
Define $k_{t-1} = f(x_{t-1})$, the output of the LSTM on the previous network activation state
Define the current error to be $\epsilon = 0$
while $t <$total number of time windows **do**
    $k_t \leftarrow f(x_t)$
    $\hat{x}_t \leftarrow d(k_t)$
    $\epsilon \leftarrow \epsilon + ||\hat{x}_t - x_t||_2^2$
    $\hat{k}_t \leftarrow A \cdot k_{t-1} + b$
    $\epsilon \leftarrow \epsilon + ||\hat{k}_t - k_t||_2^2$
    $t \leftarrow t + 1$
end while
$\epsilon \leftarrow \epsilon/t +$ L-1 regularization term of all learnable parameters
Update f, d, A, b according to the gradient of ϵ w.r.t learnable parameters

In practice this was done using Python's tensorflow Adam optimizer under default settings (tf.keras.optimizers.Adam, learning rate 1e-3) [32]. All networks used ReLU activation functions and L-1 regularization.

In summary, this meant to make k_t into a state that captures information that allows me to predict the future state using linear methods while still retaining enough information to predict the most recent network activation vector.

4.2.1 Behavioral Classification (Figure 1C)

To ensure that the Koopman state representation captured neurocognitively interesting information, I used them to predict the subject's behavior. I manually annotated participant behavior on two separate days of the week for three behaviors: watching a digital screen, socializing with someone else, or physically manipulating an object. These three behaviors were not mutually exclusive. I denoted one day a training day, the other the testing day. When training the Koopman model (Figure 3), I exclude the testing day but include the rest of the week including days that I did not annotate videos for.

Behavioral classification was done using L1-regularized logistic classifiers using Python's sklearn toolbox. Hyper parameterization was optimized on the training set using ten-fold cross-validation. The area-under-curve of the receiver-operator-curve of each network's ability to classify the desired behavior was calculated.

Figure 4 shows the behavioral classification accuracy when the Koopman state dimensionality was varied. I found that classification accuracy rose compared to the original network activations in all cases except when the Koopman state was half the dimensionality of the original features in which case they performed roughly equally.

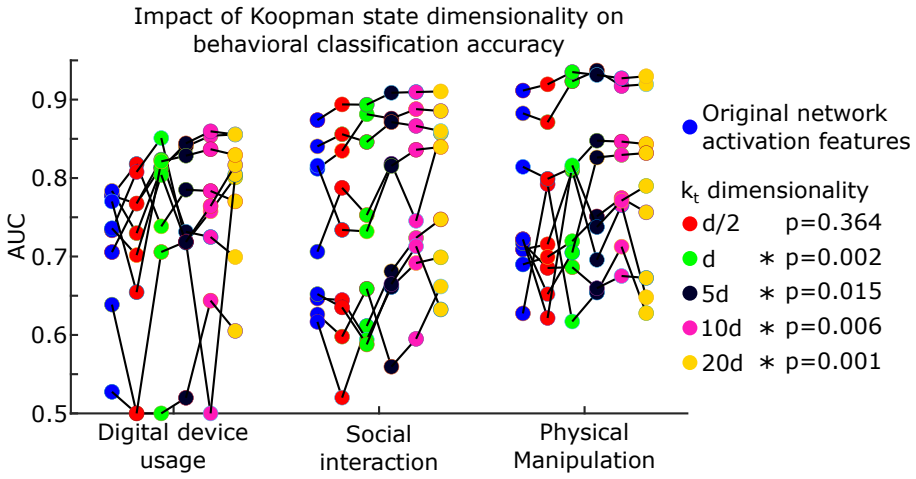

Figure 4: The classification accuracy of behavior as the dimensionality of the Koopman state was changed. Asterisks/p-values indicate whether the average classification accuracy across behaviors was higher when using the Koopman states as the features rather than the original network activations by paired t-test.

4.2.2 Behavioral trajectory visualization (Figure 2A)

For visualization purposes, I chose two subjects where we could accurately predict all three behaviors and plotted their behavioral trajectories in Koopman space. First, I calculated a subspace where all three behaviors are separated based on the linear classifiers determined in the previous section. The first axis (Koopman subspace 1) was the coefficient vector found to discriminate digital screen usage by the behavioral logistic classifier. The second axis was the part of the socialization-associated feature vector that was orthogonal to the first axis. The third axis was the part of the physical manipulation-associated feature vector that was orthogonal to both the first and second axes. This is formalized in Equations 4.7 to 4.7 where $w_1, w_2, w_3 \in \mathbb{R}^{10d \times 1}$ are the weights found by logistic classification to predict whether a subject is doing a behavior or not from their Koopman state.

$$s_1 = w_1 \tag{4.7}$$

$$s_2 = w_2 - (w_2^T \cdot s_1)s_1 \tag{4.8}$$

$$s_3 = w_3 - (w_3^T \cdot s_1)s_1 - (w_3^T \cdot s_2)s_2 \tag{4.9}$$

I then projected the Koopman state representation from both annotated days onto these three axes and plotted trajectories when the subjects were partaking in each behavior along with times where the subject was not doing any of the behaviors. No quantitative analyses were done on these projections, they were done purely for visualization.

4.2.3 Attractor state analysis (Figure 2B)

I analyzed the overall dynamical systems properties of the Koopman operator (A, b from Figure 3). The equilibrium point of a 1-st order discrete dynamical system is shown in Equations 4.10-4.12[33].

$$k_{t+1} = Ak_t + b \tag{4.10}$$

$$k_{eq} = Ak_{eq} + b \tag{4.11}$$

$$k_{eq} = \text{inv}(I - A) \cdot b \tag{4.12}$$

If $(I - A)$ is full rank, there is only a single equilibrium point. In all twenty subjects, that matrix was full rank and non-ill conditioned.

To determine whether this point was an attractor or a repulsor, I calculated the eigenvalues of A. If the magnitude of all eigenvalues are less than one, then the equilibrium point is an attractor (which it was for all subjects). This can be proven in Equations

4.13-4.18 where A has eigenvectors/values $Av_j = \lambda_j v_j$ and we define a perturbation away from k_{eq}, $k_t = k_{eq} + \epsilon$. The next time step, k_{t+1} is modeled as,

$$k_{t+1} = Ak_t + b \tag{4.13}$$

$$= A(k_{eq} + \epsilon) + b \tag{4.14}$$

$$= k_{eq} + A\epsilon \tag{4.15}$$

$$= k_{eq} + \sum_j \left(v_j^T \lambda_j \epsilon\right) v_j \tag{4.16}$$

$$||k_{t+1} - k_{eq}||_2^2 = \left|\left|\sum_j \left(v_j^T \lambda_j \epsilon\right) v_j\right|\right|_2^2 \tag{4.17}$$

$$< ||\epsilon||_2^2 : \epsilon \neq 0 \tag{4.18}$$

While dynamical systems may have multiple critical points, this single attractor by the Koopman operator is conventionally thought to represent the "global" behavior of the system (e.g. a system with both an attractor and repeller will show eigenfunctions associated with the attractor if the system gravitates towards the attractor as time approaches infinity [6]) with the caveat that interpreting systems with more than one critical point using Koopman operators is an active area of investigation [34].

To see what brain networks tended to be activated when the brain is at this "attractor state", I calculated $x_{eq} = d(k_{eq})$: the output of the decoding model (which predicts x_t from k_t) when given the equilibrium Koopman state representation. One important consideration with this approach is that the encoder (f) takes in a time series of $x_1, x_2, \ldots, x_t$ to generate each k_t. I only ask the decoder to return the last x_t. From a neuroscience perspective, this is asking "when the brain reaches the attractor state, what are its networks doing right then and there?" which sufficed for the basic question I wanted to ask. It is not a complete picture of the overall dynamics where "reaching" the attractor

state would encompass an entire time series.

I projected the output of the decoding model, x_{eq} onto the six canonical fMRI networks as defined in [12] by taking the dot product between $x_{eq} \in \mathbb{R}^{d\times 1}$ (activation of each of the participant's networks at central attractor) and a transformation matrix $C \in \mathbb{R}^{d\times 6}$. C details the proportional overlap between each of the participant's d networks and the six canonical fMRI networks (e.g. did one network fall halfway in the default mode and halfway in visual network). This resulted in a six-by-one vector for each subject, the "fMRI attractor state" which detailed which fMRI networks were activated at this attractor. I subtracted each subject's fMRI attractor state vector by its mean to ask "which networks were activated or deactivated" relative to the rest of the brain. I used t-tests on each network activation across subjects to see if any networks were consistently activated or deactivated with Benjamini-Hochberg for multiple comparisons correction. Two subjects that did not have electrodes in all six canonical networks were removed from this analysis.

I repeated this process except by averaging over networks to ask if any frequencies were activated or inactivated.

This analysis for other Koopman state dimensionalities is shown in Figure 5 where we continue to see a general trend of default mode network activation and visual network deactivation.

4.2.4 Active behavior departs from attractor state (Figure 2C)

For the nine subjects with video annotations, I calculated the distance between each window's Koopman state representation and that subject's attractor state. I then averaged these distances across windows based on whether the subject was doing one of the three behaviors, awake but not doing any of the marked behaviors (which practically meant sitting idly), and asleep. This resulted in three metrics per subject. I used paired t-tests to ask whether the average active behavior to attractor distance was larger than the awake non-active to attractor distance or the asleep to attractor distances across subjects.

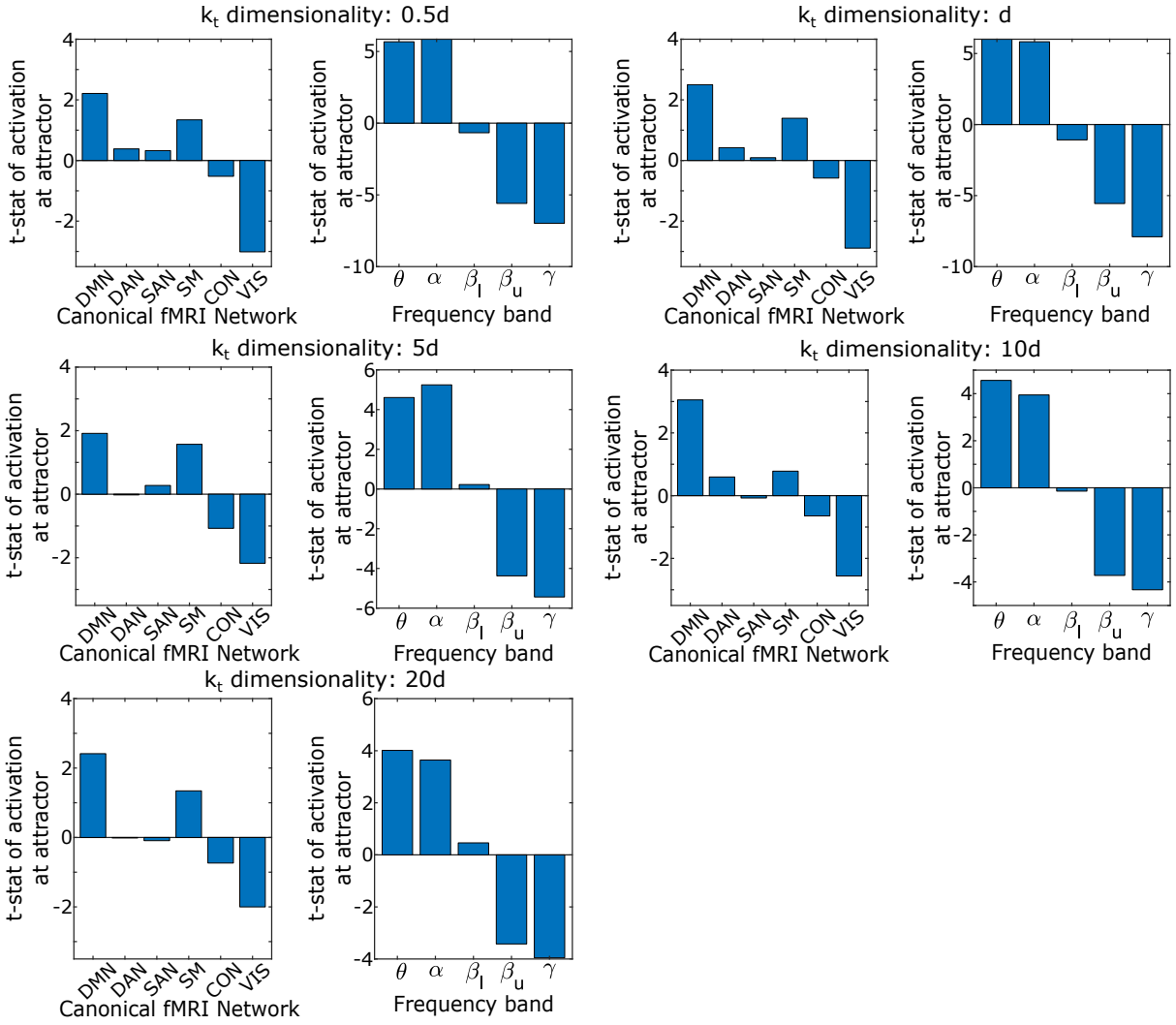

Figure 5: Which canonical fMRI networks and frequencies tended to be activated at the attractor state for different Koopman state dimensionalities.

References

[1] Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

[2] Walter J Freeman et al. *Mass action in the nervous system: Examination of the neurophysiological basis of adaptive behavior through the EEG*, volume 2004. Citeseer, 1975.

[3] Michael Breakspear. Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3):340–352, 2017.

[4] Stephen Smale. On gradient dynamical systems. *Annals of Mathematics*, pages 199–206, 1961.

[5] Bernard O Koopman. Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. Publisher: National Acad Sciences.

[6] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. Publisher: Springer.

[7] Gintaras V Puskorius and Lee A Feldkamp. Neurocontrol of nonlinear dynamical systems with kalman filter trained recurrent networks. *IEEE Transactions on neural networks*, 5(2):279–297, 1994.

[8] Misha Rabinovich, Ramon Huerta, and Gilles Laurent. Transient dynamics for neural processing. *Science*, 321(5885):48–50, 2008. Publisher: American Association for the Advancement of Science.

[9] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 26(4):1314–1328, 2006. Publisher: Soc Neuroscience.

[10] Shi Gu, Fabio Pasqualetti, Matthew Cieslak, Qawi K Telesford, Alfred B Yu, Ari E Kahn, John D Medaglia, Jean M Vettel, Michael B Miller, and Scott T Grafton. Controllability of structural brain networks. *Nature communications*, 6(1):1–10, 2015. Publisher: Nature Publishing Group.

[11] Luca Cocchi, Leonardo L Gollo, Andrew Zalesky, and Michael Breakspear. Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, 158:132–152, 2017. Publisher: Elsevier.

[12] Lucina Q Uddin, BT Yeo, and R Nathan Spreng. Towards a universal taxonomy of macro-scale functional human brain networks. *Brain topography*, 32(6):926–942, 2019. Publisher: Springer.

[13] Marcus E Raichle, Ann Mary MacLeod, Abraham Z Snyder, William J Powers, Debra A Gusnard, and Gordon L Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001. Publisher: National Acad Sciences.

[14] Eva M Palacios, Esther L Yuh, Yi-Shin Chang, John K Yue, David M Schnyer, David O Okonkwo, Alex B Valadka, Wayne A Gordon, Andrew IR Maas, Mary Vassar, et al. Resting-state functional connectivity alterations associated with six-month outcomes in mild traumatic brain injury. *Journal of neurotrauma*, 34(8):1546–1557, 2017.

[15] Andrew A Nicholson, Maria Densmore, Paul A Frewen, Richard WJ Neufeld, Jean Théberge, Rakesh Jetly, Ruth A Lanius, and Tomas Ros. Homeostatic normalization of alpha brain rhythms within the default-mode network and reduced symptoms in ptsd following a randomized controlled trial of eeg neurofeedback. *Brain Communications*, page fcad068, 2023.

[16] Walter Koch, Stephan Teipel, Sophia Mueller, Jens Benninghoff, Maxmilian Wagner, Arun LW Bokde, Harald Hampel, Ute Coates, Maximilian Reiser, and Thomas

Meindl. Diagnostic power of default mode network resting state fmri in the detection of alzheimer's disease. *Neurobiology of aging*, 33(3):466–478, 2012.

[17] Mary Helen Immordino-Yang, Joanna A Christodoulou, and Vanessa Singh. Rest is not idleness: Implications of the brain's default mode for human development and education. *Perspectives on Psychological Science*, 7(4):352–364, 2012.

[18] Alexa M Morcom and Paul C Fletcher. Does the brain have a baseline? why we should be resisting a rest. *Neuroimage*, 37(4):1073–1082, 2007.

[19] Jonathan Smallwood, Boris C Bernhardt, Robert Leech, Danilo Bzdok, Elizabeth Jefferies, and Daniel S Margulies. The default mode network in cognition: a topographical perspective. *Nature reviews neuroscience*, 22(8):503–513, 2021. Publisher: Nature Publishing Group UK London.

[20] Mahiko Konishi, Donald George McLaren, Haakon Engen, and Jonathan Smallwood. Shaped by the past: the default mode network supports cognition that is independent of immediate perceptual input. *PloS one*, 10(6):e0132209, 2015. Publisher: Public Library of Science San Francisco, CA USA.

[21] Elvis Dohmatob, Guillaume Dumas, and Danilo Bzdok. Dark control: The default mode network as a reinforcement learning agent. *Human brain mapping*, 41(12):3318–3341, 2020. Publisher: Wiley Online Library.

[22] Julia WY Kam, Jack J Lin, Anne-Kristin Solbakk, Tor Endestad, Pål G Larsson, and Robert T Knight. Default network and frontoparietal control network theta connectivity supports internal attention. *Nature human behaviour*, 3(12):1263–1270, 2019. Publisher: Nature Publishing Group UK London.

[23] Marcus E Raichle. The brain's dark energy. *Science*, 314(5803):1249–1250, 2006.

[24] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. pages 2959–2968, 2019.

[25] Hannah J Scheibner, Carsten Bogler, Tobias Gleich, John-Dylan Haynes, and Felix Bermpohl. Internal and external attention and the default mode network. *Neuroimage*, 148:381–389, 2017. Publisher: Elsevier.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[27] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[28] Enoch Yeung, Soumya Kundu, and Nathan Hodas. Learning deep neural network representations for Koopman operators of nonlinear dynamical systems. pages 4832–4839. IEEE, 2019.

[29] Yiqiang Han, Wenjian Hao, and Umesh Vaidya. Deep learning of koopman representation for control. pages 1890–1895. IEEE, 2020.

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

[31] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] J Klamka. Controllability of linear dynamical systems. *Contrib. Theory Differ. Equ*, 1:189–213, 1963.

[34] Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. Data-driven discovery of koopman eigenfunctions for control. *Machine Learning: Science and Technology*, 2(3):035023, 2021.

CHAPTER 5

Discussion

Since the discovery that we can record electrical activity from the human cortex a century ago [1], we have learned an enormous amount about the brain by taking short snapshots of how it reacts to highly-controlled experiments, stimuli, or rest. In Chapters 3 and 4, I presented two examples of how we can take this knowledge and contextualize it in understanding how our brains continuously change over very long time periods in the chaotic real-world.

In Chapter 3, I asked the question "what are some of the biggest and most basic laws that seem to govern how the brain slowly changes over time?" We found that when we "zoomed into" individual regions and networks of the brain, their dynamics in isolation showed characteristic timescales and trajectories that were consistent across the week and were related to anatomy. In other words, if during the first day of recordings, one brain region appeared to be activating according to slow wave patterns whereas another was activating according to sharp bursts, those patterns would remain consistent throughout the week. Across subjects, there were consistent anatomical differences driving these trends. Additionally, the ways that these individual actors would interact in pairs also remained consistent over time.

When we "zoomed out" and looked at all measured networks of the brain, we found that they fell into a punctuated equilibrium of stable states and chaotic-like transitions.

When the brain fell into a stable state, its networks would remain relatively static for periods lasting from minutes to hours, and we could use this information to predict what a participant was doing during that time as well as their physiological status. When the brain transitioned between states, this corresponded to when the participant's own behavior was changing (such as going from reading a book to talking with a friend). During these transitions, instead of transferring directly from one state to another, the brain would explore a variety of intermediate states in a chaotic-like fashion before settling into a new stable state. Despite this seemingly unpredictable behavior however, the overall distributions on how these transitions occurred formed remarkably consistent power laws across subjects, indicating a fundamental shared mechanism on how this seemingly-random exploration was driven.

In Chapter 4, I took advantage of the hundreds of thousands of neural datapoints we could easily collect in a single participant to investigate whether there were central forces guiding the brain's activity. Using a recurrent neural network formulation of a Koopman operator, I found that the brain's overall dynamics seemed to be driven by a central attractor state at which the brain preferentially activates the default mode network while suppressing visual sensory ones. When the brain was engaging in active behavior (such as interacting with a friend), the brain would leave this central attractor in predictable trajectories before returning back to this attractor during times of wakeful rest (times when the participant was awake but not outwardly active).

The default mode network has long been under investigation as a "resting-state baseline" state of the brain based on the observation that individuals inside an fMRI scanner who are told to "not do anything other than stay awake" activate that network. Since then, it has turned into the *default* way to study whether brain activity is perturbed in some neural disease: collect resting state brain activity from individuals with the disease, compare this activity to individuals without it. This sort of data formed a significant part of the dataset used in Chapter 1 to predict depression treatment response. It formed the entirety of the

dataset I used to study how deep brain stimulation affects brain activity in Chapter 2.

However, in recent years cognitive neuroscientists and psychologists have increasingly argued that "sitting inside an fMRI machine and resting" in practice turns into "think introspectively and be internally mindful" which is arguably a very active state of cognition. Under this interpretation, what we call "resting-state" brain activity is simply one of many tasks (such as reading, writing, listening to music, etc.) you could ask someone to do while recording their brain activity. What does that mean for its clinical significance?

By taking long term recordings of the brain in the natural environment and asking what overall dynamical patterns exist, I found that something very similar to what is seen during resting-state fMRI emerges as a central attractor point. Cognitively, we are all bombarded with a myriad of stimuli, stressors, deadlines, new memories, and more throughout the course of an average day. These all pull our brains in various directions to deal with our immediate environment. But when we remove these immediate pressures and the need to take care of a current problem or challenge in our environment, what do our brains do? We get pulled into a state of internal focus where we sort through our thoughts and memories.

While this has not typically been labeled as such, this is not inconsistent with a physiological definition of homeostasis, now applied to how the brain turns its many networks on and off. If it is a homeostatic process, then what does that mean for the plethora of neurological and psychiatric disorders that neuroscientists have identified resting-state activity changes in? Shifts in homeostatic equilibrium points are commonly associated with pathology in every other organ system. There's a reason why our blood pressure equilibrates to roughly 120/80mmHg: it is high enough pressure to perfuse our organs while not being so high of a pressure that our heart has to work too hard to pump against it. Similar logic applies to our fasting blood sugar, our thyroid hormones, our sodium levels, and more. Is activation of the default mode network while suppressing networks capturing sensory inputs optimal for inward mindfulness? If someone with post

traumatic stress disorder [2] or traumatic brain injury [3] has hypersensitive or injured brain networks that push this central attractor state off balance, is that detrimental to our long-term cognitive health by the same logic that chronic hypertension is bad for our long-term cardiovascular health?

More broadly, I hope that these two projects demonstrate the utility of analyzing neural signals over continuous time periods orders of magnitude slower and longer than what is conventionally done in neuroscience. That it is possible to use this information to understand how the brain moves in and out of different behavioral and physiological states.

If we can learn brain transitions around behavior and physiology, then how about understanding transitions around pathological states? Most neurological and psychiatric disorders change on the timescales investigated in Chapters 3 and 4. Delirium, post-traumatic stress disorder, panic attacks, dementia, and more: these are not static diseases of the brain, they wax and wane. Patients report "good days" and "bad days", not good milliseconds and bad milliseconds. There must be some slow, consistent neural pattern underlying this fluctuation. Can we detect it?

Parts of this approach have been laid out in depression treatment trials performed in [4–6] where the authors used intracranial electrodes to capture short snapshots of neural activity as a patient's depression severity fluctuated over the course of a week and correlated the two together. Together with our work, this suggests a more comprehensive study of "how does someone's brain slowly enter and leave a pathological state" should be possible.

This type of approach is highly geared towards exploiting heterogeneity in disease patterns between different patients. If the goal of cancer genotyping is to answer "what is the difference between an individual's healthy cells and their tumorous ones", then the goal of this type of study would be "what is the difference between an individual's good

days and bad days". Rather than targeting every patient with a similar set of symptoms with an identical treatment, just as we tailor cancer therapeutics to specifically target malignant cells, we can learn to target specific, personalized temporal patterns of neural activity.

In the past, these approaches were limited by their sample size. By looking for short snapshots of activity during tightly controlled experiments, realistically they could only capture a modest number of samples from each patient. [5] used roughly a few dozen samples of neural activity spread out over seven days to correlate the power of a patient's neural activity at various frequencies in different electrodes to their depression severity using linear regression. While this demonstrates the feasibility of such an approach, the complexity of the brain's electrophysiology goes beyond simple "high activity" or "low activity" in an electrode: there are complex bursts, waves, and long-term temporal patterns underlying our neural states.

Detecting these temporal patterns requires more powerful methods that require sample sizes magnitudes in excess of what is used in conventional neuroscience paradigms. Analysis of continuous long term brain dynamics offers a feasible way to accumulate dataset sizes that begin to reach parity with their counterparts in other fields of machine learning. If we had a way to record neural activity from the brain of a single participant (either through invasive or non-invasive means) over one month, there are 2.6 million seconds in that month. 2.6 million examples of how their brain fluctuates through all manner of different situations, physiological states, environments, and behaviors.

Even if most of those data points lack physiologically or behaviorally interesting labels to directly link to their neural activity, these types of datasets are commonly used for transfer learning in other fields. The popularly known ChatGPT algorithm was developed by taking the GPT-3 family of models which were trained on massive amounts of unlabeled text (text that has not been assigned any "ground-truth" label by a human) and then fine-tuned these models based on feedback from a relatively small number of human

trainers [7, 8]. Facebook's wav2vec voice recognition algorithm demonstrated that by first training algorithms to generate robust statistical representations of un-transcribed speech audio alone, they could re-train these algorithms using a miniscule amount of transcribed speech audio to outperform state of the art algorithms using orders of magnitude less labeled data [9]. I used a similar approach in Chapter 4 to use unannotated long-term continuous neural recordings to learn a dynamical state representation of the brain that better allowed us to predict and understand how the brain enters different behavioral states such as reading a book or watching YouTube on a smartphone. In general, a variety of methods in machine learning exist to do this ranging from pretraining, self-learning, self-supervised learning, fine-tuning, and more [10]. This approach can become a model for investigating neural disease.

This proposal will be enabled by recent and ongoing dramatic improvements in wireless, wearable, and implantable technologies to collect continuous neural data in the natural environment. Surgically embedded wireless neural electronics are beginning to show promise, raising the possibility of collecting high-electrode intracranial data outside of the hospital over longer periods of time than previously feasible [11, 12]. We are also seeing constant advances in our capability to record neural data using noninvasive devices [13, 14]. These technologies have been mostly advertised as ways to implement brain-computer interfaces for those with sensory or motor deficits by translating rapid neural fluctuations over milliseconds to seconds into desired commands. While these are undeniably important problems to solve in medicine, this technology also has the potential to allow us to broadly study diseases of the brain at incredible detail over very long time periods.

I hope that this thesis can serve as an early "proof-of-concept" to demonstrate the viability of these kinds of analyses.

References

[1] Hans Berger. Über das elektrenkephalogramm des menschen. *DMW-Deutsche Medizinische Wochenschrift*, 60(51):1947–1949, 1934.

[2] Saskia BJ Koch, Mirjam van Zuiden, Laura Nawijn, Jessie L Frijling, Dick J Veltman, and Miranda Olff. Aberrant resting-state brain activity in posttraumatic stress disorder: A meta-analysis and systematic review. *Depression and anxiety*, 33(7):592–605, 2016.

[3] Eva M Palacios, Esther L Yuh, Yi-Shin Chang, John K Yue, David M Schnyer, David O Okonkwo, Alex B Valadka, Wayne A Gordon, Andrew IR Maas, Mary Vassar, et al. Resting-state functional connectivity alterations associated with six-month outcomes in mild traumatic brain injury. *Journal of neurotrauma*, 34(8):1546–1557, 2017.

[4] Katherine W Scangos, Ghassan S Makhoul, Leo P Sugrue, Edward F Chang, and Andrew D Krystal. State-dependent responses to intracranial brain stimulation in a patient with depression. *Nature medicine*, 27(2):229–231, 2021. Publisher: Nature Publishing Group.

[5] Jiayang Xiao, Nicole R Provenza, Joseph Asfouri, John Myers, Raissa K Mathura, Brian Metzger, Joshua A Adkinson, Anusha B Allawala, Victoria Pirtle, Denise Oswalt, et al. Decoding depression severity from intracranial neural activity. *Biological Psychiatry*, 2023.

[6] Sameer A Sheth, Kelly R Bijanki, Brian Metzger, Anusha Allawala, Victoria Pirtle, Joshua A Adkinson, John Myers, Raissa K Mathura, Denise Oswalt, and Evangelia Tsolaki. Deep brain stimulation for depression informed by intracranial recordings. *Biological Psychiatry*, 92(3):246–251, 2022. Publisher: Elsevier.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla

Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] OpenAI. Chatgpt: Optimizing language models for dialogue, Nov 2022.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[10] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020.

[11] Thomas J Oxley, Nicholas L Opie, Sam E John, Gil S Rind, Stephen M Ronayne, Tracey L Wheeler, Jack W Judy, Alan J McDonald, Anthony Dornom, Timothy JH Lovell, et al. Minimally invasive endovascular stent-electrode array for high-fidelity, chronic recordings of cortical neural activity. *Nature biotechnology*, 34(3):320–327, 2016.

[12] Elon Musk et al. An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194, 2019.

[13] Rebecca A Frederick, Philip R Troyk, and Stuart F Cogan. Wireless microelectrode arrays for selective and chronically stable peripheral nerve stimulation for hindlimb movement. *Journal of Neural Engineering*, 18(5):056058, 2021.

[14] Kaido Värbu, Naveed Muhammad, and Yar Muhammad. Past, present, and future of eeg-based bci applications. *Sensors*, 22(9):3331, 2022.