COMPUTATIONAL METHODS TO ADVANCE INDIVIDUAL PRECISION IN BRAIN MAPPING

By

Roza Gunes Bayrak

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

August 11, 2023

Nashville, Tennessee

Approved:

Catherine Elizabeth Chang, Ph.D.

Matthew Berger, Ph.D.

Tyler Derr, Ph.D.

Mika Rubinov, Ph.D.

Dario J. Englot, MD, Ph.D.

This dissertation is dedicated to my sweet and brilliant advisor, Catherine Elizabeth Chang, and to my love and partner in this lived, breathed and written crime, Colin Blake Hansen.

ACKNOWLEDGMENTS

I am incredibly grateful for the folks who have provided their unwavering support and guidance throughout the completion of this thesis. As they say, it takes a village to finish a PhD.

First and foremost, I would like to thank my advisor, Catherine Elizabeth Chang, who has been a brilliant mentor and a great friend. You have consistently cheered me on, guided me, taught me and grew together with me. It's not easy to find an advisor who's both an expert in their field and genuinely caring, but you've exceeded all expectations. Your friendship and the depth of our connection mean the world to me.

To my committee members, Matthew Berger, Tyler Derr, Mikail Rubinov, and Dario Englot, I extend my heartfelt gratitude. Your contributions as collaborators, mentors, and advisors have been invaluable.

I am grateful to Ilwoo Lyu, Kurt Schilling, Yuankai Huo, and Bennett Landman for the scientific wisdom I have gained from our interactions. It has been amazing to interact and collaborate with many people in the Vanderbilt community. You have all inspired and taught me; thanks Baxter Rogers, Ipek Oguz, Shunxing Bao, Prasanna Parvathaneni, and many others that I can't possibly fit the list here.

The VISE and women of VISE (woV), and in particular, Michelle Bukowski, I am deeply grateful for your instrumental role in helping me cultivate confidence and Bill Rodriguez, your friendship and energy has lead to many strenuous bike rides. Thank you for being and invaluable source of inspiration and strength throughout this journey. I learnt something from everyone that I have met through woV. I have gained a treasure trove of wisdom from the diverse experiences shared by the women around me, which have equipped me with a wide range of valuable tools in navigating the challenges of academia, especially when faced with isolation and imposter syndrome.

Haley, my first friend across the hall and the last friend to share in the camaraderie. Karthik Ramadass, it is very comforting to know that there is someone who got your back, thank you. Cailey Kerley, Nhung Hoang, Pratu Kanakaraj, Neda Sardaripour, and Colette Abah, thank you for the best night-ins and night-outs. Thank you Devika and Diego Apples for either creating or being part of all of our latest adventures. Thank you, Sarah Goodale and Caroline Martin, and all the neurdies, for making every trip to the lab, one worth taking. Ahmet Cakir, Srijata Chakravorti, Francois-Xavier Carton, and so many others friends that I made along the way, thank you for making me feel belong.

Viviana Grice and Jonna Whitman, thank you for keeping me mentally and physically healthy throughout this challenging journey. Front desk SHC, Morris, Darian, and Valerie thank you for always greeting me by my first name and smiles. You've brightened my sick days. Thank you Outdoor Rec folks, for some of the best memories here at Vanderbilt.

Mom and Dad, Gamzegul and Osman Nuri, thank you for allowing me to fail, grow, learn on my own, thank you for the "I can do anything" gene and its reinforcement by example. More than I show it, I look up to you. Thank you for the copious amount of phone calls over the duration of this PhD, for always distracting me from the stresses of academia and reminding me of a world outside of it. I love you. My grandparents, Cezminur and Haydar as well as my favorite aunt, Aymelek, for their dedication to my education and unconditional love. Thank you, Nuri Erek, for believing in me and always showing up, stepping up for me. I would like to also thank to my hidden heroes, close and distant family.

Thank you Karen Panetta, Yahya Kemal Baykal, Sule Sekreter, for being great role models always being there for me. Thank you, Ilkay Aydin, for your friendship, and when I lost direction being my lighthouse. Burcu Yakisir, Cagatay Durmaz without whom I would not be whole, thank you for being you.

Last but certainly not least, I want to express my heartfelt appreciation to our furry babies, Fjord Shoebill and Rhombus Flatbread. Your presence in my life has brought immeasurable joy, serenity, and even moments of heart attacks. Thank you for all the oxytocin, dopamine and adrenaline that you have provided, my stress busters and mood lifters.

Colin Blake Hansen, my best friend, my person, your curly hair and furly (yes, that's a word now) sense of humor have been instrumental in keeping me sane throughout this journey. From the adventurous escapades we've embarked on to the cozy nights watching Gilmore Girls, I can't thank you enough for being by my side. I don't know how I would have done this (PhD) without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

The Mind Through the Lens of Neuroimaging

The proper level of discourse for cognitive theory concerns information, not the medium used to carry it.

— Stephen Palmer

The brain is a complex organ made up of billions of neurons and other cells, connected by trillions of synapses. These neurons communicate with each other through electrical and chemical signals, forming intricate networks that give rise to our thoughts, emotions, and behaviors. The mind is thought to be an emergent property of these interactions, arising from the computations and information processing that takes place within the brain. The exact mechanisms by which the brain gives rise to the mind are still not fully understood.

The study of the most complex system known is an interdisciplinary endeavor that draws upon research from various fields including philosophy, neuroscience, psychology, engineering, mathematics, linguistics, computer science, and biology. Researchers in these diverse disciplines employ a wide range of methods and approaches to explore the nature of the brain. Among the significant advancements in brain research, the development of neuroimaging techniques has played a pivotal role. Neuroimaging techniques allowed researchers to non-invasively measure brain activity in humans while they perform cognitive tasks or experience different mental states. By observing the brain in action in real-time, neuroimaging has revolutionized our understanding of the neural mechanisms underlying cognition and behavior. Neuroimaging has been instrumental in enabling scientists to observe the human brain at work in real-time without invasive procedures, uncovering fascinating insights into the spatial and temporal changes underlying a broad range of brain functions.

Despite these advances, our understanding of the mechanisms of brain function remains incomplete, with inter-individual variability representing a significant challenge. To achieve a better understanding of brain function in health and disease and to understand what makes us unique, it is crucial to interpret neuroimaging data at the individual level. This thesis aims to address some of the current challenges in individualized mapping of the human brain.

The human brain, with its multitude of cognitive processes, emotions, perceptions, and behaviors, poses a formidable challenge for researchers. However, through the integration of neuroimaging and advanced machine learning methodologies, scientists aim to unravel the complex mechanisms that underlie these functions. By analyzing brain structure and dynamic activity patterns, these interdisciplinary studies offer valuable insights

into fundamental aspects of human cognition, mental health, and neurological disorders. In this dissertation, we explore the cutting-edge advancements in studies of the brain function, examining the promises, challenges, and future directions that lie ahead in our quest to uncover the mysteries of the mind.

1 Contemporary Study of the Mind

In this chapter, I provide an overview of the technology that is relevant to this dissertation, magnetic resonance imaging (MRI). I discuss the different types of MRI techniques and their respective strengths and limitations, emphasizing the critical role of computational techniques in processing and analyzing large-scale neuroimaging data. Next, I delve into the literature that focuses on the current challenges in individualized mapping of the human brain. This includes discussions of functional and structural variability across individuals and the importance of incorporating concurrently collected information when interpreting neuroimaging data. I also explore how individualized mapping of the human brain can provide valuable insights into cognition and behavior. Finally, I provide an overview of the work that will be covered in this thesis, a set of new computational solutions to facilitate individual precision in addressing progressively more complex and interesting questions about brain function.

1.1 Neuroimaging: Non-invasive Windows Into the Brain

Modern brain imaging techniques provide non-invasive windows into the anatomy and function of the brain. Since its inception, magnetic resonance imaging (MRI) has become one of the most widely used techniques and is the workhorse of investigations of the human brain in health and disease.

In the 1970s, the discoveries made by Paul Lauterbur and Peter Mansfield have made it possible to develop modern MRI techniques based on the nuclear magnetic resonance (NMR). Diffusion-weighted magnetic resonance imaging (dMRI) has, since the 1980s, been widely used to track the diffusion of water molecules in the brain. In a paper published in 1990, Seiji Ogawa and colleagues demonstrated the blood-oxygen-level-dependent (BOLD) effect (described below) that underlies functional magnetic resonance imaging (fMRI).

Different MRI sequences can provide complementary views in neuroimaging research. For instance, structural MRI has become a standard part of clinical procedures and images brain morphology; dMRI is shedding light onto the white matter fiber organization and anatomical connectivity of the brain; and fMRI provides a window into on-going neural activity.

1.1.1 Structural Windows

Structural magnetic resonance imaging (sMRI) captures morphometric information about the brain. T1-weighted (T1w) imaging is one of the most common types of MRI techniques. T1w images are the most descriptive of

white matter and bone marrow. However, they have low sensitivity to subtle changes in tissue properties, such as water content, blood flow, or metabolism.

Diffusion-weighted magnetic resonance imaging (dMRI) uses the basic principle of diffusion to track water molecules within the microstructure of the brain. Water molecules diffuse slowly across – and fast within – the direction of white matter pathways. If water diffusion is measured in many directions via diffusion weighted MRI, it allows for mapping of the microstructure and organization of these white-matter fiber pathways (Basser and Pierpaoli, 2011). dMRI is useful for detecting acute ischemic stroke, as water molecules are restricted in the affected area as well as showing changes in tissue microstructure, such as axonal damage, inflammation, or tumor infiltration. However, dMRI can be distorted various artifacts (Le Bihan et al., 2006), including eddy currents and motion.

1.1.2 Functional Window

The main domain focus in this thesis, fMRI, is based on the blood-oxygen-level-dependent (BOLD) contrast mechanism (Ogawa et al., 1990). Through fMRI, fluctuations in neural activity are detected by subtle changes in blood oxygenation elicited by neural activity. The hemodynamic response to neural activity can be summarized as follows. An increase in local neural activity results in a transient increase in deoxyhemoglobin (deoxy-Hb), which may be visible as an initial dip in the BOLD response. Concurrently, neurovascular signals trigger vasodilation, leading to enhanced blood flow, which over-oxygenates the local tissue. This excessive flow of oxyhemoglobin (oxy-Hb) changes the relative concentration levels, resulting in an increase in the MRI signal intensity. The effect is slow, depending on the nature of the stimulus it peaks around 3-5 seconds after the arrival of the stimulus (Hillman, 2014). This relative concentration change in the deoxy-Hb is measured through fMRI (see Figure 1.1) and is reflected in the intensity changes of T2- and T2*-weighted images.

Task, Resting-State, and Naturalistic fMRI Paradigms

fMRI research is largely based on experimental paradigms that aim to probe neural activity during tasks or which examine task-free activity. ask studies may be described as a ”first wave” of fMRI research. In task fMRI, controlled task paradigms attempt to isolate specific cognitive processes and subsequently investigators seek to identify brain regions with associated cognitive functions. Due to its strong interpretability, task fMRI has been widely adopted and provides insights about the mind in addition to the brain (Finn, 2021). However, one limitation of task-fMRI is that it ignores a large fraction of variance that arose from task-unrelated, seemingly spontaneous signals. Another limitation is that task fMRI paradigms are often highly controlled and may

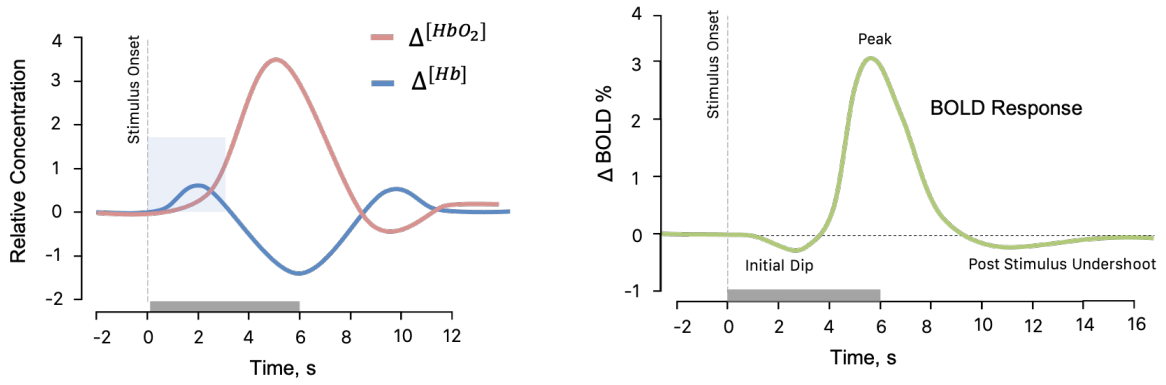


Figure 1.1: Neural activity (grey bar) causes extraction of oxygen from blood and triggers neurovascular signaling that is followed by an increase in blood flow and oxygenation. When a stimulus arrives, cellular oxygen metabolism briefly increases the relative concentration level of deoxy-Hb (blue box), visible as an initial dip. This is followed by an inflow of blood to the tissue due to hyperemia caused by vasodilation, resulting in the transient increase of the oxy-Hb to deoxy-Hb ratio. This effect increases the T2*-weighted MRI signal relative to its baseline value, and it is what is measured with the BOLD signal. Figure adapted from (Cinciute, 2019).

not accurately reflect real-world cognitive processes. A second wave of fMRI experiments came with resting-state fMRI, where subjects lie still in scanner with no externally imposed task. This paradigm has generated significant breakthroughs, including the identification of large-scale networks and their interactions, providing a comprehensive view of brain organization. Resting-state scans are easy to acquire in comparison to task paradigms, as no meticulous study design or specialized equipment is necessary. Consequently, resting-state fMRI enables the examination of variations in brain dynamics within non-normal populations including infants and those with significant cognitive or physical impairments. This further allows identification of neural patterns that extend into clinical domains such as in neurodegenerative diseases (Hohenfeld et al., 2018), aging (Ferreira and Busatto, 2013) and neurodevelopmental disorders (Uddin et al., 2010). One limitation of resting-state fMRI is that it can be difficult to interpret. This could be mitigated by obtaining first-person reports of mental experiences from participants, either retrospectively or in real-time, however this still provides only a few anchor points for observations. Resting-state fMRI, though useful, is inherently limited as it does not offer insights into the modulation of brain activity in response to specific stimuli. Finally, a new shift in fMRI paradigms has led to tasks in which subjects engage with naturalistic stimuli (i.e. audio narratives, movies or video games) which exhibit richer context. These stimuli are more dynamic and ecologically valid than traditional tasks. The shift towards naturalistic paradigms in fMRI research provides the controllability of a lab environment while reflecting real-world cognitive processes (Finn, 2021). Moreover, naturalistic tasks have the advantage of capturing both shared and idiosyncratic components in neural responses, offering insights into individual differences in brain function. Naturalistic paradigms present many opportunities for high-throughput

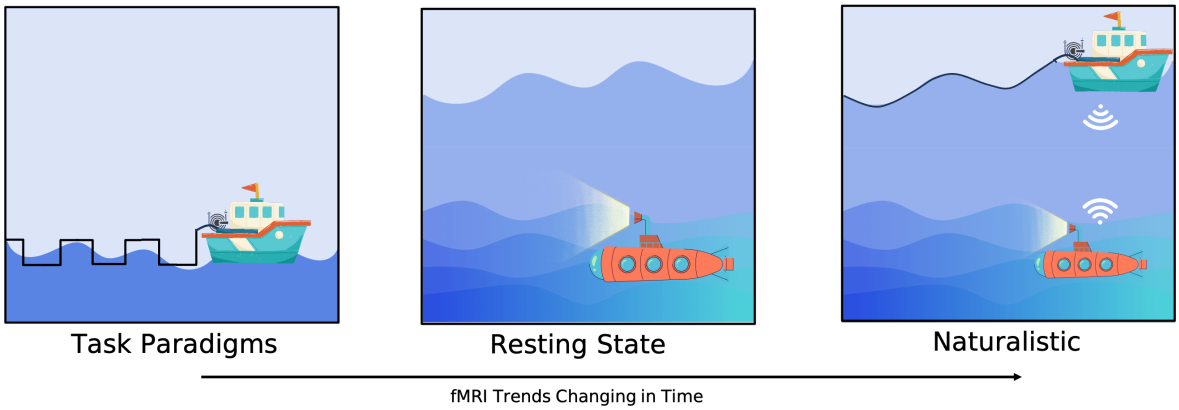


Figure 1.2: fMRI paradigms. Figure adapted from (Finn, 2021).

functional characterization of the human brain (Zhang et al., 2021c).

The Brain and Body Connection

More common than not, the brain is studied in isolation. However, we also know that brain and body are closely interconnected and continuously influencing each other. The brain-body connection is a complex phenomenon that involves the interaction of neural, vascular, and metabolic processes. fMRI relies on the fact that cerebral blood flow and neuronal activation are coupled and measures a surrogate for brain activity by detecting changes associated with blood flow. This means that our measures contain both confounds and sources of information that need to be carefully considered. One significant factor that influences BOLD signals is peripheral physiological signals. These signals, which include measures of heart rate and respiration could be recorded during fMRI scans and have been observed to co-vary with what we expect as neural components. Historically, these signals were regarded as mere confounds and were removed from the BOLD data. However, it is now evident that the relationship between peripheral physiological signals and neural activity is more complex than initially thought (Chen et al., 2020; Bright et al., 2020; Yoo et al., 2022; Nashiro et al., 2022). For example, our thought processes can alter the rate of breathing (Zelano et al., 2016), and in return, breathing can regulate behavior (Molle and Coste, 2022).

1.1.3 A Necessary Evil: Preprocessing

Preprocessing refers to a series of data preparation steps that are applied to raw or unprocessed data before further analysis or interpretation. In the context of neuroimaging, preprocessing involves specific techniques and procedures applied to structural MRI (sMRI), diffusion MRI (dMRI), or functional MRI (fMRI) data. Preprocessing is necessary in neuroimaging for several reasons:

1. It aims to enhance the quality and reliability of the acquired data. It helps address common artifacts and sources of noise, such as motion-related distortions, physiological noise, scanner drift, or imaging artifacts. By minimizing these unwanted effects, preprocessing improves the signal-to-noise ratio, enhances the accuracy of subsequent analyses, and increases the likelihood of detecting true neural signals.

2. It allows for data standardization, enabling meaningful comparison across subjects or groups. Techniques like spatial normalization or registration align data to a common anatomical space, enabling group-level analyses and statistical comparisons. Preprocessing also reduces inter-subject variability by minimizing the impact of individual differences in brain anatomy or imaging acquisition.

3. It aims to mitigate confounding factors that can distort or bias the analysis results. Steps such as motion correction, physiological noise removal, or artifact detection and correction help reduce the impact of confounds on the data, allowing for more accurate and reliable interpretation of the neural signals of interest.

4. It is often tailored to specific analysis techniques or models. By applying appropriate preprocessing steps, the data can be prepared in a manner that aligns with the assumptions and requirements of the chosen analysis methods. This optimization improves the sensitivity and validity of subsequent analyses, leading to more accurate interpretations and scientific conclusions.

However, prepossessing can introduce biases and errors that are hard to detect and correct, reduce the variability and complexity of the data, leading to overfitting and false positives, affect the reproducibility and replicability of the results (Bhagwat et al., 2021; Schilling et al., 2021a; Marek et al., 2022). Different choices of parameters and software can lead to different outcomes. For example, a preprocessing variability can be introduced with as simple of a choice as normalization method. Normalization is a commonly used in preprocessing and aims to remove inter-subject variability in brain structure and function. However, different normalization methods can lead to different outcomes and affect the results of neuroimaging studies.

Here, I provide a brief overview of the common preprocessing steps in each modality.

Structural MRI

Structural MRI (sMRI) data preprocessing involves several important steps to improve the quality and accuracy of the data. They include techniques such as bias field correction, removal of non-brain voxels, registration to a common space, tissue segmentation, and surface reconstruction.

It is essential to assess the quality of the acquired sMRI data. This step involves checking for artifacts, signal dropout, motion-related issues, and other factors that may impact data quality. sMRI images often suffer from intensity inhomogeneities known as bias fields. Bias field correction methods are applied to normalize the image intensities and remove these artifacts. The non-brain voxels are removed from the sMRI. This process, known as brain extraction or skull stripping, helps reduce partial volume effects and eliminates the contamination of brain measurements with extracranial signals. Image registration techniques are used to align it with a standardized anatomical space, facilitating inter-subject comparison and group analyses. Tissue segmentation aims to classify the brain image into different tissue types, such as gray matter, white matter, and cerebrospinal fluid. Accurate tissue segmentation is essential for subsequent analyses and quantification of brain structures. For advanced analyses of cortical structures, such as cortical thickness or surface-based morphometry, cortical surface reconstruction is performed. This step involves generating a representation of the brain's outer surface and mapping the structural data onto it.

Diffusion MRI

Diffusion MRI (dMRI) data preprocessing involves several steps to improve the quality and reliability of the data. While there is no consensus on what should be included in a preprocessing piepeline, key steps commonly applied are distortion (susceptibility, eddy current, inhomogeneity and motion) correction, bias field correction, removal of non-brain voxels, registration, diffusion modeling and fiber tracking.

Distortion correction refers to the process of correcting geometric distortions in an image caused by factors such as magnetic field inhomogeneities, susceptibility effects, or eddy currents. These distortions can result in misalignment between different image volumes or even signal drop-outs. The bias field refers to spatial intensity variations that can occur in MRI images. These variations may arise due to imperfections in the MRI scanner's magnetic field, coil sensitivity, or other factors. Correction methods are employed to compensate for these distortions. The non-brain voxels are removed from the dMRI data to focus the analysis on the brain regions of interest. Similar to structural MRI, registration is applied to dMRI data to align it with a standardized anatomical space (i.e. Talairach, MNI), establishing correspondence between voxels across images and facilitating inter-subject comparison and group analyses. Finally, advanced diffusion models, such as high-angular-resolution diffusion imaging (HARDI) or constrained spherical deconvolution (CSD), can be

employed to capture complex fiber orientations. Fiber tracking algorithms are then applied to reconstruct white matter pathways and estimate connectivity.

The quality and accuracy of results in dMRI are impacted by the effectiveness of these preprocessing steps. Inaccurate motion correction can introduce artifacts and affect the quality of estimated diffusion metrics. Errors in gradient distortion correction may lead to misalignment and incorrect estimation of fiber orientations. Careful consideration and optimization of each preprocessing step are crucial for obtaining reliable and meaningful insights into white matter connectivity and microstructure from dMRI data.

Functional MRI

Preprocessing fMRI data involves several steps aimed at enhancing the quality and reliability of the data before subsequent analysis. While there is no consensus on preprocessing, key steps involved are slice-timing correction, coregistration, spatial normalization or registration, brain extraction, spatial smoothing, temporal filtering and normalization, artifact detection and correction, and commonly dimensionality reduction.

Slice timing correction corrects for differences in image acquisition timing between slices within each volume. It ensures that the signal across slices is synchronized. Failure to correct for slice timing differences can introduce systematic errors in the temporal dynamics of the data. Coregistration aligns each volume in the fMRI time series to a reference volume, minimizing motion-related artifacts. Excessive motion can introduce spurious signal changes and distort connectivity patterns, potentially leading to erroneous results. Proper motion correction reduces these artifacts and improves data quality. Spatial smoothing involves applying a spatial filter to the fMRI data, usually a Gaussian (2-8mm FWHM), which reduces high-frequency noise and enhances the signal-to-noise ratio. Spatial smoothing can help improve the detection of activations and enhance the statistical power of subsequent analyses. However, excessive smoothing may blur the data and lead to loss of fine-grained spatial details. Temporal filtering involves filtering the fMRI time series to remove effects such as low-frequency drifts and high-frequency noise. A common approach is to use a bandpass filter to retain the frequency range associated with the hemodynamic response. Temporal filtering can help improve the sensitivity to neural activity of interest and reduce noise-related confounds. Artifact detection and correction is possibly the most controversial step in fMRI preprocessing. Various artifacts, such as physiological noise, scanner artifacts, and susceptibility-related distortions, can affect fMRI data quality. Techniques, which are either data-driven or based on parametric modeling, can be employed to identify and remove these artifacts. As one of the final processing steps, dimensionality reduction techniques could be applied to high-dimensional spatio-temporal data. Dimensionality reduction techniques assigns voxels to specific regions or networks, thereby reducing the dimensionality from the voxel level to the parcel level.

The quality and accuracy of the results are influenced by the effectiveness of these preprocessing steps. Failure to adequately correct for motion can lead to spurious correlations and reduced sensitivity to subtle effects. Inaccurate spatial normalization can introduce misalignment and distort statistical inference. Improper spatial smoothing or excessive filtering may lead to biased results or loss of important information. Therefore, careful consideration and optimization of preprocessing steps are crucial for obtaining reliable and meaningful results from fMRI data.

1.2 Machine Learning

Magnetic resonance imaging (MRI) is a valuable tool for investigating the human brain and its functions. However, analyzing MRI data can be challenging due to factors such as high dimensionality, complexity, and noise. Advanced machine learning techniques play a pivotal role in addressing these challenges and analyzing MRI data. These techniques enable the identification of complex patterns and relationships within the data that may not be apparent through traditional analytical approaches. Machine learning algorithms can handle the high dimensional data and effectively model the complex interactions between brain regions. They can aid in tasks such as classification, prediction, feature selection, and connectivity analysis, ultimately contributing to the study of brain function.

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions. Machine learning techniques can be broadly classified into two main categories: supervised and unsupervised or self-supervised. Supervised learning involves learning a function that maps inputs to outputs based on labeled examples, while unsupervised or self-supervised learning involves discovering patterns or structures in unlabeled data.

Some of the benefits of using advanced machine learning techniques to study MRI are:

- They can reduce the dimensionality of fMRI data and select the most relevant features for analysis. For example, principal component analysis (PCA) can transform fMRI data into a lower-dimensional space that captures the most variance in the data.
- They can discover hidden patterns in data that reveal new insights into brain function. For example, canonical correlation analysis (CCA) can identify the linear combinations of fMRI data and behavioral or cognitive variables that are maximally correlated with each other.
- They can model the nonlinear and dynamic relationships between structural and functional data. For example, neural networks (NNs) can learn complex nonlinear functions that map cortical sheet fatures to functional brain maps, while recurrent neural networks (RNNs) can capture the temporal dependencies

and dynamic changes in data over time.

- They can discover meaningful patterns and relationships within brain imaging data without relying on explicit task labels. For example, self-supervised learning can be used to predict the missing parts of a brain image based on the surrounding information to recover corrupted data.

Advanced machine learning techniques have become indispensable tools in exploring MRI data and pushing the boundaries of our understanding of the human brain. These techniques offer unique advantages that overcome the challenges associated with MRI data analysis, which traditional approaches may struggle to address. By leveraging machine learning algorithms, researchers aim to effectively handle the high dimensionality, complexity, and noise inherent in MRI data. These techniques enable the extraction of intricate patterns, relationships, and features that may not be discernible through conventional analytical methods. The ability to automatically uncover hidden insights and capture complex interactions within the data opens up new avenues for studying brain structure, function, in health and in disease. Moreover, machine learning empowers researchers to integrate multimodal data, develop personalized models, and automate various steps in the analysis pipeline. This combination of advanced machine learning and MRI data analysis has the potential to revolutionize our understanding of the human brain and drive advancements in neuroscience and clinical applications.

1.3 Big Data

"Big data" is a term that refers to the collection and analysis of large-scale and complex datasets. Researchers leverage the power of big data analytics to integrate and interpret multiple sources of information, such as neuroimaging, genetics, behavior, and cognition, to discover new patterns and insights, test hypotheses and theories on a larger and more diverse scale, and validate and refine existing models and methods with more accuracy and reliability.

Large-scale datasets offer several unique advantages over traditional small datasets, enabling researchers to explore patterns and mechanisms underlying brain function and dysfunction in ways that were previously unattainable. Small datasets are more susceptible to chance variations and noise, which can make it challenging to generalize findings to the broader population. In contrast, large-scale datasets encompass a more extensive range of subjects, including individuals from different demographics, age groups, and clinical conditions. This diversity allows researchers to investigate individual differences and identify patterns that may be specific to certain populations or brain disorders. It also helps in obtaining more reliable and robust results by reducing the impact of outliers and confounding factors. With a larger sample size, researchers can detect smaller effect sizes with greater accuracy. This is crucial in the field of neuroscience, where the underlying brain processes can be

complex and subtle. Large-scale datasets enable researchers to identify more nuanced relationships between variables, uncovering fine-grained patterns that were previously hidden due to limited statistical power.

Additionally, large-scale datasets foster collaboration and data sharing among researchers. By pooling data from different research institutions and initiatives, the collective knowledge and resources can be leveraged to accelerate scientific progress. Of note, this poses significant challenges and ethical issues for neuroscience research. Some of these challenges include the management, storage, and sharing of large and heterogeneous datasets; the quality, validity, and reproducibility of big data analysis. However, fostering responsible data governance can balance the benefits and risks of using big data for neuroscience research (Poline et al., 2012; Gorgolewski et al., 2016; Fothergill et al., 2019).

2 What are the challenges?

2.1 Individual Precision

Neuroimaging research has revolutionized our understanding of the human brain, providing insights into its functions and dynamics. Nevertheless, the complex nature of the brain itself, along with the medium through which it is captured, namely MRI, introduces challenges that necessitate individualized approaches. Understanding the factors that contribute to these individual differences, based on neuroimaging data, has broad implications for advancing our understanding of the neural underpinnings of individual variability. While the utilization of individualized approaches in neuroscience research has garnered increasing interest (Bijsterbosch et al., 2018; Wang et al., 2015; Kong et al., 2021), to date, fMRI research has been concerned primarily with data summaries at the population level. Certainly this approach has benefited group comparisons, e.g. disease progression, severity classification, or early diagnosis in clinical populations, but it falls short when more individualized characterization of brain function is needed (Dubois and Adolphs, 2016). Consequently, there is a critical need for methods that enable the identification of nuanced distinctions between individuals. To this end, the contributions in this thesis focus on developing tools and computational models for neuroimaging data with precision at the level of individuals.

One of the main goals in medicine is enhancing the diagnosis, prognosis, and treatment of neurological and psychiatric disorders. Clinical neuroimaging provides a medium to identify individualized biomarkers that can capture the heterogeneity and complexity of brain diseases. Identifying biomarkers specific to each patient, capturing the unique characteristics of their condition has broad implications. By doing so, we can tailor our diagnostic and treatment strategies to meet the personalized needs of each individual, optimizing outcomes and minimizing potential side effects. Individual precision in neuroimaging also holds great potential in the realm of treatment evaluation and monitoring. By utilizing advanced imaging techniques, we can objectively assess

the efficacy of interventions and track changes in brain function and structure over time. This allows us to fine-tune treatment approaches, optimizing outcomes and ensuring the best possible care for the patients.

Through advanced imaging techniques, the neuroimaging community aims to develop methodologies that can identify individualized biomarkers and capture the complexity of brain disorders. This information then can be used to guide clinical decision making with the ultimate goal of improving patient outcomes, transforming the diagnosis and treatment of neurological and psychiatric disorders, and advancing the field of neuroscience.

2.2 Open Science

In the rapidly advancing fields of machine learning and neuroimaging, ensuring the quality and reliability of research results is of paramount importance. Open science plays a vital role in promoting transparency, accountability, and advancement in scientific research. By embracing these principles in neuroimaging research, we can foster collaboration, drive innovation, and strengthen the credibility of our findings. Open science is a philosophy and a set of practices aimed at making scientific research more transparent, accessible, and inclusive. It encompasses the principles of openness, transparency, reproducibility, and collaboration. In the context of neuroimaging research, open science entails making research data, methods, code, and publications freely available to other researchers and the public. By removing barriers to accessing and replicating research, open science enhances the credibility and impact of scientific discoveries.

Reproducibility and Verification: Openly sharing data, analysis code, and methods allows other researchers to reproduce and verify research findings. This process strengthens the scientific rigor and reliability of neuroimaging studies, enabling the validation of results and facilitating the identification of potential errors or biases.

Collaboration and Innovation: Open science fosters collaboration among researchers by providing a platform for sharing ideas, methodologies, and datasets. Collaboration fuels innovation, as researchers can build upon each other's work, explore novel research questions, and leverage diverse perspectives to advance the field of neuroimaging.

Transparency and Accountability: Open science promotes transparency by providing a clear view into the research process. Researchers are accountable for their methods, ensuring that experiments are conducted ethically and data is analyzed accurately. This transparency also allows for scrutiny and constructive feedback from the scientific community, ultimately improving the quality of research.

Maximizing Research Impact: Openly accessible research enables a broader audience to benefit from scientific findings. Patients, clinicians, policymakers, and the general public can gain insights and make informed decisions based on the latest neuroimaging research. Additionally, open science facilitates the dissemination

of knowledge in developing countries or under-resourced institutions, promoting global scientific progress.

Open science is crucial for the advancement of neuroimaging research. By embracing these principles, we can enhance the credibility, transparency, and impact of our findings, while promoting collaboration and innovation. Overcoming challenges and embracing a culture of open science will pave the way for personalized and effective diagnostic tools, treatments, and interventions in the realm of neuroimaging.

Variability in the neuroimaging analyses have substantial effects on scientific conclusions. This emphasizes the importance of validating and sharing complex analysis workflows (Botvinik-Nezer et al., 2020). One of these efforts is community driven, open source data preprocessing pipelines (Esteban et al., 2019a,b). Complex datasets should be analyzed using several analysis pipelines, preferably by more than one research team, to improve reproducibility of data analysis outcomes.

3 Outline of Thesis

Deriving insights about brain function from neuroimaging data is challenging and researchers are limited by the theoretical and experimental tools at their disposal (Lynn and Bassett, 2019). The large variation in brain anatomical structures and functional organization across people (Suárez et al., 2020; Finn and Constable, 2022) poses a major challenge. Additionally, modality-specific challenges arise; for example, in fMRI, a key challenge is that the signals only indirectly relate to neural activity and are influenced by physiological signals with complex spatial and temporal correlation to those of neurally generated fMRI signals (Bright et al., 2020; Chen et al., 2020). Here, we investigated how peripheral physiology uniquely influences brain function and its recordings (Chapter III). Chapter II proposes a learning frameworks to infer heart rate and respiration variation directly from fMRI data itself.

Another general challenge in analyzing MRI data arises from its high dimensionality and low signal-to-noise ratio. In this thesis, we offer individualized parcellation approaches first to robustly reduce the spatial dimensional of functionally uniform brain regions and to delineate accurate fiber tract pathways. Functional parcellations are studied through interactive analytics (presented in Chapter IV) and advanced machine learning techniques (detailed in Chapter V). Fiber tracking is studied through manual delineation protocols (addressed in Chapter VI). Finally, I summarize the main contributions, propose future work, and describe studies and collaboration not covered in the main dissertation.

3.1 Publications

The work described in the body of this thesis has been in peer-reviewed journals or full-length conference proceedings. Additionally, there is three more journal submissions are currently going through internal review

prior to submission for publication.

Chapter II

- **Roza G. Bayrak**, Jorge A. Salas, Yuankai Huo, Catie Chang. "A Deep Pattern Recognition Approach for Inferring Respiratory Volume Fluctuations from fMRI Data.", **NIH Award**, In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020.

- **Roza G. Bayrak**, Colin B. Hansen, Jorge A. Salas, Nafis Ahmed, Ilwoo Lyu, Yuankai Huo, Catie Chang. "From brain to body: Learning low-frequency respiration and cardiac signals from fMRI dynamics.", In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021.

- **Roza G. Bayrak**, Colin B. Hansen, Jorge A. Salas, Nafis Ahmed, Ilwoo Lyu, Mara Mather, Yuankai Huo, Catie Chang. "Tracing peripheral physiology in low frequency fMRI dynamics", (under revision) (2022).

Chapter III

- **Roza G. Bayrak**, Nafis Ahmed, Ruoqi Yang, Mara Mather, Catie Chang. "Physiological Signatures Across the Brain", (in preparation) (2023).

Chapter IV

- **Roza G. Bayrak**, Nhung Hoang, Colin B. Hansen, Catie Chang, Matthew Berger. "PRAGMA: Interactively Constructing Functional Brain Parcellations.", **Best Short Paper Honorable Mention**, IEEE Visualization Conference (VIS), 2020.

Chapter V

- **Roza G. Bayrak**, Ilwoo Lyu, Catie Chang. "Learning Subject-Specific Functional Parcellations from Cortical Surface Measures", PRIME Workshop In International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2022.

- **Roza G. Bayrak**, Ilwoo Lyu, Catie Chang. "Cortical sheet features featuring function", (in preparation) (2023).

Chapter VI

- **Roza G. Bayrak***, Francois Rheault*, Xuan Wang, Kurt G. Schilling, Jasmine M. Greer, Colin B. Hansen, Cailey Kerley, Karthik Ramadass, Lucas W. Remedios, Justin A. Blaber, Owen Williams, Lori L. Beason-Held, Susan M. Resnick, Baxter P. Rogers, Bennett A. Landman. "TractEM: Evaluation of Protocols for Deterministic Tractography White Matter Atlas." Magnetic Resonance Imaging (2021).

CHAPTER 2

Tracing peripheral physiology in low frequency fMRI dynamics

Many studies of the human brain using functional magnetic resonance imaging (fMRI) lack physiological measurements, which substantially impacts the interpretation and richness of fMRI studies. Natural fluctuations in autonomic physiology, such as breathing and heart rate, provide windows into critical functions including cognition, emotion, and health, and can heavily influence fMRI signals. Here, we developed *DeepPhysioRecon*, a Long-Short-Term-Memory (LSTM)-based network that decodes continuous variations in respiration amplitude and heart rate directly from whole-brain fMRI dynamics. Through systematic evaluations, we investigate the generalizability of this approach across datasets and experimental conditions. We also demonstrate the importance of including these measures in fMRI analyses. This work highlights the importance of studying brain-body interactions, proposes a tool that may enhance the efficacy of fMRI as a biomarker, and provides widely applicable open-source software.

1 Introduction

The brain and body are closely coupled and continuously influencing each other. Brain-body interactions underpin key functions, including cognition and emotion, as well as the overall health of an organism (Barrett and Simmons, 2015; Shokri-Kojori et al., 2018; Azzalini et al., 2019; Koban et al., 2021).

Functional magnetic resonance imaging (fMRI) is a powerful and widely used technique in human brain research. While fMRI studies do not routinely incorporate physiological signals measured from the body, there is a growing trend toward acquiring continuous physiological measurements (such as heart rate and respiration) during fMRI scans. One motivation stems from considerations about interpretation and reproducibility of fMRI, a pressing issue in the neuroscience field (Botvinik-Nezer et al., 2020). Since fMRI is based on measurements of blood oxygenation, fMRI signals are influenced not only by spatially local changes in neural activity (Ogawa et al., 1992; Kwong et al., 1992; Bandettini et al., 1992), but also by any bodily physiological process that modulate blood oxygenation (Murphy et al., 2013). If not modeled, the presence of these additional effects can complicate the inferences drawn from fMRI as well as drive variability across results, such as in the mapping of large-scale brain networks (Birn et al., 2006; Xifra-Porxas et al., 2021).

Moreover, although physiological effects in fMRI are often regarded merely as confounds, several lines of work indicate that they also provide valuable information. For example, breathing and heart rate exert spatially structured, dynamic influences on cerebral blood oxygenation that closely mirror the spatial structure of core

neuronal networks, suggesting a close connection (and potential interactions) between the regulation of blood flow and neuronal responses (Chen et al., 2020; Bright et al., 2020). Indeed, blood-flow responses to heart rate variability have potential to strengthen neural connectivity within networks involved in emotion regulation, highlighting the bi-directional connection between brain and body (Nashiro et al., 2022). Further, removing physiological components of fMRI signals has been shown to reduce test-retest reliability of individual differences in functional connectivity (Dubois and Adolphs, 2016).

Two major physiological drivers of the fMRI signal arise from natural, slowly varying (<0.15 Hz) fluctuations in respiration volume (RV) and heart rate (HR). RV and HR are thought to influence fMRI signals through altering blood pressure, autonomic tone, and arterial carbon dioxide (Wise et al., 2004; Birn et al., 2006; Shmueli et al., 2007; Chang et al., 2009; Power et al., 2017; Tong et al., 2019; Picchioni et al., 2022). RV and HR variations are found to account for substantial variance in fMRI signals and to have a larger impact on brain functional connectivity measures compared to physiological effects that are directly synchronized with the breathing and cardiac cycles (Xifra-Porxas et al., 2021). Further, the influence of RV and HR substantially overlaps with neuronally mediated BOLD responses, spanning the same low frequency range (0.01-0.15 Hz) and overlapping with widely distributed functional networks (Birn et al., 2006). Therefore, the ability to precisely identify such physiological components of BOLD is crucial for their use as either 'noise' or valuable 'signal'.

However, it is not always possible to acquire clean external physiological measures during fMRI (Glasser et al., 2018), and many existing datasets lack such measures altogether (ADNI (Jack Jr et al., 2015); UK Biobank (Bycroft et al., 2018); HCP 7T Release (WU-Minn, 2017)). Accordingly, computational approaches have been developed for detecting RV and HR effects in the absence of physiological monitoring. Data-driven methods include global signal regression and ICA, yet currently, these methods cannot unambiguously identify RV and HR components of fMRI signals without the use of recorded physiological waveforms for reference. Recent proof-of-concept studies (Bayrak et al., 2020; Salas et al., 2021) indicated that the RV time course can be directly reconstructed from fMRI signals, but did not examine HR. Since HR has been associated with functional circuits for emotion regulation and correlates with dynamic variation in large-scale functional networks, the ability to reconstruct an HR signal directly from fMRI would contribute an additional rich source of information to fMRI studies. Data-driven techniques have been proposed to infer cardiac phase information directly from fMRI (Beall and Lowe, 2007; Ash et al., 2013; Aslan et al., 2019), from which low-frequency variation in heart rate can be derived. Here we aim to integrate HR estimation in the same framework as RV, with the goals of: (1) improving usability through a single end-to-end model, and (2) to leveraging known covariation between RV and HR to improve HR estimation.

Here, we develop a computational approach for inferring slow changes in respiratory volume and heart rate directly from the fMRI signal. Motivated by the inter-dependence of RV and HR, we propose to jointly learn these signals using a multi-task learning (MTL) architecture, and train models to simultaneously learn RV and HR signals. This work aims to enrich the information content of existing and prospective neuroimaging datasets with missing or corrupted physiological information. An earlier form of this work has been presented (Bayrak et al., 2021).

2 Results

In Section 2.1, we introduce the proposed *DeepPhysioRecon* framework. We then show the agreement between measured and decoded RV and HR signals (Section 2.2), and in Section 2.3, we demonstrate how *DeepPhysioRecon* can enable investigating these low-frequency physiological effects in the absence of measured respiration and cardiac data. In Section 2.4, we assess model generalizability and performance across datasets and experimental conditions. Section 2.5 and 2.6 explore the interpretability of our learning framework. Finally, in Section 2.7 we investigate *DeepPhysioRecon* as a potential tool for denoising RV and HR measures derived from corrupted physiological recordings.

2.1 DeepPhysioRecon Framework

We developed a generalizable deep learning framework to estimate RV and HR from fMRI data. We hypothesized that RV and HR share information that can foster their mutual learning, such that joint learning of RV and HR may enhance model accuracy and generalizability. The proposed network architecture (Figure 2.1) is composed of a bidirectional Long-Short Term Memory (bi-LSTM) block followed by two linear layers. An LSTM network allows for learning an explanatory feature space with temporal dependencies, while linear layers hone the focus on each physiological signal separately.

To reduce computational demands and improve signal-to-noise ratio, we employ atlases for dimensionality reduction. As input to the model, fMRI time-series are extracted from 4 different atlases constructed from several imaging modalities, and include cortical, subcortical, white matter, and ascending arousal network regions. The output of the linear layers contains the estimated RV and HR signals, which are of the same length as the input fMRI scans, and sampled at the same rate. The predicted signals are evaluated against the measured RV and HR signals using Pearson correlation.

2.2 Decoding Low-Frequency Peripheral Physiological Signals from Resting State Data

We first demonstrate the applicability of our joint learning approach using the publicly available Human Connectome Project (HCP) resting-state (rs) fMRI dataset. In the HCP rs-fMRI scans, subjects lie quietly with no

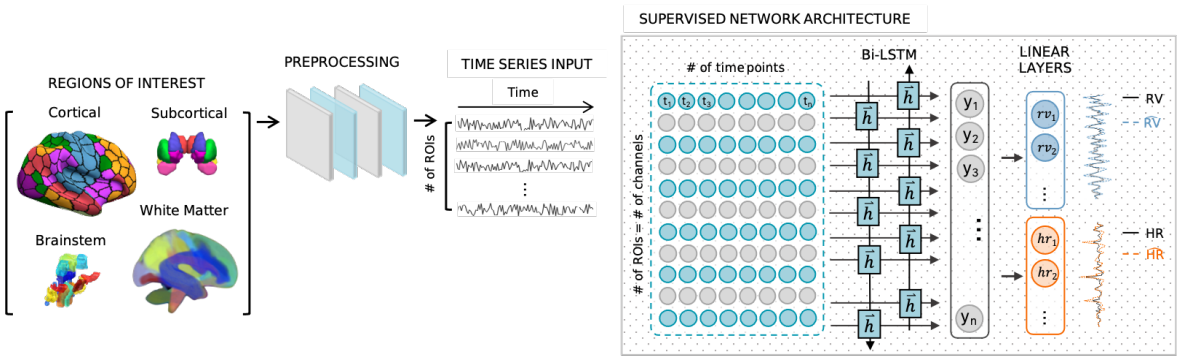


Figure 2.1: **DeepPhysioRecon Pipeline**. The pipeline for estimating respiration volume (RV) and heart rate (HR) signals from fMRI time-series dynamics is shown. Regions of interest are defined using 4 published atlases that had been constructed from different imaging modalities, comprising areas in cerebral cortex, white matter, subcortex, and the ascending arousal network. ROI time-series signals are extracted from the fMRI volumes, detrended, bandpass filtered and downsampled. The preprocessed signals are provided to a candidate network as input channels. A bidirectional LSTM network architecture is adapted for joint estimation. The output of linear layers are RV and HR signals.

externally imposed task or stimulation, and each scan lasted for $\sim$ 14.4 min. Cardiac and respiration data are continuously monitored throughout these scans. RV and HR are extracted as the temporal standard deviation of the raw respiration waveform, and the mean beat-to-beat interval, respectively, within a 6-sec sliding window centered on each fMRI time point. Further details of the fMRI and physiological signal preprocessing are described in Section 5.2.

Models were trained using preprocessed HCP rs-fMRI data from 375 subjects scanned 4 times (1500 scans). The resulting models are referred to as *rs-models*. Five-fold cross validation was used. Figure 2.2a shows the results for one example scan, where decoded signals aligned with the measured (ground truth) signals with a Pearson correlation of r = 0.884 for RV and r = 0.749 for HR. The resulting model accuracy for all folds shown in Figure 2.2b. The proposed framework reconstructs RV and HR signals on test cohorts with high agreement against those calculated from the measured physiological data, with a median r $\sim$ 0.689 for RV, and median r $\sim$ 0.627 for HR. To note, the 5 cross-validation folds produced models with highly similar performance on the test partitions (detailed in Section 5.4).

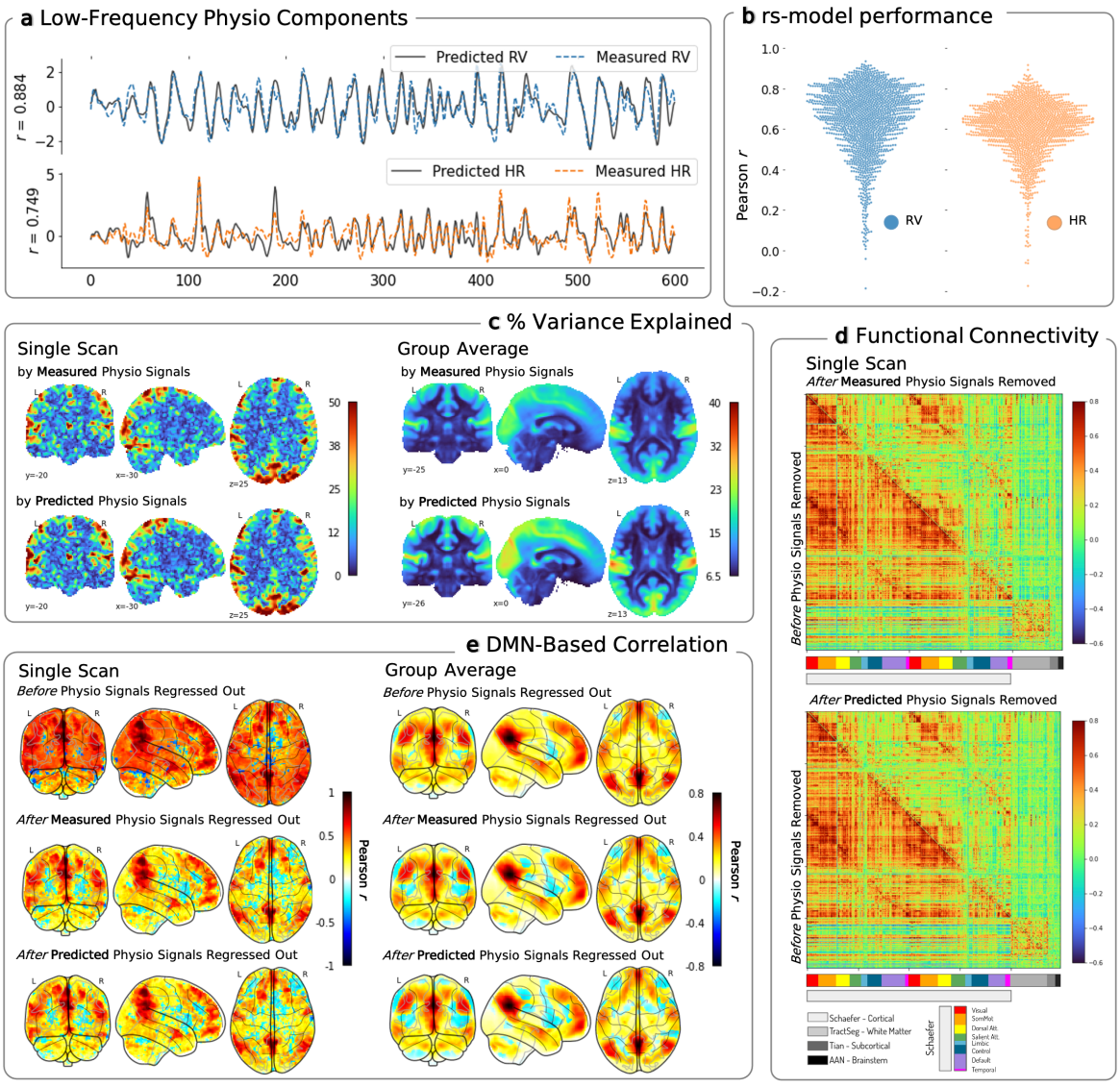


Figure 2.2: **Resting-state model performance on withheld test data.** (a) Low-frequency physiological signals predicted by our model are overlaid on the measured signals for one example scan. Accuracy is measured using Pearson correlation coefficient between measured and predicted signals. (b) Models are trained and evaluated with resting-state data under a 5-fold cross validation paradigm. Each marker represents the Pearson correlation (r) scores between measured and predicted signals for respiration variation (RV – shown in blue) and heart rate (HR – shown in orange), pooled across the withheld resting-state test sets. Single scan and group effects observed on resting-state fMRI data. **Single scan and group effects observed on resting-state fMRI data**. (c) Percent (%) variance explained maps shown for selected slices, indicate the percentage variance explained at each brain voxel by the measured and predicted physiological signals, for one example scan and averaged across the subject group. (d) Functional connectivity matrices highlight the change in ROI-to-ROI correlation after the indicated signals were removed from the ROI time-series data for one example scan. (e) DMN-based correlation maps at the voxel level show the seed-based correlation across the brain regions before and after the measured (as well as predicted) physiological signals are regressed out from the fMRI data, again for one example scan and averaged across the subject group.

2.3 Impact on fMRI signals and network connectivity

We next demonstrate how these decoded physiological signals can improve widely conducted fMRI analyses, and compare the results against analyses conducted using measured ("ground truth") physiological signals. Figure 2.2 provides examples of the impact of decoding RV and HR at the single-scan and the group levels.

To gauge the influence of RV and HR on fMRI, we first show the proportion of temporal variance explained by a linear combination of RV and HR in the fMRI signal at each voxel, across the whole brain. For this example subject, the variance explained by the predicted RV/HR signals reaches 50% (Figure 2.2c) – i.e., there are parts of the brain where half of the fMRI signal fluctuation is accounted for by RV/HR. When averaged across the entire set of 1500 scans, the variance explained by RV and HR signals across the brain ranged from 6.5 to 33% (Figure 2.2c); note that colorbar limits were set to a larger range to maximize image contrast.

We next assess the impact of these decoded physiological signals on functional connectivity analysis. In functional connectivity analysis, correlations between the fMRI time courses of different brain areas are investigated as a proxy for neuronal interactions (Van Den Heuvel and Pol, 2010; Noble et al., 2019; Zhang et al., 2021a), but non-neuronal influences (such as RV and HR) can obscure underlying connectivity patterns. When physiological signals are recorded, they are typically removed from the data to control for their influence on functional connectivity. If *DeepPhysioRecon* is successful, regressing out the decoded (versus measured) RV and HR signals should have a similar impact upon functional connectivity.

We therefore assess the functional connectivity between all 497 ROIs drawn from the aforementioned cortical, white matter, subcortical, and ascending arousal network atlases, both before and after projecting out the estimated physiological subspace. Further details of the projection are provided in Section S3.2.1. The Pearson correlation between signals from each pair of ROIs is used to construct a 497 x 497 symmetrical connectivity matrix. As a complementary analysis, we also map the whole-brain, voxel-wise connectivity with respect to a reference ("seed") region in the default mode network (DMN), both before and after removing the estimated physiological subspace. We observe that both the ROI-based (Figure 2.2d - Single Scan) and seed-based (Figure 2.2e - Single Scan) correlations are altered with the removal of physiological signals, and that comparable results are obtained using the measured and decoded physiological waveforms, as shown for an example scan. When averaged across all 1500 scans, regressing out the predicted and measured physiological signals had a more mild effect overall but also exerted similar effects on the DMN, enhancing negative correlations while sustaining positive correlations (Figure 2.2e - Group Average), as expected (Chang and Glover, 2009; Chai et al., 2012). Similar effects of measured and predicted physiological signals on functional connectivity were also obtained at the group level Figure (S2.2).

2.4 Generalizability across tasks and acquisitions

The above results indicate that RV and HR can be decoded from rs-fMRI data. To what degree do these models directly generalize to other fMRI paradigms or acquisition parameters? To test this question, we applied a randomly selected model from the 5-fold CV on resting-state data, without any additional training or fine-tuning, to data acquired during seven different task paradigms from the HCP dataset. Further, as an even stronger test of model generalizability, we also evaluated the performance of the HCP resting-state model, again without any additional training, on an external dataset that was acquired on a different scanner and with different acquisition parameters, including temporal resolution (results from models built from each of the 5 folds are reported in Figures S2.3 and S2.4). Further details about the datasets and preprocessing are described in Section 5.1, and 5.2.

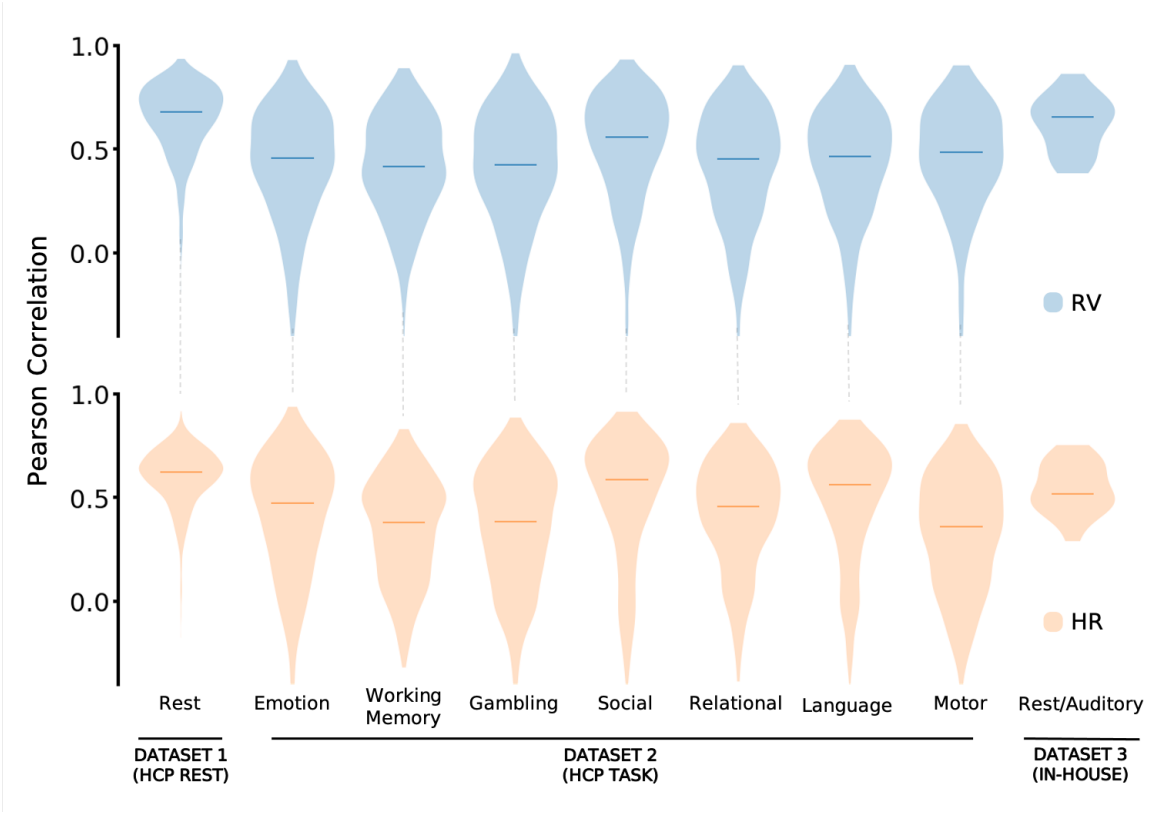


Figure 2.3: **Generalizability of rs-model to scans acquired under different experimental conditions and acquisition parameters.** From the five models trained on resting-state data under the 5-fold cross validation paradigm, a model is randomly selected for this assessment. Seven tasks from Dataset 2 (HCP: emotion, working memory, gambling, social, relational, language and motor), and Dataset 3 (in-house: rest/auditory) were used to assess generalization of the model. Each plot represents the aggregated Pearson correlation scores between measured and predicted signals for respiration variation (RV – shown in blue) and heart rate (HR – shown in orange) in the withheld test set. Median r is indicated by the horizontal line.

The results (see Figure 2.3) suggest that for the held-out datasets, the model trained on the HCP resting-

state data captures a transferable feature space. The agreement between measured and predicted signals for RV was high in the in-house dataset (median r $>$ 0.6), and moderate for all HCP tasks. For HR, the performance was also high for the in-house dataset. Moreover, HR had notably high performance on the HCP social and language tasks (median r $\sim$ 0.6), and a moderate agreement for the other HCP tasks.

Finally, we asked whether training models on task data could improve the performance on datasets involving tasks that are different from those on which the model is trained. This question also probes whether the relationship between fMRI and physiology is largely independent of task condition. To investigate this question, we preserve the same framework used for the rs-model (including hyperparameter values and model architecture) and train a new model, which we will refer to as the **task-model**, using only four of the seven tasks (working memory, social cognition, emotion processing, and gambling tasks). The task-model is then deployed to test generalizability on the three remaining tasks (relational processing, language, motor tasks), as well as on the HCP resting-state scans and the in-house dataset. The results show (see Figure 2.4) that overall, the task-model is able to learn both RV and HR with moderate agreement against ground truth (median r $\sim$ 0.5). When tested on the HCP tasks, the task-model showed an improvement in performance compared to the rs-model (Table S2.5).

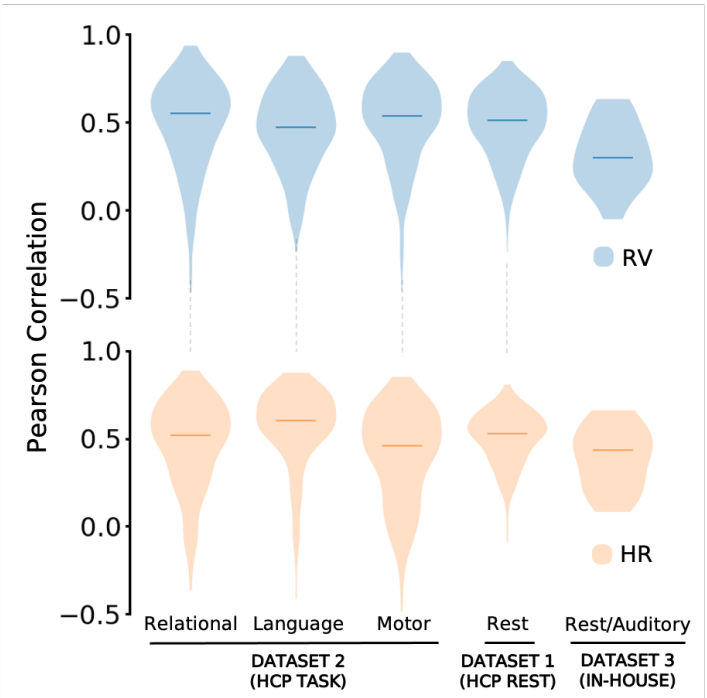


Figure 2.4: **Generalizability of task-model on test cohorts acquired with different task conditions and acquisition parameters.** Three tasks from Dataset 2 (HCP: relational, language and motor), Dataset 1 (HCP: resting-state) and Dataset 3 (in-house: rest/auditory) were used to assess generalization of a model trained on data consisting of 4 other task conditions. Each plot represents the aggregated Pearson correlation scores between measured and predicted signals for all scans in a set. Median r is indicated by the horizontal line.

2.5 Spatially constraining what the network sees

Our learning framework can flexibly accommodate any set of atlas regions for dimensionality reduction. Although we chose a large number of brain regions when fitting the original model in order to maximize brain coverage, we hypothesized that certain brain regions convey more physiological information than others. For example, regions near major blood vessels may be more strongly influenced by physiology, and thus provide greater predictive value. To investigate this hypothesis, three linked experiments are designed. In the first, we use only one input ROI times-series signal at a time, generating what we will refer to as **individual ROI models**. These models are trained in a manner identical to that of the original (multi-ROI) model. The model performance for each scan is assessed using Pearson correlation between the predicted and measured physiological signals, and an average score for the entire test cohort is noted (see Section 5.4). Figure 2.5a shows these average scores mapped onto their corresponding regions for all 497 regions, representing the predictive ability of each individual ROI. The best performance of individual ROI models reaches r $\sim$ 0.5 (note that colorbar limits were set to a smaller range to maximize image contrast).

Next, we assess whether the number of regions included in training affected the model performance. Here, models are trained using successively increasing numbers of ROIs (from 1% to 100%), rank-ordered by the results of the individual ROI analysis. These models will be referred to as **percent ROI models**. While the improvement in model performance continuously increased with the number of included ROIs (Figure 2.5b, i.e. for RV median Pearson r $\sim$ 0.45 and r $\sim$ 0.65 respectively with 1% and 25% of ROIs), the rate of increase slows after the model includes approximately 25% of the top ROIs, and reaches median Pearson r $\sim$ 0.7 with 100% of the regions.

Given that brainstem and cerebellar regions are often absent from fMRI acquisitions, and yet are implicated in cardiovascular regulation, the last experiment assesses the performance when using all regions except those that spatially overlap with brainstem and cerebellum. Excluding brainstem and cerebellar regions resulted in comparable performance (Figure 2.5c).

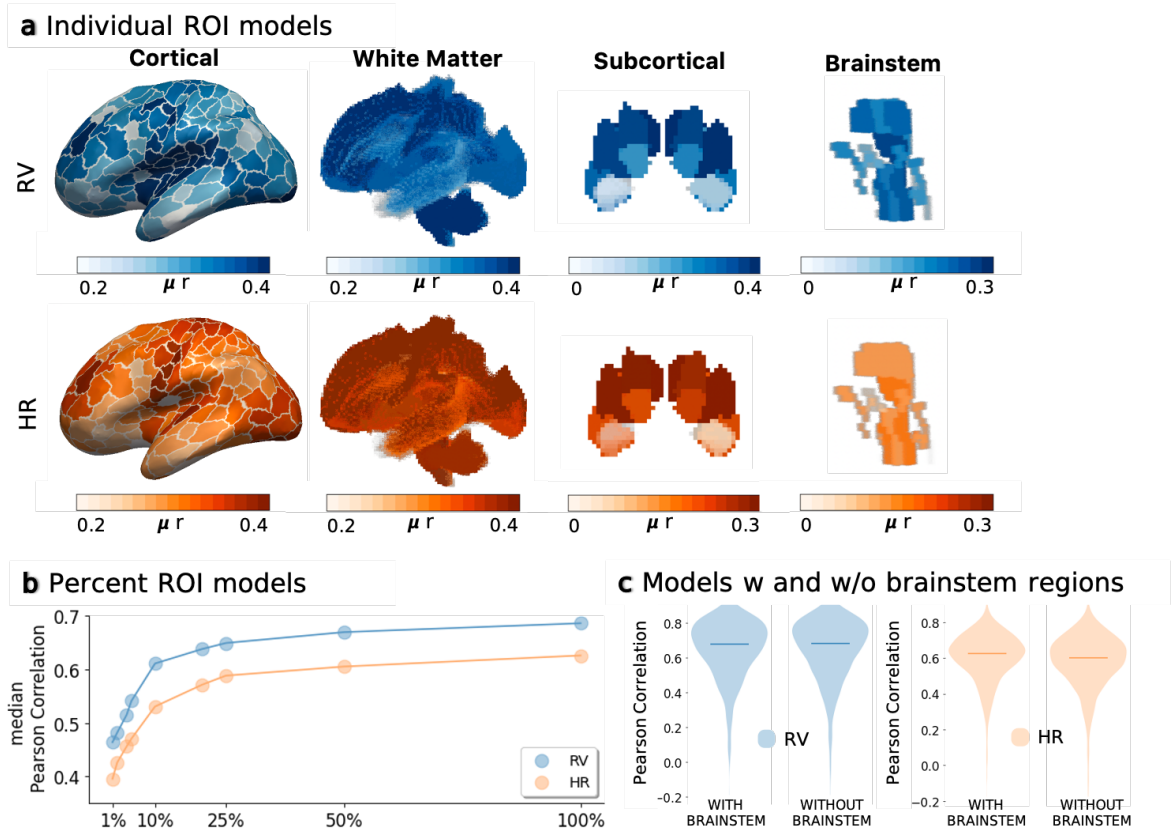


Figure 2.5: (a) **Individual ROI models**. Using resting-state data, models were trained separately for each individual ROI. Predictiveness of a given ROI was assessed using Pearson correlation between measured and predicted signals. Results were visualized by projecting the mean r onto (left to right) cerebral cortex (400 ROIs), white matter bundle regions (72 ROIs), subcortical regions (16 ROIs) and ascending arousal network (9 ROIs). (b) **Percent ROI models**. A set of models are trained with an increasing number of ROIs. Starting from the top 1% of the total ROIs, rank-ordered by results of the individual ROI analysis, models are trained using successively increasing numbers of ROIs, up to 100% of ROIs. (c) **Model comparison without brainstem and cerebellar regions**. A model is trained using 477 ROIs, excluding regions that spatially overlap with brainstem and cerebellum. Each plot represents the Pearson correlation scores between measured and predicted signals for all scans in the unseen test data.

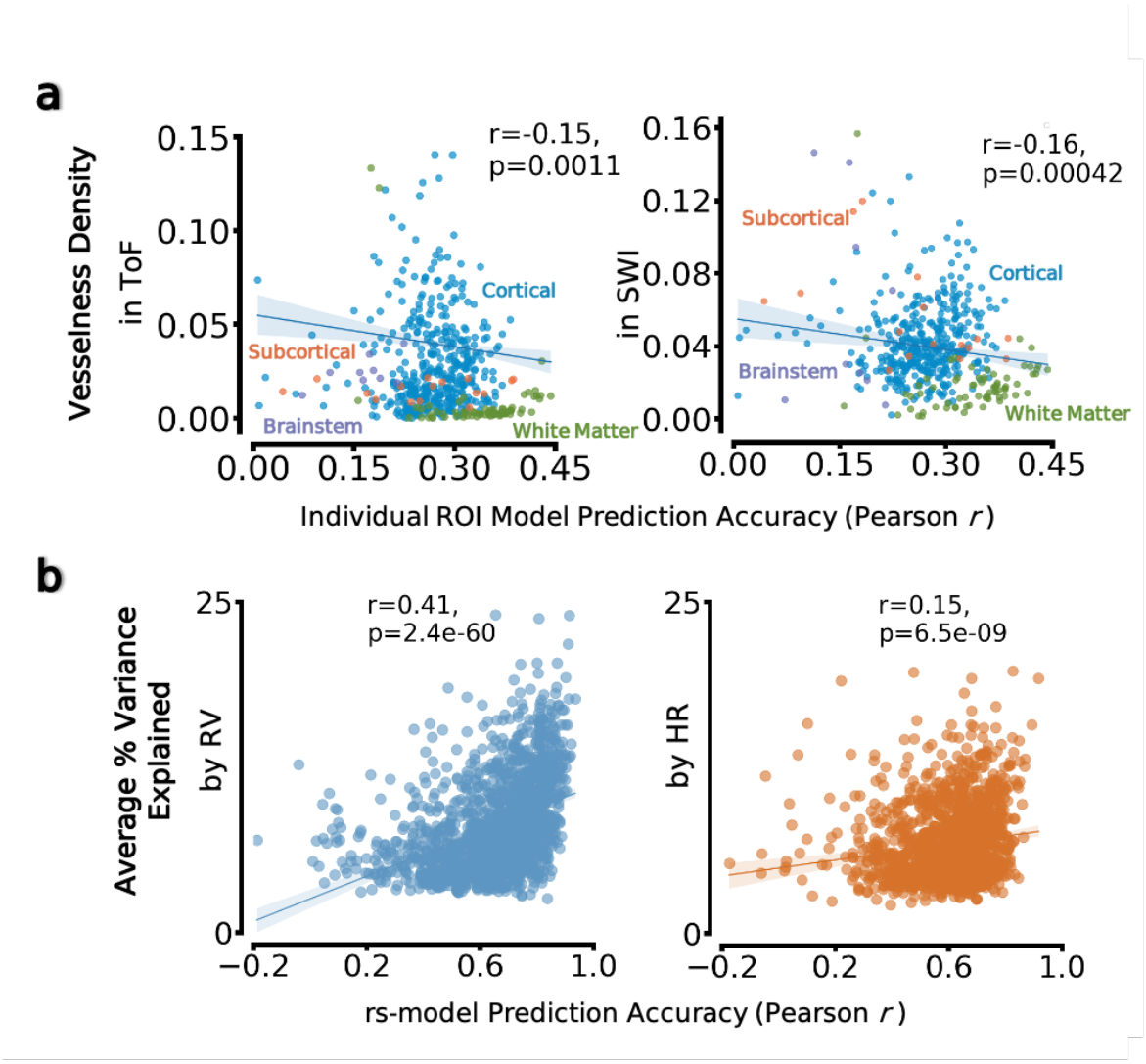


Figure 2.6: (a) **Vessel density as a factor that may underlie regional prediction accuracy.** In each plot, the x-axis represents model accuracy for each individual-ROI model. The y-axis represents the average vessel density for Time-of-Flight (ToF - shown in green) and Susceptibility Weighting Imaging (SWI - shown in purple) measures, averaged across all voxels within each ROI. (b) **Percent variance accounted for by physiology in fMRI data, as a factor that may underlie model prediction accuracy.** In each plot, the x-axis represents cross-validated model accuracy calculated using Pearson correlation scores between measured and predicted signals for respiration variation (RV – shown in blue) and heart rate (HR – shown in orange) in the withheld resting-state test set. The y-axis represents the percentage of temporal variance explained in the fMRI data, averaged across all brain regions, by the respective (measured) physiological signal.

2.6 Factors driving model performance

We further investigated the various factors that may drive or affect model performance. For this, we first examine whether the results are driven by BOLD signal in regions near blood vessels. To investigate this relationship

between the ”vesselness” of a region and its relative predictive ability, we carried out linear regression analysis between Time of Flight (ToF) and Susceptibility Weighting Imaging (SWI) measurements of vessel density (Bernier et al., 2018) in an ROI and the corresponding individual ROI model prediction accuracy. The results are shown in (Figure 2.6a). Interestingly, the vessel density metric is negatively associated with predictive value, showing that the higher the vessel density, the less accurate the model predictions from the fMRI signal within that region.

Another candidate factor could relate to the magnitude of RV and HR fluctuations exhibited by the subject in a given scan, as larger physiological variations would modulate brain hemodynamics more strongly. To investigate the relationship between the magnitude of fMRI physiological responses and model accuracy, we carried out a linear regression analysis between the amount of fMRI temporal variance explained in a scan (averaged across all voxels in the brain) and the model prediction accuracy. The results (Figure 2.6b) indeed indicate a positive correlation between the two, suggesting that the amount of physiologically induced BOLD variation during a scan impacts the ability to predict RV/HR from an fMRI scan.

2.7 Potential for denoising physiological waveforms

Low correlation scores can emerge when there are artifacts in the measured physiological recordings, coloring the performance during testing and therefore necessitating close inspection. We speculate that the signals predicted from *DeepPhysioRecon* may have the potential to ‘fix’ some of the artifacts in the measured physiological data. In Figure 2.7, we provide one example in which possible motion-induced artifacts can be observed in the raw cardiac signal, whose effects propagate to the ”ground truth” low-frequency HR signal. We observe that the predicted HR smooths over these artifacts, suggesting that the model may be able to clean up RV/HR measures through this data-driven learning process. Another observation is that in some cases wherein the predicted RV/HR signals exhibited low correlation with the ground-truth RV/HR signals, the predicted signals explained a larger variance in the fMRI data compared to the measured physiological signals. Two such examples are shown in Figure 2.8 (Subject A, Subject B). Notably, although having stronger magnitudes, the spatial maps associated with the predicted physiological signals retain subject-specific patterns that are also present in the maps associated with the measured physiological signals. These examples may also indicate that the model can generate physiological estimates that are cleaner than the recordings. In cases where the measured physiological signals appear to be clean and have high correlation with the predicted waveforms (Subject C), close correspondence is also seen in their effects on fMRI data.

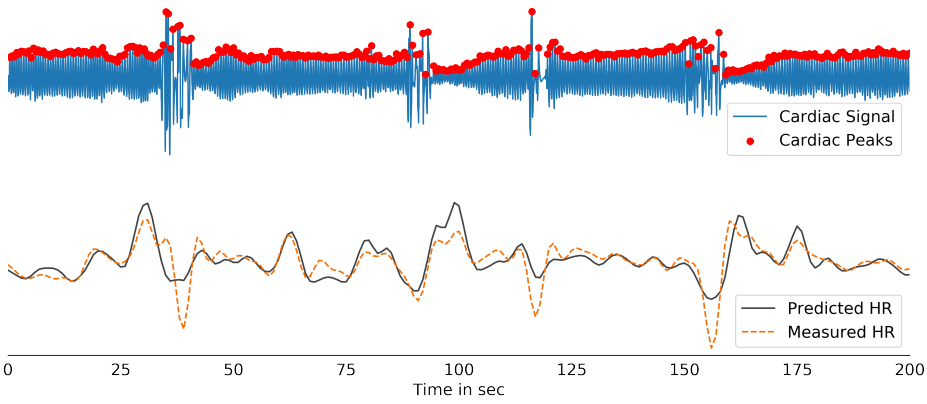
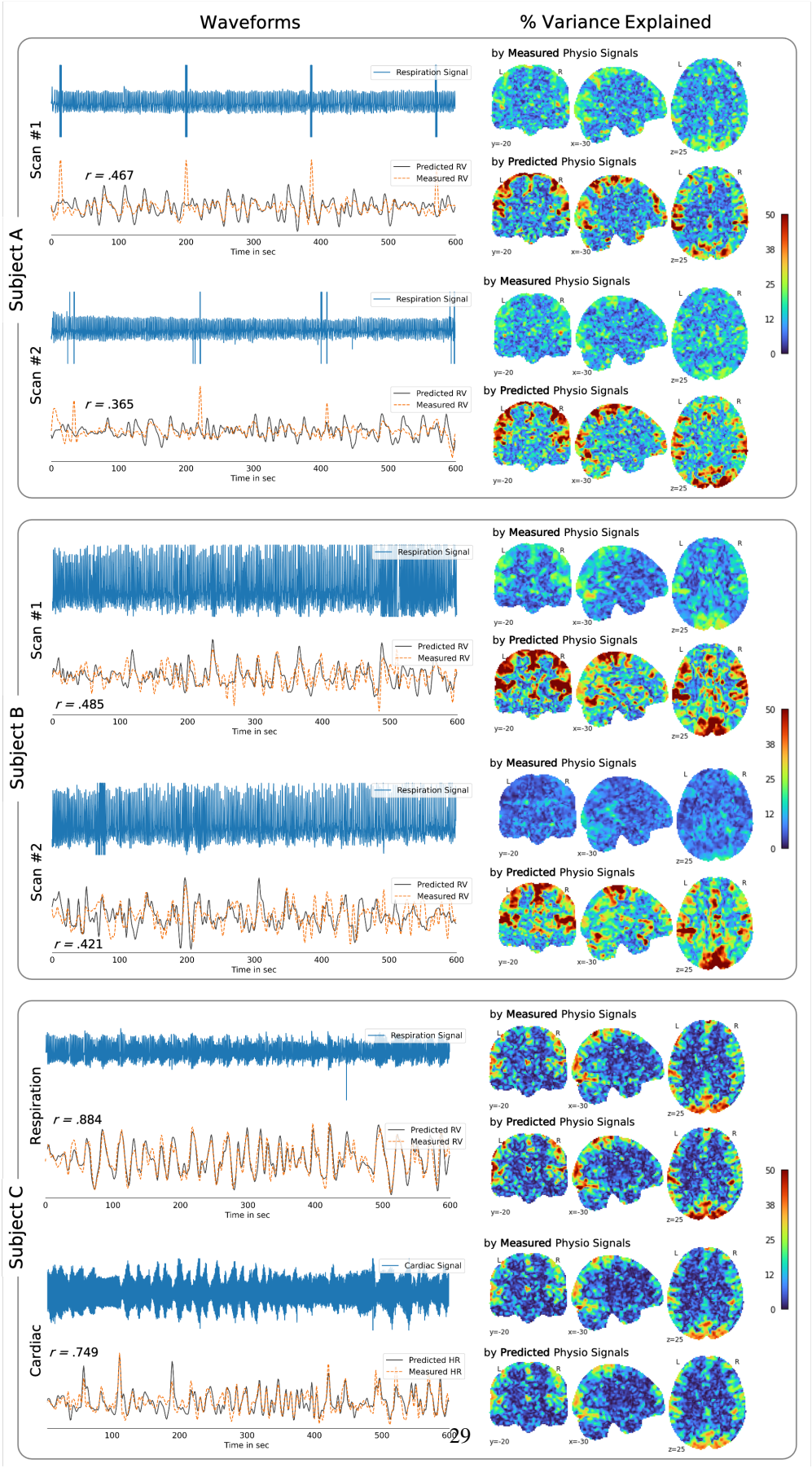


Figure 2.7: An example recording is shown. The transient artifacts in the HR waveform (bottom; orange) are likely due to motion in the raw PPG signal (top). DeepPhysioRecon may help to 'fix' noisy physiological recordings, as indicated by the predicted HR signal (bottom; black).

3 Discussion

This work presents a framework for jointly inferring respiration and heart rate fluctuations directly from fMRI dynamics. The *DeepPhysioRecon* framework provides these two key low-frequency physiological signals (RV, HR) to datasets that fully or partially lack external physiological measures, or which lack measurements of sufficient quality. We demonstrate that the proposed models of RV and HR, trained on resting-state fMRI data, generalize across datasets with varying experimental conditions and significant acquisition differences. Consistent with the literature (Birn et al., 2006; Chang et al., 2009; Power et al., 2017; Kassinopoulos and Mitsis, 2019), we also find that RV and HR can, in certain scans, account for large amounts of temporal variance in fMRI signals. By including the predicted RV and HR signals in functional connectivity analyses, we also show that large-scale network maps are altered in agreement with the changes introduced by incorporating the measured RV and HR signals. The proposed models are tested not only in independent participants within each dataset, but also across datasets acquired with different experimental conditions, MR scanners, and protocols.

Models trained and tested on resting-state fMRI data succeeded in decoding RV and HR signals with high accuracy. These models were also found to generalize well between experimental conditions, different subjects, and even across significantly different acquisition parameters, including different temporal resolutions (TR=0.72s to 2.1s). Figure 2.3 indicates that models that were directly applied to different fMRI conditions (i.e. resting-state models applied to task, and task models applied to resting-state or unseen task conditions), could infer information in a new set of experimental conditions. This ability to transfer models between fMRI conditions suggests that the relationship between fMRI and physiology could be largely brain state-invariant. However, as expected, models trained using task fMRI data exhibited improvements on decoding task data in



(Caption on next page.)

Figure 2.8: Low correlation between measured ('ground truth') and predicted RV signals may not accurately reflect the quality of reconstruction. In the selected examples, artifacts are observed in the raw time-series signals. For example, data from Subject A is shown to have periodic artifacts likely related to the equipment, and data from Subject B exhibits clipping artifacts. These are then carried over to the derived low-frequency RV and HR waveforms. We observe that the predicted waveforms appear to mitigate some of these artifacts, suggesting that the model may be able to clean up physiological signal measures through this data-driven learning process. The percent variance explained maps indicate that the predicted RV signals accounted for a much larger proportion of the fMRI signal variations compared to the measured RV signals. On the contrary, in the example of Subject C, the respiration and cardiac recordings from the same scan (without any major artifacts) are shown. The variance explained maps indicate that measured and predicted RV signals accounted not only for a similar proportion of the fMRI signal variations but spatial distribution of these maps closely aligned.

comparison to those trained on resting-state data (Figure 2.4; Table S2.5), though the degree of performance improvement varied across tasks. Likewise, models trained on resting-state data performed better on held-out resting-state data than they did on task data.

One possible explanation for this increase in accuracy is that within a particular condition (task vs. rest), it is expected that the distribution (signal and/or noise) matches more closely. For the case of task, there may be task-related brain activity that shows up in more systematic ways both in the fMRI data and physiological recordings. Therefore, we can speculate that the model learns to isolate these patterns when trained with task data. Further, the success of rs-models to more accurately decode an external dataset with different acquisition parameters, compared to the t-model, may also relate to the conditions used in the external (in-house) dataset. The in-house dataset tasks involved only simple sensory stimulation (auditory tones) that were presented at long-inter-stimulus intervals (ISI) for some scans, or at very short ISI but continuously throughout the experiment for others. Both of these designs will have more contribution from spontaneous fluctuations, resembling the resting-state conditions.

Nonetheless, we found that even when training and testing on the respective conditions, model performance was higher on the rs-fMRI data than for task fMRI data. One explanation may be that the resting-state scans were longer than the task scans. It has been shown that increasing the scan duration may increase the reliability of fMRI connectivity estimates (Birn et al., 2014), so perhaps training models with long duration ($\sim$ 15 min) scans (as in the HCP resting-state data) may offer advantages over training with the shorter-duration (2-5 min) scans.

We suggest that depending on the goal of the study, different training datasets should be considered. If the goal is to create the data in its absence (i.e., to reconstruct physiological signals in a dataset that does not include any physiological measurements), a large cohort including variety of fMRI conditions and acquisition parameters may give a more robust model. Training models using the same task as the one for which predic-

tion is needed can further increase accuracy, as the models may be trained using the same dataset and same condition. While this will reduce the generalizability of a model across conditions, it may allow for better condition-specific reconstruction as well as for 'fixing' corrupted data within a dataset (i.e., cases in which the dataset has physiological measurements but contains corrupted samples). However in the latter case, we would extra caution against overfitting to the data, which can be achieved in various ways. For instance, initializing the weights by a model pretrained with large-scale fMRI data across a broad set of conditions can provide a good starting point.

Our neural network architecture was designed to handle many unique challenges related to fMRI. First, bi-directional LSTM networks support varying-length input and multivariate output. This allows for learning on varying-length input signals and for predicting RV and HR simultaneously. Given that fluctuations in respiration and heart rate evoke delayed responses in the BOLD signal (i.e., the BOLD effects occur asynchronously and persist beyond the duration of the physiological response) (Birn et al., 2008; Chang et al., 2009), a neural network that takes past and future information into account – such as bi-LSTM – means that no hemodynamic lag needs to be explicitly considered. Further, the ability to integrate over multiple time-scales may offer benefits over fixed-windowed methods (Salas et al., 2021).

Given that RV and HR are strongly coupled, we had hypothesized that learning the features of one (i.e. RV) could boost the accuracy of predicting the other (i.e. HR) even with a few learning instances. However, experiments with pre-training a network based on RV and fine-tuning these weights to predict HR (compared to training with random weights), see Section S2.1: Handling Missing Data During Training, resulted in only a small performance increase, and further experiments may be warranted. Commonly, transfer learning is performed by freezing the early layers (high-level features) and training the remaining layers (low-level features) to specialize on another task. Since our network architecture comprises only two levels, a bi-LSTM layer and linear layers, fine-tuning only the LSTM layer may not be enough to transfer the knowledge between RV and HR and needs further investigation with different architectures. This direction is of particular interest given that the measured HR signal is often noisy.

Here, we adopted a large dataset to train and evaluate our models. Although an initial quality-check was performed on physiological recordings, visual inspection indicated that many still contained artifacts such as clipping and imperfect heart-beat detection. However, substantial improvement in performance beyond recent published work (from medians of approximately $r \sim 0.5$ (Bayrak et al., 2020; Salas et al., 2021) to $r \sim 0.7$ for RV in the current study) was attained. (Of note, fMRI data has low signal-to-noise ratio. Given that meaningful fluctuations comprise about 1-4% in the data, correlations at and above 0.5-0.6 between fMRI and an external measure are considered to be relatively strong in the neuroscience field.) In the previous studies, we

used a small, visually inspected subset of the HCP data consisting only of RV signals. The aforementioned improvement suggests that a larger, albeit less strictly vetted, dataset was effective for learning physiological patterns. Nonetheless, while a wide variety of signals (including some low-quality examples) may be advantageous during training, low-quality physiological data colors measurements of performance during testing, as these physiological recordings are taken as ground truth. In other words, low correlation scores can emerge when there are artifacts in the physiological recordings, necessitating close inspection as in Figure 2.8.

White matter regions in the brainstem were shown to be very predictive of breathing as well as heart rate. While much remains to explore in terms of regional predictiveness of physiological signals beyond the current heuristic approach, the strong performance of tractography-based ROIs suggest that incorporating parcellation boundaries derived from other imaging modalities (e.g., diffusion-weighted MRI) and non-gray-matter regions could benefit fMRI prediction tasks. Further, results of the percent ROI models (Section 2.5) demonstrate that performance could be boosted by increasing the number of ROIs that are included in the training. While this finding supports the need for comprehensive atlases with finer subdivisions, the exponentially decaying projection in Figure 2.5b suggests that dimensionality reduction using atlases sufficiently captures low-frequency physiological information.

The present work is directly inspired by reports showing that low-frequency physiological processes can impact fMRI analyses (Xifra-Porxas et al., 2021; Power et al., 2017; Birn et al., 2006; Wise et al., 2004). Consistent with prior work, we found that modeling physiological effects could alter maps of functional connectivity, and that RV and HR signals can account for substantial temporal variation in certain brain regions. Previous studies have also found that in task conditions, changes in cerebral oxygenation, captured in part by respiration and heart rate signals, may modulate the magnitude of observed task activation responses (Bright et al., 2020). These findings suggest that accounting for RV and HR signals can advance individual-level precision in neuroscience and medicine using fMRI. As such, it is important to have general, versatile models that can reconstruct these features in the common scenario of missing or noisy physiological recordings.

Notably, in addition to its use for 'denoising' - i.e., removing RV and HR effects from fMRI signals, the proposed method may also open possibilities for studying *neural* processes relating to autonomic regulation in data where physiological signals have not been monitored. Specifically, we may expect that a component of the reconstructed RV and HR signals may track large-scale neural activity that is linked with physiological modulations (e.g., (Tu and Zhang, 2022)). Based on our experiments with individual ROI models, we observed that cortical regions with the highest individual predictive power tended to span somatomotor, salience, and ventral attention regions, which have known connections with autonomic regulation (Mckay et al., 2003; Beissner et al., 2013). While these autonomic signals may not be of interest for some studies, they may inform work on

brain-body interactions.

The current results suggest a number of future directions. One springs from limitations in the use of population-level atlases for defining fMRI parcellations. The present study used atlases that had been derived from populations, yet individuals can exhibit variability in the spatial boundaries of functional regions (Kong et al., 2021). Thus, individual- or cohort-specific parcellations may enable higher precision in extracting personalized physiological information from fMRI. The issue of precision is amplified when working with datasets from different stages of human life, including infant or aging datasets, which do not conform with the topographical features of the young cohort that is used to train networks in this work. Another limitation that can be addressed in future work is model interpretability, a common problem with machine learning models. Methods such as Shapley Values (Ghorbani and Zou, 2019) can offer explanations for model predictions and provide insights into the model development process. Architectural modularity (multiple layers) could be considered to allow transfer learning (Chollet, 2021), since the ability to freeze layers that capture high-level information may further boost the performance of joint learning.

4 Conclusions

Modeling physiological variability is increasingly recognized as crucial for both (1) improving the sensitivity of fMRI to neural effects, as well as (2) providing valuable information about cerebrovascular health, brain states, and emotion regulation. Since high-quality physiological signals are often missing from fMRI datasets, the present study fills this gap by introducing a generalizable tool for reconstructing two key physiological signals (RV, HR) from fMRI data. The proposed framework was found to be robust across a broad range of experimental conditions and imaging protocols, indicating that it can enrich a broad array of fMRI datasets with missing physiological information. This study motivates future work on methodological advances for modeling RV/HR, and enables retrospective studies of physiological effects in health and disease, leveraging the large body of fMRI databases that have been acquired without physiological signals.

5 Methods

5.1 Datasets

Models were trained on the Human Connectome Project (HCP) data (Van Essen et al., 2013). Our models can be divided into two categories, resting state models (rs-models) and task models (t-model). The rs-models were trained using HCP resting state data using 5-fold cross-validation. To evaluate the generalizability of rs-models across subjects both within (resting state) and between (task vs. resting state) conditions, rs-models were tested on both HCP task data as well as a separate, in-house dataset with different acquisition parameters.

Similarly, the t-model was trained on HCP task datasets, and tested on held-out HCP task data (with different task conditions), HCP resting state data, and the in-house dataset. The individual datasets and acquisition parameters are described below. For more details, please refer to the published articles (WU-Minn, 2017; Goodale et al., 2021).

5.1.1 Dataset 1

A set of resting-state fMRI (rs-fMRI) scans was drawn from the publicly available HCP 1200 subject release. fMRI scans in this release were acquired using a simultaneous multi-slice EPI sequence with the following parameters: TR = 0.72 s, duration of 14.33 mins, voxel size of 2 mm isotropic, TE = 33.1 ms, multi-band factor = 8, flip angle = 52 deg and 72 slices. During the resting-state scans, subjects were instructed to keep their eyes open and fixate on a cross-hair, and subjects underwent 4 scans (two runs on one day, and two runs on a second day). A subset of the HCP rs-fMRI dataset, consisting only of those subjects whose physiological signals were reported to have passed a quality assessment in both Power et al. (Power et al., 2020) and Xifra-Porxas et al. (Xifra-Porxas et al., 2021), was included in this study. This procedure resulted in N = 375 subjects (with all 4 runs, totalling 1500 scans).

5.1.2 Dataset 2

Our study also drew upon task fMRI (t-fMRI) scans from the HCP 1200 subject release. These scans were acquired with the same parameters as the rs-fMRI data, described above, except for the duration of the scans. The task dataset comprises seven tasks: Working Memory, Social Cognition, Emotion Processing, Gambling, Relational Processing, Motor, Language. These tasks were designed to activate a variety of brain networks, and each subject underwent 2 scans per task (both runs on the same day). signals was one that had excluded that subject from the All scans from t-fMRI datasets were initially considered. A quality assessment was employed, using automated selection criteria to remove scans with poor-quality physiological data (this script will be found here upon acceptance: http://github.com/deep-physio-recon/qa). Briefly, these criteria checked for clipping of waveforms (values clamped at 0 and 4095), unrealistic heart rates (mean heart rates below 30 bpm or above 97 bpm, or constant at 48 bpm (Xifra-Porxas et al., 2021), and missing waveforms. A subset of fMRI scans for each task, consisting of only those scans whose physiological data were labeled as 'clean' by the above criteria, was included in this study. For detailed information about task fMRI scans, please see Table S2.2.

5.1.3 Dataset 3

A different, in-house dataset was used as an external validation dataset. fMRI data were acquired with a multi-echo, gradient-echo EPI sequence, with the following parameters: TR = 2.1 s, duration of 24.5 mins, voxel size of 3 mm isotropic, echo times of [13.0, 29.4, 47.5 ms], flip angle = 75 deg, 30 axial slices (for detailed acquisition information see (Goodale et al., 2021)). The scans were acquired under 5 different conditions: cued deep breaths, eyes-open auditory task, eyes-closed auditory task, passive eyes-closed auditory task (i.e. no button press) and eyes-closed rest (no stimuli presented). In the auditory tasks, auditory tones were delivered at long inter-stimulus intervals, and subjects were instructed to press a button as quickly as possible. We selected a random subset of these scans that additionally underwent visual inspection to have clean physiological data. This yielded a total of 23 scans (drawn from 11 subjects).

5.2 Preprocessing

5.2.1 fMRI

Both resting-state and task fMRI scans had undergone the HCP minimal preprocessing pipeline (Glasser et al., 2013). Beyond this, we applied linear and quadratic detrending to remove slow scanner drifts. This was followed by band-pass filtering, within the low-frequency range of (0.01 - 0.15 Hz) and temporal downsampling by a factor of 2 (which does not result in further information loss, since the Nyquist criterion is satisfied). The HCP minimal preprocessing pipeline included motion coregistration, and while an additional step of regressing out head motion parameters is an integral part of many fMRI preprocessing pipelines, we did not regress out head motion parameters when training our models since they may carry predictive information about physiological fluctuations. For example, recent evidence has shown that apparent head motions can be introduced by respiration (Power et al., 2019), and we have previously found that retaining head motion slightly improved the reconstruction of respiration variation signal (Salas et al., 2021). Therefore, head motion was retained in the training data, to preserve information relevant to predicting RV and HR. The in-house data were preprocessed according to a multi-echo ICA pipeline described in (Goodale et al., 2021). All fMRI scans were already spatially aligned to the common MNI152 space.

5.2.2 Physiological Recordings

From the pulse oximetry signal, heart rate (HR) was extracted as the inverse of the mean inter-beat-interval in sliding windows of 6 seconds centered at each fMRI time frame (TR). Likewise, the respiration variation (RV) signal was calculated as the temporal standard deviation of the raw respiration waveform in a window of 6 sec centered at each TR (Chang et al., 2009). Both RV and HR were then band-pass filtered (0.01 – 0.15 Hz) and

were resampled to temporal resolution of 1.44 seconds, in a manner identical to the HCP fMRI data, the fMRI data that is used to train the models.

5.2.3 Normalization

fMRI signals and respiration belt data were both acquired with arbitrary units and may carry scan- and subject-specific amplitude differences. Unless stated otherwise, all time-series signals, including heart rate, were temporally normalized to zero mean and unit variance.

5.2.4 Dimensionality Reduction

When training a neural network on voxelwise, 4D whole-brain data, downsampling or patch/window-based implementations are typically required due to GPU memory limitations. Extracting time courses that are averaged within functional or anatomically defined regions of interest (ROIs) enables computationally efficient modeling. The use of ROI time courses also renders the approach less sensitive to the spatial resolution of the acquired fMRI data. Therefore, here we carried out dimensionality reduction by parcellating the brain into regions of interest based on four published atlases. Of note, our learning framework can flexibly accommodate any set of atlas regions for dimensionality reduction. The four published atlases used here were derived from multiple imaging modalities: a cerebral cortex atlas that was derived from rs-fMRI data provided 400 cortical regions embedded within 7 larger functional networks (Schaefer et al., 2018); the Pandora TractSeg white matter atlas, which was derived from diffusion MRI data and included 72 uni-/bi-lateral white matter regions (Hansen et al., 2021); the Melbourne subcortex atlas, which provides a multi-modal segregation of 16 subcortical regions (Tian et al., 2020); and an ascending arousal network (AAN) atlas that includes 9 regions located in the brainstem (Edlow et al., 2012). Dimensionality and noise were reduced by extracting the mean fMRI time series from all voxels within each ROI.

All four atlases were already registered to MNI152 space, and when necessary, were also resampled to 2 mm isotropic voxels to match the resolution of the preprocessed fMRI scans. While the Pandora white matter atlases (Hansen et al., 2021) includes various options, a probabilistic atlas created using TractSeg method (Wasserthal et al., 2018) on the HCP cohort was utilized in this study. We thresholded the probabilistic atlas (at 95%) to exclude voxels with lower confidence and minimize the overlap between white matter and gray matter regions. The 7 Tesla HCP Scale I atlas from Melbourne subcortex atlases (Tian et al., 2020) was selected, with the assumption that 7T may provide better spatial precision and the Scale I granularity could support generalizability to external cohorts.

5.3 Network Architecture and Implementation Details

Deep neural networks have shown remarkable success for image and time-series data (Sharma and Singh, 2017), including in the field of fMRI. The success of these networks stems from their ability to find a non-linear representation of the data and to make meaningful connections between spatial and temporal information implicit in the data. The proposed framework consists of a bidirectional LSTM (bi-LSTM) network followed by dropout and two linear layers. Long short term memory networks (LSTM) were selected as a candidate approach based on their capacity to automatically learn the temporal dependencies present in time series and their capability of operating on data of varying lengths (Van Houdt et al., 2020). Linear layers are commonly used for inferring an objective-specific feature space, and here they were used to infer unique RV and HR estimates from LSTM hidden units. The framework is illustrated in Figure 2.1. The input to the networks consist of ROI time-series signals provided as different channels (# of channels = # of ROIs). The bi-LSTM network is followed by a dropout layer and is connected to two linear layers, which output the estimated RV and HR time series. These estimated signals are the same length as the corresponding input ROI time series.

We conducted a grid search to select hyperparameters for the resting-state models. The size of hidden states was selected from values of [32, 64, 512, 1000, 1024, 2000, 2048], batch size from values of [1, 2, 8, 16, 32, 64], for dropout rates of [0.1, 0.3, 0.5, 0.6], for learning rates from [1.0e-2, 1.0e-3, 1.0e-4, 1.0e-5] and for the decaying learning rate with decay rate of [0.01, 0.1, 0.2, 0.5].

All models were then trained with the following, empirically chosen hyperparameters: a hidden state (h) size 2000 (i.e., depth), a batch size of 16, dropout rate 0.3 and were trained with decaying learning rate (lr) of 1.0e-3 with patience 2 and decay rate of 0.5, saving only the best models according to validation performance. The models were trained using ADAM optimizer with default parameters. The experiments were performed on an NVIDIA RTX 2080Ti GPU. Programs were implemented with Python using the Pytorch deep learning library.

5.4 Model training and evaluation

Resting state (rs-) models. Only for the model training on the HCP resting-state dataset (Datasets: Dataset 1), we used 5-fold cross validation, with 68% for training and 12% for validation and 20% for testing. A subset of 375 subjects (1500 scans) were used for training by rotating the partitions (Figure S2.5; left), and the resulting performance (pooled over the 5 testing partitions) is reported. In addition, to evaluate the generalizability across subjects both within (resting state) and between (task vs. resting state) conditions, rs-models were tested on HCP task (Dataset 2) and in-house (Dataset 3) datasets. Since the 5-fold CV on the rs-fMRI data resulted in 5 different models, one of these models was randomly selected to be applied to the task and in-house datasets

(Figure S2.5; right). For detailed information about number of scans used for training and testing, please see Table S2.3.

Training task (t-fMRI) model. t-fMRI model is trained using HCP task fMRI dataset (Dataset 2). More specifically, models were trained using 4 tasks (working memory, social cognition, emotional processing, and gambling tasks), and assessed using the remaining 3 tasks (relational processing, motor and language tasks). Of note, the subjects that were included in the training t-fMRI model were excluded from testing cohorts (Figure S2.6; left), in order to assess generalizability across different subjects. In addition, to further evaluate the generalizability across subjects and between (task vs. resting state) conditions, t-models were also tested on HCP resting (Dataset 1) and in-house (Dataset 3) datasets (Figure S2.6; right). For detailed information about number of scans used for training and testing, please see Table S2.4.

5.5 Impact on fMRI signals and network connectivity

5.5.1 Percent Variance

To assess the degree to which our reconstructed RV and HR signals could account for fMRI signal fluctuations across the brain, we examined the percentage of temporal variance explained in each fMRI voxel signal by the reconstructed RV and HR signals. The predicted time courses were first convolved with a previously determined transfer function (for RV, respiration response function (Birn et al., 2008); for HR, cardiac response function (Chang et al., 2009)) that captures the forward mapping between physio and fMRI fluctuations, as well as their time and dispersion derivatives (Chen et al., 2020) to allow for small deviations in latency and shape from the canonical model (Henson et al., 2002). The percent variance explained was defined as the fraction by which a voxel's original temporal variance would be reduced after projecting out (via ordinary least squares) a linear combination of the aforementioned regressors, and multiplying by 100. For each scan, the RV and HR predictions used in this analysis were obtained from cross-validated rs-models, where the model that was applied to derive a subject's predicted RV/HR signals was one that had excluded that subject from the training. This analysis used the HCP resting-state data that has been bandpass filtered and downsampled by a factor of 2 (see Section 5.2). The percent variance maps calculated from individual scans were then averaged to get the population-level mapping of the RV and HR effects.

5.5.2 Connectivity Analysis

For both the ROI-based and seed-based connectivity analyses, we used resting-state scans that have undergone the HCP ICA-FIX preprocessing pipeline (Griffanti et al., 2014). The rationale behind using ICA-FIX data for this analysis is to show the utility of physiological signals even beyond ICA-FIX denoising, as substantial

RV and HR effects may remain. Beyond ICA-FIX, we applied linear and quadratic detrending to remove slow scanner drifts. This was followed by band-pass filtering within the low-frequency range of (0.01 - 0.15 Hz) and temporal downsampling by a factor of 2.

ROI-based. To assess the impact of the physiological signal of interest on regional correlations, we pursued a functional connectivity analysis. Functional connectivity commonly refers to similarities in brain activity signals between regions, and is calculated as the pairwise (Pearson) correlation of ROI time series and represented as a symmetrical matrix. We assessed the pairwise functional connectivity between all 497 ROIs drawn from cortical, white matter, subcortical, and ascending arousal network atlases (Section 5.2.4) both before and after projecting out the estimated physiological subspace using the same basis functions described above (see Section S3.2.1), as well as 6 rigid-body head motion parameters and their derivatives. The Pearson correlation between each pair of ROIs were calculated and used to construct 497 x 497 symmetrical connectivity matrices where each element represents temporal similarity score between two ROIs.

Seed-based. Seed-based analysis is one of the most common ways to explore functional connectivity within the brain (Van Den Heuvel and Pol, 2010). In a seed-based correlation analysis, connectivity is calculated as the correlation between the time course of a selected reference ("seed") region to all other voxels in the brain. The resulting connectivity map represents the Pearson correlation scores for each voxel, indicating how well each voxel's time series correlates with the time series of the seed. In this analysis, we pursued seed-based connectivity with respect to a seed region in the default mode network (DMN). For comparison, these analyses were repeated after projecting out the measured and predicted RV and HR waveforms, using the basis sets and derivatives described above (see Section S3.2.1), as well as 6 rigid-body head motion parameters and their derivatives.

S2 Supplementary Material

S2.1 Handling Missing Data During Training

The goal of training models that jointly learn RV and HR was to exploit the close coupling between RV and HR. Yet often times, only one of these physiological signals is acquired, and/or one of the signals is corrupted at particular times during the recording. Thus, we also ask whether RV and HR share common features that could enable the models to be trained on datasets that contain only one of these signals, and yet still be able to predict both (RV and HR) signals. Two experiments are designed to answer this question. In the first, we control the respective contributions of RV and HR to the loss function, and in the second, we simulate missing data.

To control the contribution of RV and HR to model training, we pursue a model training strategy whereby

the weighting of each term in the loss function is varied while the model is trained to jointly estimate both signals. To this end, we include a hyperparameter λ which governs the respective contributions of each term. As shown in Figure S2.1a, even a small contribution from either term (when λ is 0.0001) is sufficient to learn the corresponding waveform. When there is no contribution from one of the terms (i.e. $\lambda = 0$, corresponding to no RV, or $\lambda = 1$, corresponding to no HR), the models did poorly in learning the signal (RV or HR) whose input was not taken into account (see Figure S2.1a).

$$L = \lambda * L_{RV} + (1 - \lambda) * L_{HR} \tag{2.1}$$

In the next experiment, we asked whether we could predict HR, which tends to be noisier than RV, based on a model trained primarily with RV signal outputs. To simulate this scenario, we pretrained a model using all available RV data and fine-tuned these models using a varying number of HR labels, which imitated the case of missing HR data. We find (Figure S2.1b) that the models indeed benefited from initialization with pretrained RV-weights in the case where limited HR labels were made available in the training. As expected, the advantage was less pronounced when more HR labels were used.

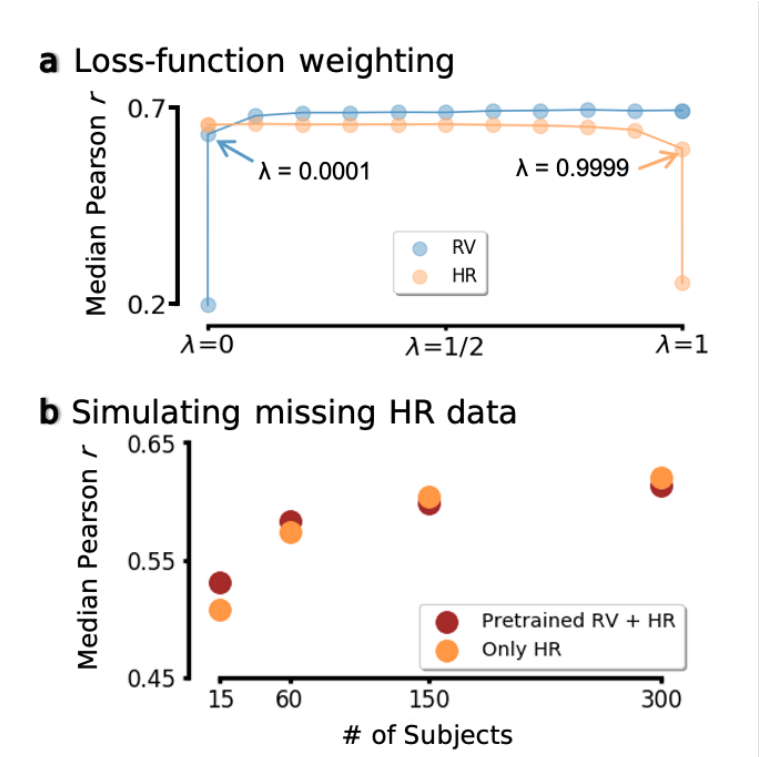


Figure S2.1: (a) **Dependence of prediction accuracy on loss-function weighting.** For the deep bi-LSTM model, 13 models with different weights λ governing loss terms were trained (note the unequal divisions of the x-axis). Prediction performance is calculated as the mean Pearson correlation coefficient across the test set. (b) **Simulating missing HR data.** Models are trained with the goal of estimating HR signals in the scenario of missing or few HR recordings during fMRI scans. Model performance is shown when using a varying number of HR labels together with pretrained RV weights (red); or when using a varying number of HR labels without pretrained RV weights (orange).

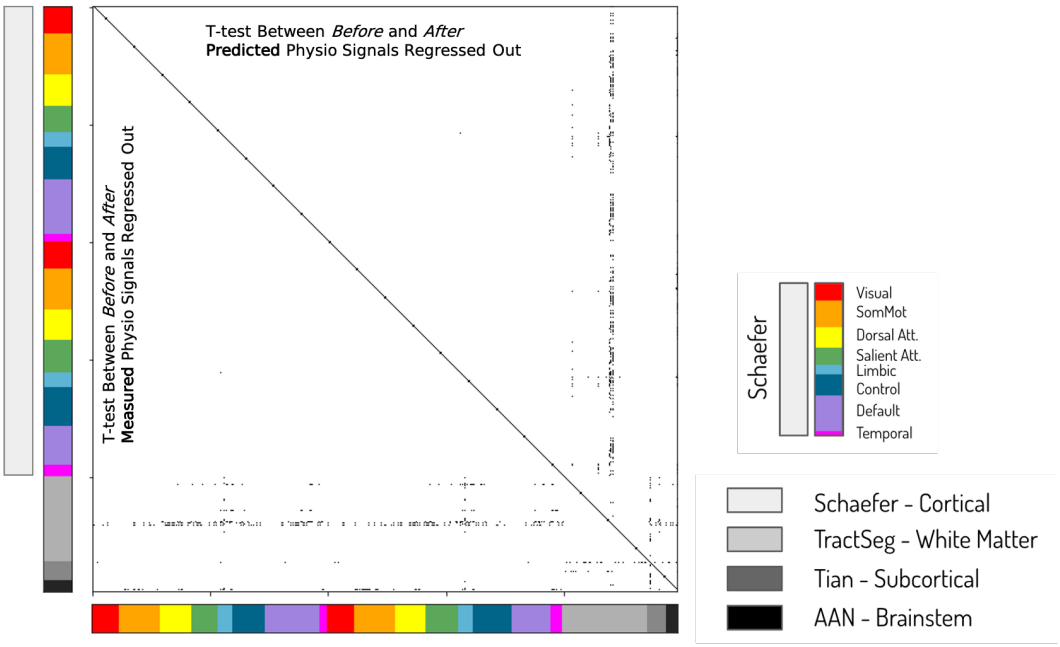


Figure S2.2: Effects of regressing out the measured (lower triangle) and predicted (upper triangle) physiological signals on group-level functional connectivity. First, we conduct t-tests on functional connectivity values at each pair of ROIs, calculated before versus after removing **measured** physiological signals from fMRI data, and then repeat the experiment for **predicted** physiological signals. The resulting values are binarized at the statistical significance threshold of $p<0.05$, with false discovery rate (FDR) correction for multiple comparisons.

S2.2 Impact of predicted signals on functional connectivity at the group level

Here, we determine brain regions whose functional connectivity, at the group level, is significantly altered by modeling physiological information, comparing the use of measured versus predicted physiological signals. First, we conduct t-tests between functional connectivity values at each pair of ROIs, calculated before versus after removing **measured** physiological signals from fMRI data, and then repeat the experiment for **predicted** physiological signals. We provide the resulting p-values to a multiple comparison test, the false discovery rate (FDR) (Genovese et al., 2002). The resulting maps are binarized using threshold 0.05 and shown in Figure S2.2. We observed that, despite a high degree of individual variability in the influence of physiological regression on FC, the group-level FC matrices were similar in many regions except those indicated in this plot.

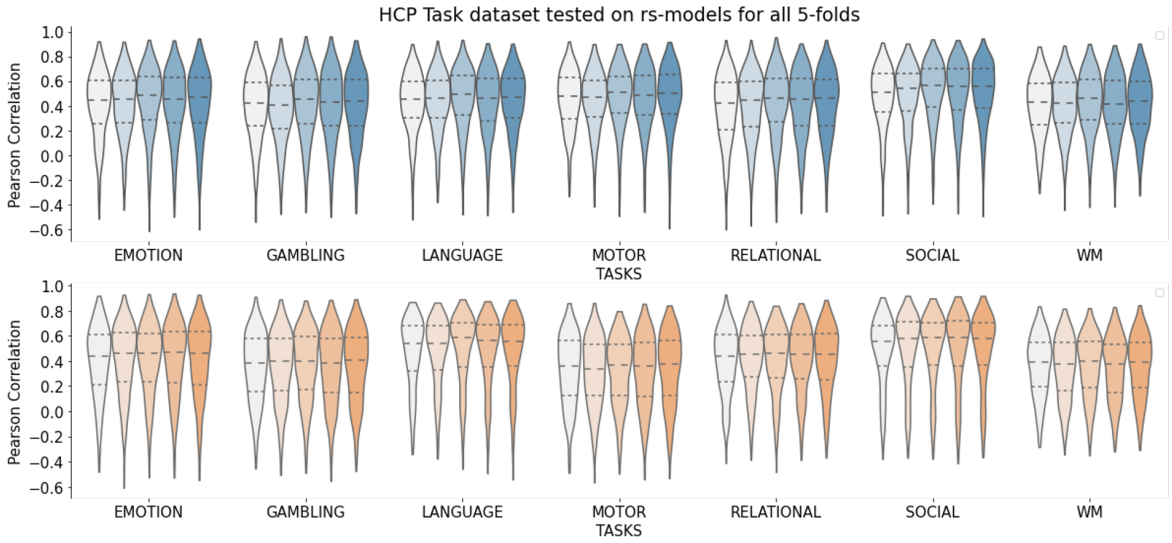


Figure S2.3: Prediction of physiological signals in the HCP task dataset, using each of the 5 models produced by 5-fold CV on the training (HCP resting-state) data. Similar performance across the 5 folds indicates that the training resulted in reliable models. RV – shown in blue, HR – shown in orange.

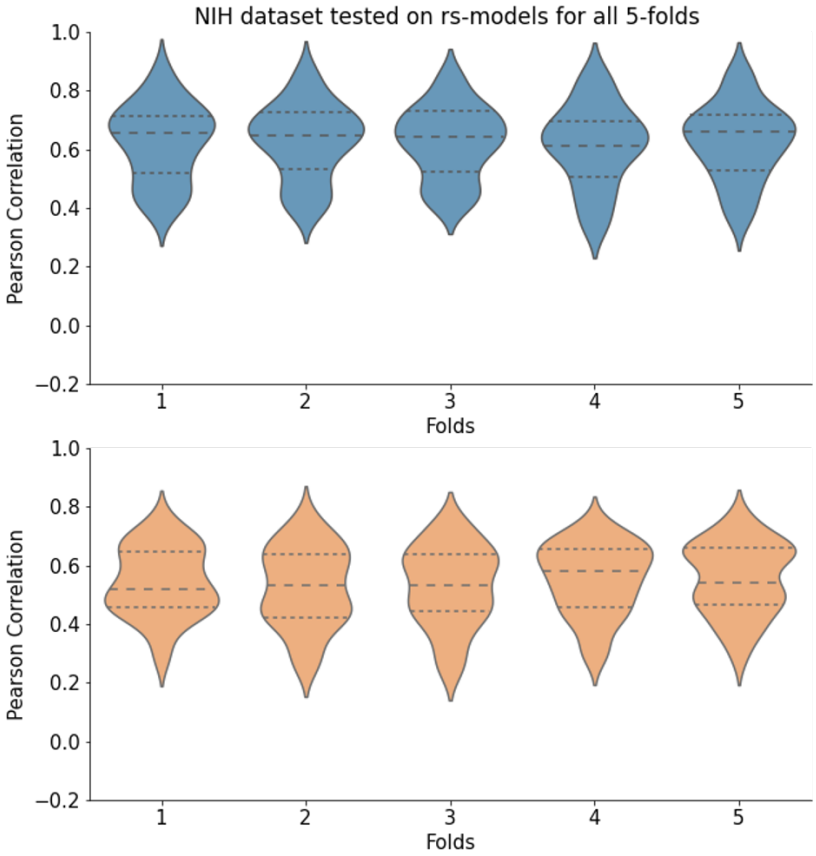


Figure S2.4: Prediction of physiological signals in the in-house dataset, using each of the 5 models produced by 5-fold CV on the training (HCP resting-state) data. RV – shown in blue, HR – shown in orange.

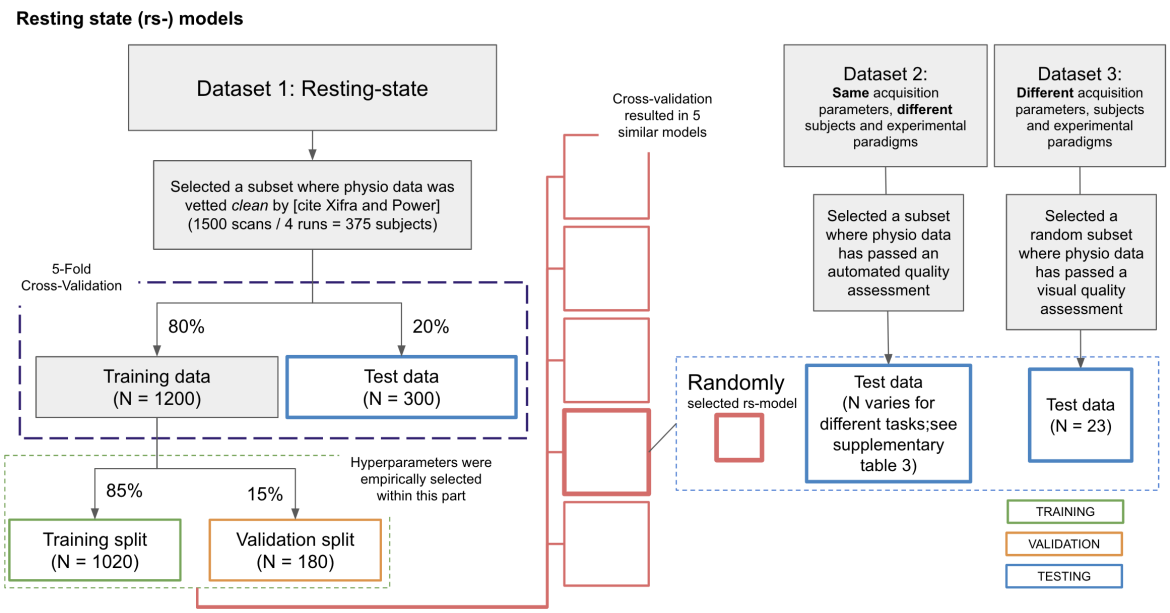


Figure S2.5: Resting state model training.

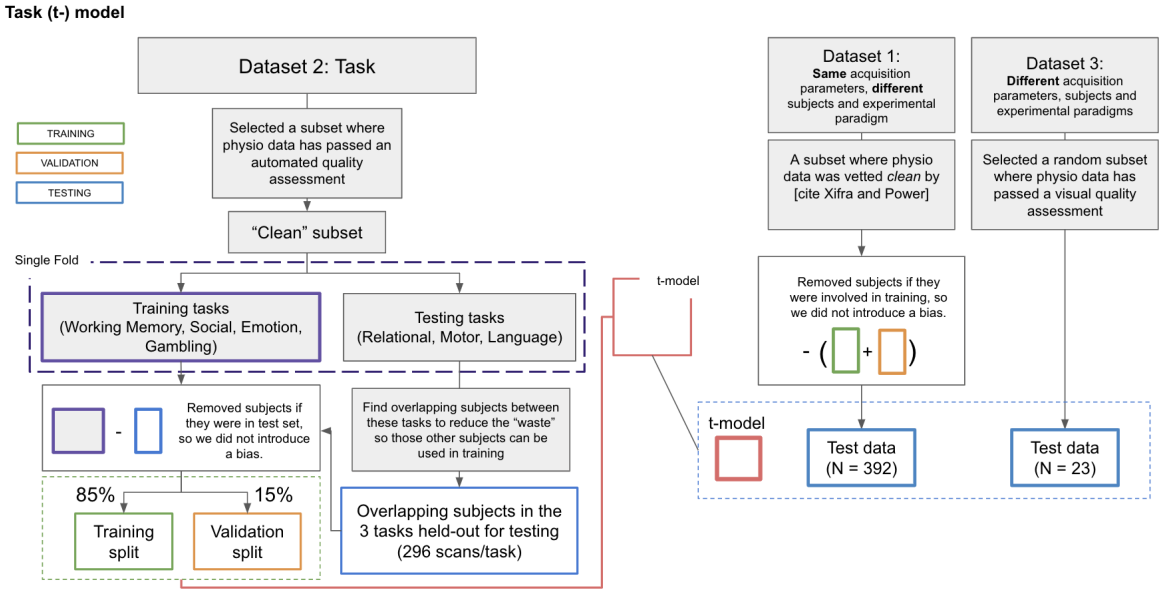


Figure S2.6: Task model training.

	Training	Testing		
		Held-out	Out-of-sample	Out-of-distribution
Resting State Model	Resting State	Resting State	Working Memory, Social Cognition, Emotion Processing, Gambling, Relational Processing, Motor, Language	In-house
Task Model	Working Memory, Social Cognition, Emotion Processing, Gambling	Relational Processing, Motor, Language	Resting State	In-house

Table S2.1: Datasets used for training and testing our models.

Resting-State	Working Memory	Social Cognition	Emotion Processing	Gambling	Relational Processing	Motor	Language	In-House
1500	977	1018	1209	1136	1027	649	976	23

Table S2.2: Number of scans that passed quality checks.

Training	Testing							
Resting-State	Working Memory	Social Cognition	Emotion Processing	Gambling	Relational Processing	Motor	Language	In-House
1200	620	634	735	716	645	394	632	23

Table S2.3: Training and testing data for resting state (rs-) models.

Training				Testing				
Working Memory	Social Cognition	Emotion Processing	Gambling	Relational Processing	Motor	Language	Resting State	In-House
652	664	840	797	296	296	296	392	23

Table S2.4: Training and testing data for task (t-) model.

	RV			HR		
	Relational Processing	Motor	Language	Relational Processing	Motor	Language
rs-model	0.448	0.469	0.488	0.453	0.559	0.362
t-model	0.552	0.474	0.534	0.516	0.600	0.458

Table S2.5: Model performance comparison between resting state (rs-) and task (t-) models when applied to task data. Performance is measured as the Pearson correlation score between measured and predicted physiological signals in test set. Median scores are reported in this table.

CHAPTER 3

Physiological Signatures Across the Brain

While low-frequency systemic physiological (e.g., respiration and cardiac) fluctuations are often treated as a nuisance component in fMRI research, an increasing number of studies indicates that they contain meaningful untapped information about brain physiology and autonomic function. Previous work has shown that emotions and personality could be linked to the interactions between brain regions and physiological signals, and that respiratory variations may track neuronal signals. Yet, the behavioral relevance of systemic, low-frequency BOLD effects, is largely underexplored. To probe this relationship, we examined large-scale patterns of low-frequency respiratory volume (RV) and heart rate (HR) across the brain and whether they relate to individual differences in battery of behavioral and cognitive traits. The first canonical variate of the whole-brain physiological signature resembling the XX network, was significantly related to an axis of positive and negative cognitive and personality traits, as well as emotion and behavior. These findings indicate that the physiological signals encompass valuable information pertaining to human cognition and behavior. This work also has important implications for current challenges regarding artifact removal practices in neuroimaging.

1 Introduction

The intricate interplay between the brain and body constitutes a fundamental aspect of biological systems, shaping cognition, emotion, and the overall health of an organism (Barrett and Simmons, 2015; Shokri-Kojori et al., 2018; Azzalini et al., 2019; Koban et al., 2021). In human brain research, functional magnetic resonance imaging (fMRI) is a powerful and extensively utilized technique. fMRI captures spatially localized changes in neural activity (Ogawa et al., 1992; Kwong et al., 1992; Bandettini et al., 1992). However, fMRI signals also reflect various influences that impact blood oxygenation throughout the body including physiological processes (Glover, 2011; Murphy et al., 2013). Natural, slow fluctuations (below 0.15 Hz) in respiration volume (RV) and heart rate (HR), the two key physiological factors, impact fMRI measurements by affecting blood oxygen concentrations, leading to widespread correlations with variations in the fMRI signal throughout the brain.

fMRI studies often consider physiological effects as a nuisance, as they can significantly impact the results of fMRI investigations. One motivation behind this perspective is the need to address the challenges of reproducibility in fMRI research, which is a pressing issue in the field (Botvinik-Nezer et al., 2020). Failure to account for these physiological effects can complicate the interpretation of fMRI data and contribute to variability across results, including the mapping of large-scale brain networks (Birn et al., 2006; Xifra-Porxas et al.,

2021). Therefore, it is crucial to identify and characterize the influence of physiological signals on fMRI data to ensure accurate quantification and interpretation of fMRI findings.

However, growing evidence suggests that studying physiological components of fMRI data, rather than removing or considering them as nuisance effects, may offer valuable insights into the function of the brain and body. Recent studies have revealed that the dynamic influences of RV and HR on cerebral blood oxygenation exhibit spatial patterns that closely resemble core neuronal networks, indicating a close relationship and potential interactions between blood flow regulation and neuronal responses (Chen et al., 2020; Bright et al., 2020). Intriguingly, the interactions between the central and autonomic nervous systems are known to contribute to shaping individual characteristics (Shokri-Kojori et al., 2018). Moreover, it has been observed that blood flow responses associated with heart rate variability can enhance neural connectivity within networks involved in emotion regulation (Nashiro et al., 2022) and lead to positive changes in mood (Yoo et al., 2022). Dubois and Adolphs (2016) has shown removing physiological components from fMRI signals diminish the test-retest reliability of individual differences in functional connectivity, further highlighting the significance of these components. These findings indicate the potential importance of physiological signals and their associations with brain function and behavior, suggesting the need to retain and investigate these signals rather than simply remove them.

Yet, the behavioral relevance of systemic, low-frequency BOLD effects, and their large-scale patterns across the brain, are largely underexplored. This study aims to fill this gap by systematically examining the patterns of peripheral physiological influences in fMRI signals and their association to individual differences in behavior and cognition. Here, we explore the potential of these physiological signatures to serve as predictors of cognitive and behavioral variables. Specifically, we investigated relationships between the spatial distribution of these patterns and phenotypic information, seeking to unravel the functional significance of signals that have traditionally been treated as mere noise. This study contributes toward advancing our understanding of the complex interplay between the brain and the body, cognitive processes, mental health, and disease susceptibility, with implications for personalized approaches in neuroscience and healthcare.

2 Results

The objective of our study was to explore the complex interplay between the peripheral physiology, brain hemodynamics and human behavior. Through multivariate analyses, we sought to uncover meaningful relationships and identify potential underlying mechanisms that contribute to the observed variations across the population.

In this work, we used a total of 375 subjects from the Human Connectome Project (HCP) S1200 release. These subjects were selected based on their physiological signals meeting quality standards, as verified in

previous studies conducted by (Power et al., 2020; Xifra-Porxas et al., 2021). In the HCP resting-state dataset, each subject was scanned twice on two consecutive days, resulting in a total of four scanning sessions per subject. This design allowed for the examination of potential within-subject variability across scans.

In addition to the imaging and physiological data, a set of 51 behavioral and cognitive measures was included in the analysis. These measures were selected based on the exclusion criteria outlined in the study conducted by (Li et al., 2019). Furthermore, the selected measures were obtained from the 'unrestricted' version of the cognitive assessments available within the HCP dataset. To ensure independence between the datasets used for model development and evaluation, we took the family structure into account. The family structure information is provided in the 'restricted' version of the data.

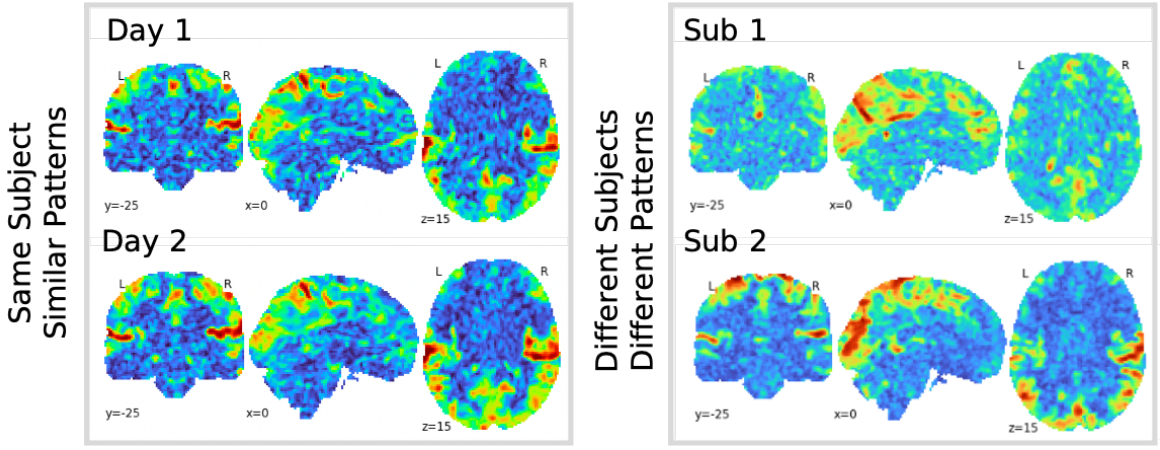


Figure 3.1: Physiological signal patterns demonstrate high within-subject reliability across different scanning sessions, and distinct patterns between individuals.

2.1 BOLD patterns of peripheral physiology

To represent the contribution of RV and HR to the temporal fluctuations in fMRI signals, we examined the percentage of temporal variance explained in each fMRI voxel across the brain. To achieve this, the predicted time courses were first convolved with previously determined transfer functions (for RV, (Birn et al., 2008); for HR, (Chang et al., 2009)). These transfer functions captured the forward mapping between physiological signals and fMRI fluctuations, accounting for their time and dispersion derivatives to accommodate small deviations in latency and shape from the canonical model (Henson et al., 2002). The percent variance explained measure quantified the extent to which the removal of these RV and HR regressors reduced the physiological-related variance in each voxel's time course, see Figure 3.2, see Section S3.2.1. PVE was computed separately for each of the four runs of resting-state fMRI data, which encompassed two days and two phase scans each day.

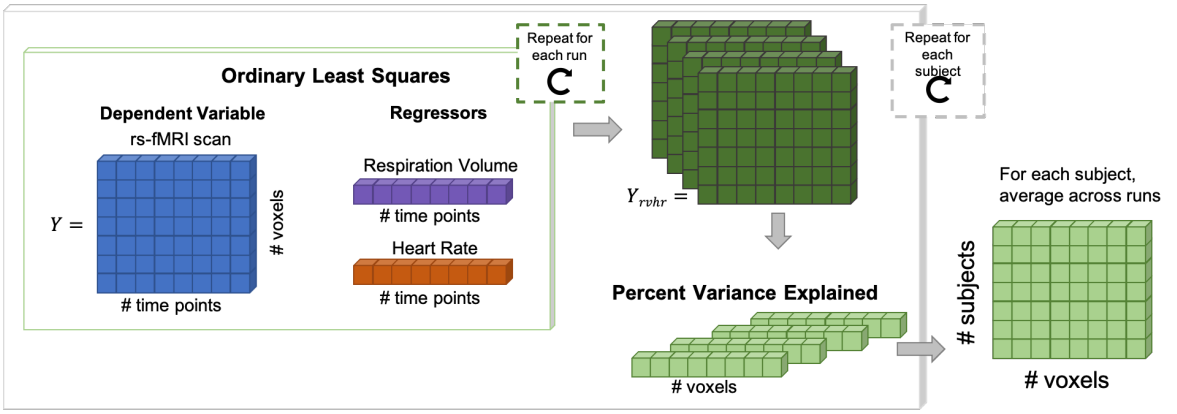


Figure 3.2: **Pipeline to calculate the percent variance explained (PVE) maps.** For each subject scan, these maps measure the degree to which the removal of respiration variation (RV) and heart rate (HR) regressors decreased the amount of physiological-related variability in the time course of each voxel.

2.2 Physiological patterns are highly structured and appear to differ across individuals

In close examination, physiological patterns, represented by PVE maps, exhibit a high degree of structure. Through visual analysis, shown for example in Figure 3.1, it has become evident that distinct physiological signatures exist, highlighting individual differences in the organization of these patterns. These differences may arise from various factors, including genetics, environmental influences, and personal characteristics. The highly structured nature of physiological patterns suggests that they reflect inherent biological variability and contribute to the uniqueness of each individual's brain function.

2.3 Physiological patterns are stable across sessions and days

To assess the stability of physiological signal patterns, we employed a correlation analysis to examine the similarity of percent variance explained (PVE) maps within and between scanning days. First, we calculated spatial Pearson correlation coefficients between PVE maps derived from the two scans acquired on the same day for each individual. Additionally, we examined the between-days consistency of physiological signal patterns by averaging the two same-day scans and computing Pearson correlation coefficients between the resulting PVE maps. This step allowed us to assess the robustness of physiological signal patterns across different days, capturing any systematic variations that may arise between scanning sessions.

To test the significance of the observed correlations, a permutation test was employed. In this test, the PVE maps were randomly shuffled across all subjects. Specifically, wherein the instances of PVE maps of a subject were assigned to the four pve maps of another subject. This permutation process was repeated for 5000 iterations. Results are shown in Figure 3.3. The analysis revealed that the physiological patterns exhibit high within-subject reliability compared to a null distribution. This finding suggests that physiological signal

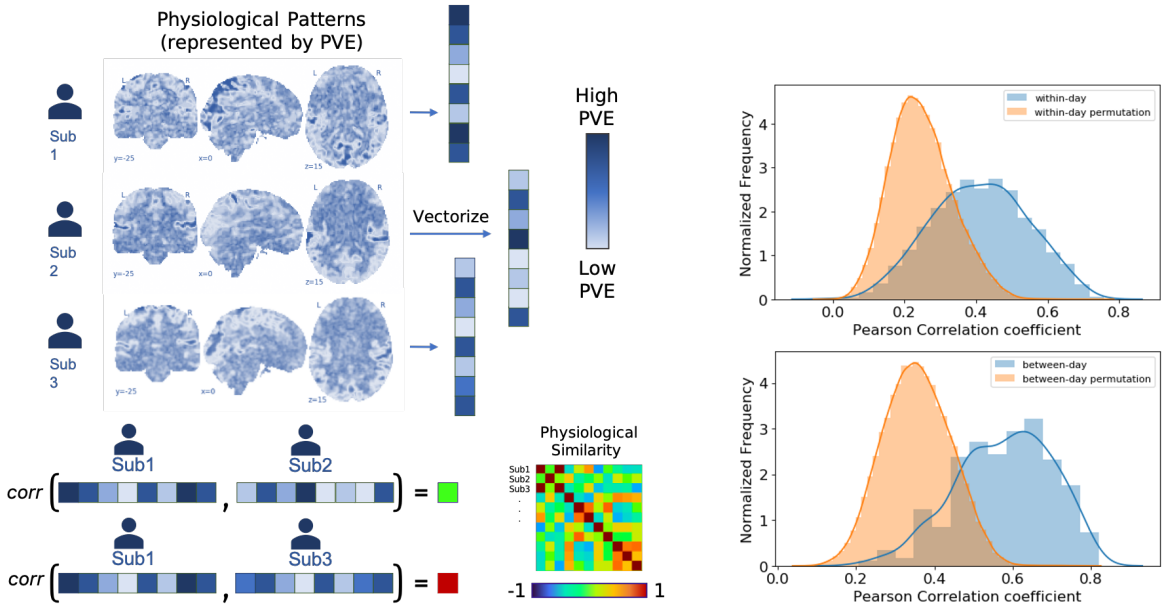


Figure 3.3: (left) **Illustration of similarity assessment between percent variance explained (PVE) maps across subjects.** (right) **Intra-subject similarity assessment within and between days.** Within-day and between day similarity is showcased to examine the consistency of BOLD physiological patterns for each subject over time. Histograms marked with blue color demonstrate the comparison of these patterns, providing insights into the stability of observed patterns across different scans. Histograms marked with orange color depict a null distribution resulting from randomizing the data distribution.

patterns show greater consistency and stability across different scanning sessions conducted on different days. Interestingly, when comparing session-averaged scans acquired on different days, a higher correlation was observed among whole-brain physiological patterns compared to individual scans acquired on the same day.

2.4 Physiological patterns may be heritable

There is growing evidence to suggest that functional topography may be heritable, indicating a genetic influence on the functional organization of the brain (Anderson et al., 2021; Barber et al., 2021). Building upon this line of research, we conducted a set of analyses investigating BOLD physiological patterns. First, we probed the similarity between family members. Second, we conducted a voxel-wise heritability analysis.

For the family similarity analysis, we averaged the four PVE maps for each individual within our cohort, resulting in a single PVE map per subject. Subsequently, we calculated the correlation coefficient between the PVE map of each subject and the PVE maps of all other subjects, generating a symmetric matrix of size 375x375. For visualization, we clustered families together, starting with the largest families (i.e., families of four) and proceeded to the lowest (i.e., families of two). It is important to note that these families exclusively consisted of siblings due to the age range of the HCP dataset (22-35 years). This family-based grouping allowed

us to investigate within-family and between-subject similarities in whole brain-wide physiological patterns.

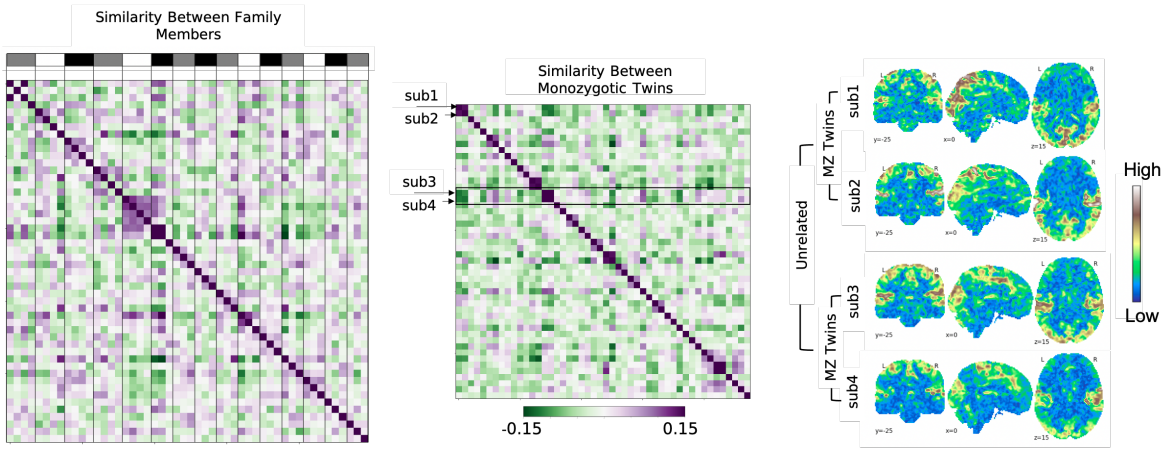


Figure 3.4: **Heritability of physiological signal patterns.** (Left and middle panels) The similarity matrices visually represent the relationship between BOLD physiological patterns among family members and other subjects. Each row and column in the matrix corresponds to an individual participant, where family members are grouped together. The matrix is color-coded to reflect the degree of similarity, with purple hues indicating higher similarity and green hues indicating lower similarity. By examining the matrix, we gain insights into the shared physiological patterns within families and their distinctiveness compared to patterns observed in other subjects. This analysis provides valuable information on the potential influence of genetic and environmental factors on the observed spatial physiological patterns. (Right panel) The twin samples of PVE maps illustrate the relationship within and between twins. Monozygotic (MZ) twins exhibit stronger correlations within the same family and exhibit distinct spatial variations compared to the two sets of other twins.

Our findings demonstrate a notable pattern of higher correlation within certain families, particularly among monozygotic twins, in comparison to the overall population. This observation strongly suggests a genetic influence on the similarity of physiological patterns within family units. The elevated correlation observed among monozygotic twins provides compelling evidence for a genetic component that contributes to the shared physiological patterns between genetically identical individuals.

2.5 Physiological patterns correlate with personality and cognition

Canonical Correlation Analysis (CCA) is a natural and suitable method for exploring the underlying relationships between two sets of variables. To investigate potential relationships between BOLD physiological patterns and behavioral measures, we performed this multivariate analysis. Prior to CCA, to mitigate potential scan-specific variations, the PVE maps from all four runs were averaged. In order to address potential confounding factors, we implemented a deconfounding procedure similar to Smith et al. (2015); Li et al. (2019) and detailed this procedure in Section S3.2.3. The deconfounded PVE maps were then standardized (subject-wise z-scored), ensuring that all subjects contribute equally to the analysis and eliminating any potential bias caused by differences in overall magnitudes across subjects, thereby emphasizing the spatial configurations.

Here, to mitigate the risk of overfitting in the CCA analysis and reduce the dimensionality of the brain and behavior data, we employed Principal Component Analysis (PCA). The z-scored PVE maps were used as input for PCA. For this, we randomly selected a subset of 200 subjects from our dataset.

In the CCA, two sets of variables - here, 30 principal components derived from the deconfounded PVE maps (brain) and 2 principal components derived from the HCP behavioral data, were analyzed together to identify the most correlated patterns or "modes" of variation between them. The modes were calculated as in (Smith et al., 2015). To assess the significance of the first two modes, we employed permutation testing with 5000 iterations. The results are shown in Figure 3.5. We observed a near significant first canonical mode between the physiological patterns and the phenotypic profiles of the population (r=0.482, $p<0.088$). For visualization and interpretation, the brain canonical variables were summarized into 14 networks, divided into 90-ROIs (Shirer et al., 2012). Analysis of these summary maps revealed that the first brain canonical mode closely mirrored spatial patterns associated with the distributed networks. Figure 3.5 also presents the top 10 positive and negative CCA behavioral components. Notably, positive weights were predominantly associated with cognitive traits, indicating a relationship between these traits and positive brain components. Conversely, negative weights were primarily observed in more negative cognitive and self-report measures. Although the second mode did not reach significance (r=0.343, $p<0.5$), it exhibited some interesting patterns. Figure 3.6 demonstrates the second canonical mode,notably, positive weights in this mode were primarily associated with positive emotional measures, while the negative axis reflected negative self-report measures.

2.6 Relation to other low-frequency BOLD patterns

2.6.1 Fractional amplitude of low frequency fluctuations (fALFF)

Notably, since our data were band-pass filtered in the 0.01-0.15Hz range prior to the above analysis, the physiological (PVE) maps represent the fractional contribution of physiology to the total low-frequency power. However, the total low-frequency power in an fMRI signal (relative to the entire bandwidth) has itself been previously examined in fMRI research and shown to have relationships with disease and cognition (Zhang et al., 2021b). Thus, for comparison, we also examine this quantity, which has been termed "fALFF" (fractional amplitude of low frequency fluctuations). fALFF (Zou et al., 2008) measures the relative contribution of low frequency fluctuations of each voxel within a specific frequency band here, taken as (0.01-0.15 Hz) to the overall power spectrum. fALFF was computed separately for each of the four runs of resting-state fMRI data, which encompassed two days and two scans each day. To mitigate potential scan-specific variations, the fALFF maps from all four runs were averaged. In order to address potential confounding factors, we implemented a deconfounding procedure identical to the one applied to BOLD physio patterns detailed in Section S3.2.3.

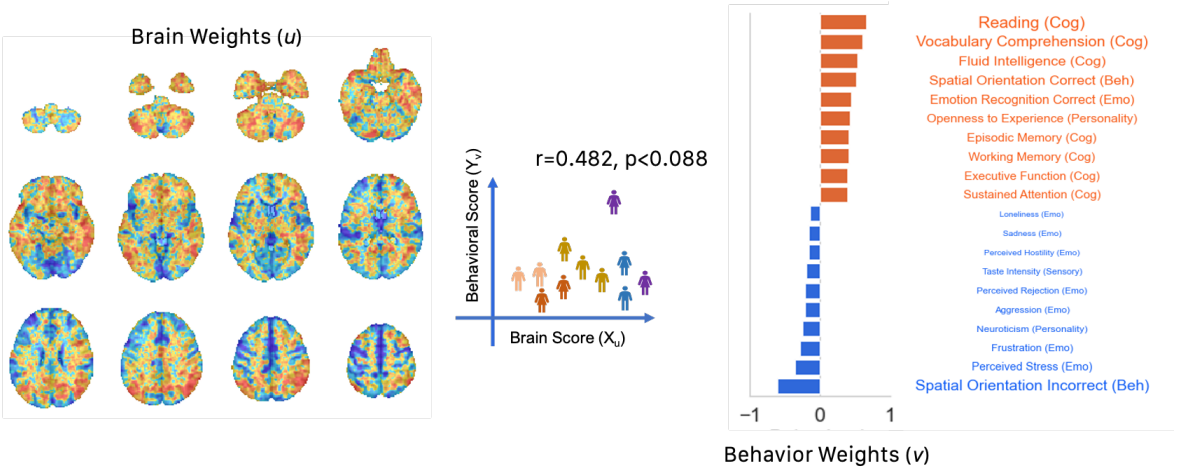


Figure 3.5: **Deconfounded BOLD Physiological Patterns 1st CCA Mode**. Panel (a) demonstrates the first CCA mode. The mode represents a composite pattern that captures the covariance between the deconfounded BOLD physiological patterns and behavioral measures. The brain weights are projected onto the 3D space, providing a spatial representation of the physiological patterns associated with the analyzed behaviors. The behavioral weights show the contribution of different behavioral factors to the observed physiological patterns. Panel (b) demonstrates the weights that are associated with the for 14 networks from Findlab atlas and the 90-ROI subdivisions of these networks.

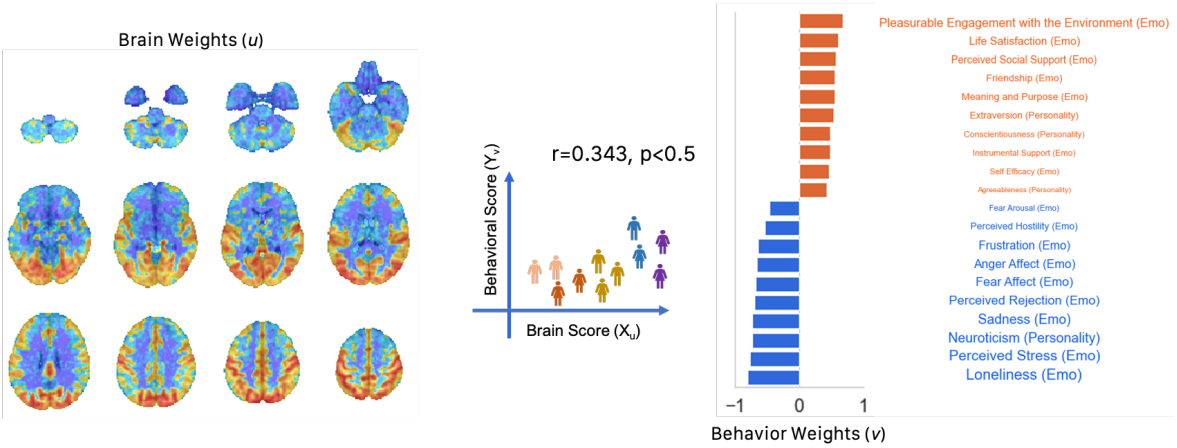


Figure 3.6: **Deconfounded BOLD Physiological Patterns 2nd CCA Mode**. Panel (a) demonstrates the second CCA mode. The mode represents a composite pattern that captures the shared information and covariance between the deconfounded BOLD physiological patterns and behavioral measures. The brain weights are projected onto the 3D space, providing a spatial representation of the physiological patterns associated with the analyzed behaviors. The behavioral weights show the contribution of different behavioral factors to the observed physiological patterns. Panel (b) demonstrates the weights that are associated with the for 14 networks from Findlab atlas and the 90-ROI subdivisions of these networks.

The deconfounded fALFF maps were then standardized (subject-wise z-scored), ensuring that all subjects contribute equally to the analysis and eliminating any potential bias caused by differences in subject-specific scales. Finally, the z-scored fALFF maps were used as input for canonical correlation analysis (CCA).

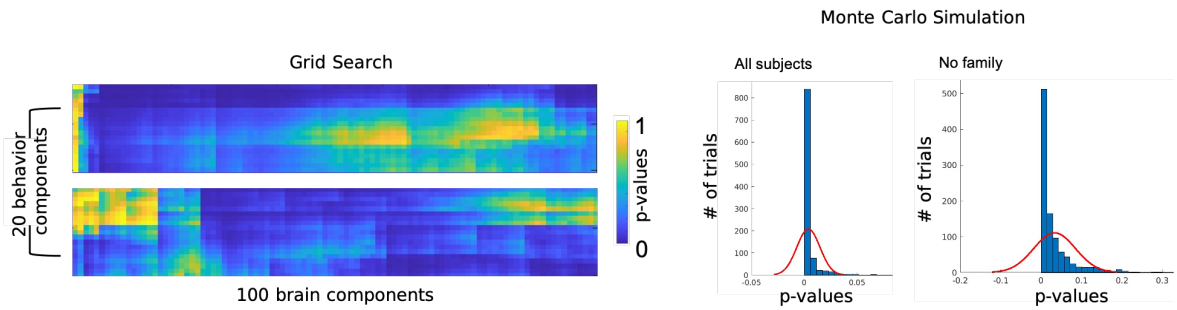


Figure 3.7: **Sensitivity Analysis for Significance of CCA Modes** To assess the robustness of the significant CCA modes, a grid search was performed on both behavior and brain components. We divided the dataset into two folds where family relationships were taken into consideration, and repeated the grid search on both subsets. This process revealed large regions where the modes exhibited statistical significance. To further validate these findings, a Monte Carlo simulation was conducted. Over 1000 iterations, 80% of the data was randomly selected, and the CCA analysis was repeated for the first mode. This simulation was carried out for all subjects in the dataset, and a subset specifically selected to exclude individuals with family relationships (only one member from each family included).

The results are shown in Figure S3.3 and S3.4. We identified a statistically significant first canonical mode between the patterns of fALFF and the phenotypic profiles of the population ($r = 0.530, p < 0.006$). Interestingly, the top behavioral components closely aligned within the same modes identified in the analysis of BOLD physiological patterns. However, the brain components captured a different spatial structure, suggesting distinct associations between the behavioral and brain measures.

2.7 The impact of deconfounding

To evaluate the impact of deconfounding on our analysis, we conducted a CCA analysis on the BOLD physiological patterns without incorporating the confounding variables. This allowed us to examine whether the inclusion of these confounders influenced the results. Upon analyzing the data, we observed a reduction in the significance levels of both the first mode ($r = 0.467, p < 0.153$) and the second mode ($r = 0.309, p < 0.794$) compared to the initial analysis. Despite the decreased significance, these modes still exhibited interesting patterns. Notably, the network-averaged positive and negative weights were similar to the deconfounded BOLD physiological patterns. Our findings suggest that the inclusion of confounds may have influenced the significance levels of the modes but did not fundamentally alter the overall patterns observed in the BOLD physiological data.

3 Discussion

The present findings of our study suggest that the variability observed in physiological signals captured by BOLD signals may reflect unique characteristics of individual subjects. Importantly, this variability appears to

be associated with behavioral and cognitive traits. These results highlight the potential significance of investigating physiological signal components in fMRI studies. By considering and analyzing these measures, we gain valuable insights into the complex interplay between brain function, individual differences, and behavioral outcomes.

Furthermore, our findings have important implications for the common practice of physiological signal regression, i.e., removing physiological signals from fMRI data based on the assumption that they are mere 'noise'. Our study suggests that these signals are closely tied to brain function and have connections to behavior. Removing them through regression may lead to the loss of valuable information. This highlights the need to carefully consider the potential influence of pre-processing strategies on the outcome of fMRI analyses – which may depend closely on the goal of one's study — and calls for further exploration of alternative approaches that preserve and incorporate the meaningful physiological signals into the analysis pipeline.

Despite the aforementioned concerns, the inclusion of 'denoising' techniques using physiological signals in data preprocessing may have indirect effects, such as reducing head-motion artifacts. Indeed, certain aspects of respiration variation may relate to head motion. Notably, recent evidence has highlighted the introduction of apparent head motions due to respiration (Power et al., 2019). Thus, depending on the specific research questions at hand, researchers should exercise careful consideration in selecting whether to exclude physiological signals during data preprocessing.

We have shown through a similarity analysis that subject-specific physiological patterns are highly stable across repeated sessions. Of note, a possible deviation in the observed difference in within-day and between-day similarity may be attributed to variations in vigilance levels. Specifically, in the HCP dataset, it has been observed that individuals tend to experience increased drowsiness during the second scan compared to the first scan [cite]. Recognizing that vigilance levels can vary across scanning sessions, we aim to mitigate some of the unwanted variability by averaging all four subject maps before conducting the CCA. By averaging multiple scans, we aim to mitigate some of the unwanted variability arising from fluctuations in vigilance levels. Further, averaging the two scans may also mitigate effects of differing acquisition methods (phase-encoding directions) between these scans in the HCP data.

CCA on both PVE and deconfounded PVE maps showed similarities in their first two modes. In both, the first mode was closely aligned with cognitive abilities and a second mode was strongly linked to traits of emotion. In the first mode, positive weights were found in several networks that are involved in higher-order cognition (e.g., in executive control, visuospatial, language). This finding suggests that when there is a more pronounced alignment between the body's internal rhythms (physiological processes) and the brain hemodynamics at rest, higher cognitive capabilities may be present. Interestingly, while we often see opposing

behavior between default-mode and executive control networks, both of them go in the same direction with regard to the sign of their CCA weights on this mode. In the second mode, positive weights are identified in the dorsal attention (visuospatial) network, which is involved in externally oriented cognition. Prior findings have shown that larger activity in cognitive networks facilitates constructive emotion regulation, whereas activity in the internally generated default-mode network has been correlated to higher depressive traits (Goodman et al., 2021) and maladaptive rumination (Hamilton et al., 2011).

The investigation of physiological patterns between twins has yielded preliminary evidence of heritability. Through qualitative analysis, we observed similarity of these patterns in subjects who are related to one another. To further explore this phenomenon, additional quantitative analysis is warranted. When conducting CCA on related subjects (the entire dataset) compared to unrelated subjects (where all family members but one were removed), we observed an increase in the significance of the identified modes, although these results are not presented here. This finding suggests that the presence of related individuals enhances the robustness and strength of the detected modes, underscoring a possible genetic connection. It is important to note that by removing the related subjects from the analysis, we unintentionally eliminated a portion of the genetic variability present in the dataset.

Significant modes were also identified when an alternate metric of fMRI fluctuations, fALFF, was investigated. The relationship between BOLD physiological patterns (PVE) and fALFF is not straightforward. In the present study, the physiological patterns were derived from fMRI data that underwent band-pass filtering in the range of (0.01-0.15 Hz). Hence, the resulting maps provide information about the fractional contribution of physiology to the *total low-frequency power* within our data. On the other hand, fALFF quantifies the relative contribution of low-frequency fluctuations to the *overall power spectrum*. Since effects of RV and HR predominantly occur within the low-frequency band, there is some connection between physiological effects and fALFF: in particular, an increase in RV and/or HR within a voxel has the potential to amplify the total low-frequency content of the signal relative to the entire bandwidth, leading to an increase in fALFF values. However, it is crucial to highlight that the RV/HR PVE maps generated in this study do not capture the fractional contribution of RV/HR to the total bandwidth. Instead, these maps are normalized based on the total power specifically within the low-frequency range of 0.01-0.15 Hz; thus, we may not expect the maps to align directly with one another. This normalization approach allows us to examine the relationship between fALFF and the fractional contribution of RV/HR within the low-frequency band in a distinct manner. Examining RV/HR variance in relation to the full power spectrum may also form an interesting avenue of future study.

S3 Supplementary Material

S3.1 Dataset

A set of resting-state fMRI (rs-fMRI) scans was drawn from the publicly available HCP 1200 subject release. fMRI scans in this release were acquired using a simultaneous multi-slice EPI sequence with the following parameters: TR = 0.72 s, duration of 14.33 mins, voxel size of 2 mm isotropic, TE = 33.1 ms, multi-band factor = 8, flip angle = 52 deg and 72 slices. During the resting-state scans, subjects were instructed to keep their eyes open and fixate on a cross-hair, and subjects underwent 4 scans (two runs on one day, and two runs on a second day). A subset of the HCP rs-fMRI dataset, consisting only of those subjects, whose physiological signals have passed a quality assessment in both (Power et al., 2020) and (Xifra-Porxas et al., 2021), was included in this study. This procedure resulted in N = 375 subjects (with all 4 runs, totalling 1500 scans). Details of the data collection can be found elsewhere (Van Essen et al., 2013). Details about behavioral measures can be found in HCP S1200 Data Dictionary and Van Essen et al. (2013).

S3.1.1 Preprocessing

Resting-state fMRI scans had undergone the HCP ICA-FIX preprocessing pipeline (Griffanti et al., 2014). A moderate spatial smoothing was applied with a 4 mm full-width at half-maximum (FWHM) kernel. This was followed by band-pass filtering within the low-frequency range of (0.01 - 0.15 Hz) and temporal downsampling by a factor of 2.

To preprocess the physiological data, we extracted the heart rate (HR) from the pulse oximetry signal by computing the inverse of the mean inter-beat interval within sliding windows of 6 seconds centered at each fMRI time frame (TR). Similarly, the respiration variation (RV) signal was obtained by calculating the temporal standard deviation of the raw respiration waveform within a 6-second window centered at each TR, following the method described by Chang and Glover (2009). Both the RV and HR signals were then subjected to band-pass filtering using the frequency range of 0.01 to 0.15 Hz, which was identical to the filtering applied to the fMRI data.

S3.2 Data Preparation

S3.2.1 Percent variance explained (PVE) maps

PVE quantifies the extent to which the inclusion of the RV and HR regressors reduce the physiological-related variance in each voxel's time course. To calculate pve at each voxel, we first colvolved the raw RV and HR time courses with previously determined transfer functions. For RV, the respiration response function (Birn et al., 2008) was used, while for HR, the cardiac response function (Chang et al., 2009) was utilized. These transfer

functions captured the forward mapping between physiological signals and fMRI fluctuations, accounting for their time and dispersion derivatives to accommodate small deviations in latency and shape from the canonical model (Henson et al., 2002). After these regressors are calculated, we calculated the fraction by which a voxel's original temporal variance would be reduced after projecting out a linear combination of the RV and HR regressors using ordinary least squares regression, then multiplied these values by 100.

S3.2.2 Fractional amplitude of low frequency fluctuations (fALFF) maps

fALFF measures the power of low-frequency brain activity relative to the total power of all frequency components in a given brain region. The fALFF analysis specifically focuses on low-frequency oscillations in the fMRI signal. The frequency band of fALFF is often within the range of (0.01-0.1 Hz). However, in this work to be able to make a fair comparison between PVE maps and fALFF maps, we used a 0.01-0.15 Hz band since we had filtered the fMRI data with a 0.15Hz low-pass cutoff prior to analysis of physiological patterns. In our RVHR analysis, we assess the magnitude of RVHR components relative only to the 0.01-0.15 Hz band, whereas in fALFF, we're looking at how much of the signal is in the same frequency band compared to the entire frequency range. To calculate fALFF (Equation S3.2.2), we first compute the full power spectrum, representing the strength of the signal at all frequencies ($<$0.6944 Hz), then the power spectrum within the low-frequency range (i.e., 0.01 to 0.15 Hz). The ratio of to the power spectrum across the entire frequency range. This ratio provides a relative measure of the amplitude of low-frequency fluctuations compared to the overall fluctuation intensity in the brain.

$$fALFF = \frac{\sum_{fi_{\min}}^{fi_{\max}} \mathrm{P}(x,f)}{\sum_{f_{\min}}^{f_{\max}} \mathrm{P}(x,f)} \tag{3.1}$$

,

where fi denotes frequency range of interest, f denotes full frequency spectrum and x represents the fMRI data.

fALFF has been used in neuroscience research to investigate alterations in spontaneous brain activity associated with various neurological and psychiatric conditions. It provides insights into the functional connectivity and integrity of brain networks, allowing researchers to identify specific frequency bands that might be abnormal or related to particular cognitive or clinical states.

S3.2.3 Deconfounding

Deconfounding is commonly applied in order to address potential confounding factors and improve the specificity of multivariate CCA analysis. Here, we applied a deconfounding workflow similar to Smith et al. (2015);

Li et al. (2019). Briefly, a set of confounds underscored in both studies were identified. We excluded certain confounds that Smith et al. (2015); Li et al. (2019) used, in order to minimize the potential loss of physiological-related information on our subject measures, namely systolic and diastolic blood pressure and hemoglobin levels. Next, head motion (mean RelativeRMS), weight, height, cube-root of total brain volume and cube-root of intracranial volume subject measures were regressed out from both the voxel-wise data and the behavioral variables using ordinary least squares (OLS).

S3.2.4 Normalization

To mitigate the potential variations in PVE maps due to the amplitude specific to each scan and subject, a temporal normalization process was applied. This involved normalizing all spatial maps within each subject to zero mean and unit variance. Furthermore, behavioral data underwent normalization within each measure to account for the varying ranges of values. This step was necessary to prevent any single measure from dominating the subsequent CCA analysis.

S3.2.5 Sensitivity Analysis

To ensure the robustness and reproducibility of the CCA results, we performed a Monte Carlo simulation analysis using 30 brain and 2 behavioral components. In this simulation, we randomly selected 80% of all subjects (without replacement) and performed CCA analysis on this subset. By randomly selecting subsets of the subjects from the dataset, we mimicked different possible combinations of subjects that could have been included in the analysis. This sampling process was repeated 1000 times. This allowed us to obtain 1000 sets of CCA results, each corresponding to a different combination of subjects. By computing and aggregating the statistical significance of each iteration, we evaluated the reliability of our findings Figure 3.7.

Additionally, on a separate post-hoc analysis, to assess whether the number of principal components have large impact on the analysis, we conducted a grid search. For this, we utilized two subsets of PVE maps, each consisting of half of the subjects. Family structure was taken into account by ensuring participants from the same family were kept within the same set and not split across the subsets. Results are reported in Figure 3.7. As the general trends in this grid search, 25-70 brain and 1-5 behavioral components, this confirmed that our 30 behavior and 2 behavioral components for the CCA analyses were sufficient to make robust inferences.

S3.3 CCA

Canonical correlation analysis (CCA) is a natural and suitable method for exploring the underlying relationships between two sets of variables. By maximizing correlations between different combinations of variables from each set, CCA allows for a comprehensive examination of these associations. We computed pairs of canonical

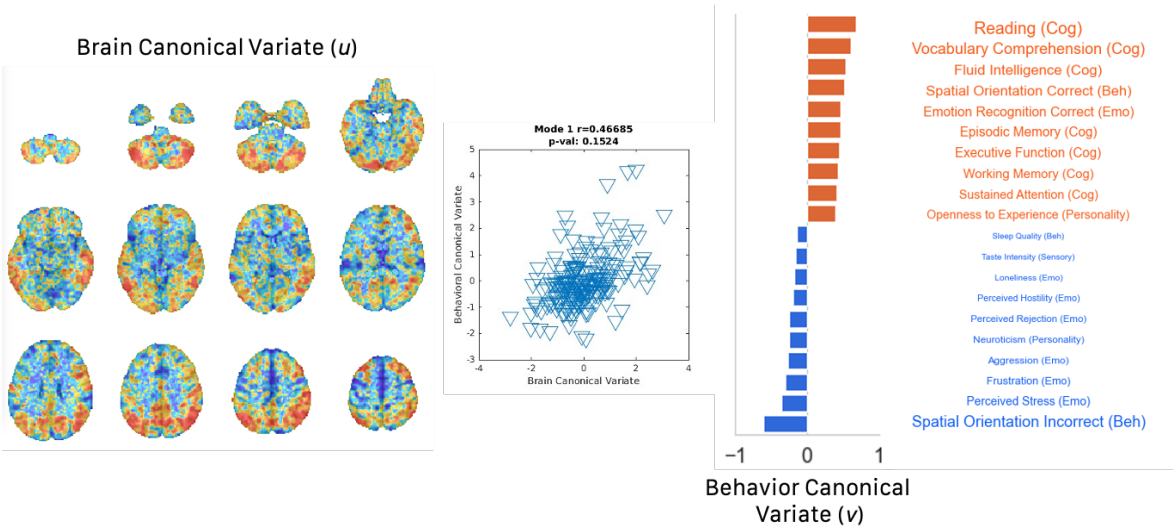


Figure S3.1: **BOLD Physiological Patterns 1st CCA Mode**. Panel (a) demonstrates the first CCA mode. The mode represents a composite pattern that captures the shared information and covariance between the BOLD physiological patterns and behavioral measures. The brain weights are projected onto the 3D space, providing a spatial representation of the physiological patterns associated with the analyzed behaviors. The behavioral weights show the contribution of different behavioral factors to the observed physiological patterns. Panel (b) demonstrates the weights that are associated with the for 14 networks from Findlab atlas and the 90-ROI subdivisions of these networks.

variates that demonstrate a shared co-variation pattern between phenotypic information and whole-brain RV and HR patterns across individuals. Each of these paired variates represents a distinct mode of co-variation. To ensure robustness of results and minimize false positives, each CCA analysis was tested for statistical significance determined via 5000 iteration permutation test. In the Canonical Correlation Analysis (CCA) employed in this study, we investigated a collection of variables. The aim was to identify the patterns or ”modes” of variation that exhibited the highest correlation between the brain (BOLD physiological patterns) and behavioral variables. The calculation of these modes followed the methodology outlined in Smith et al. (2015).

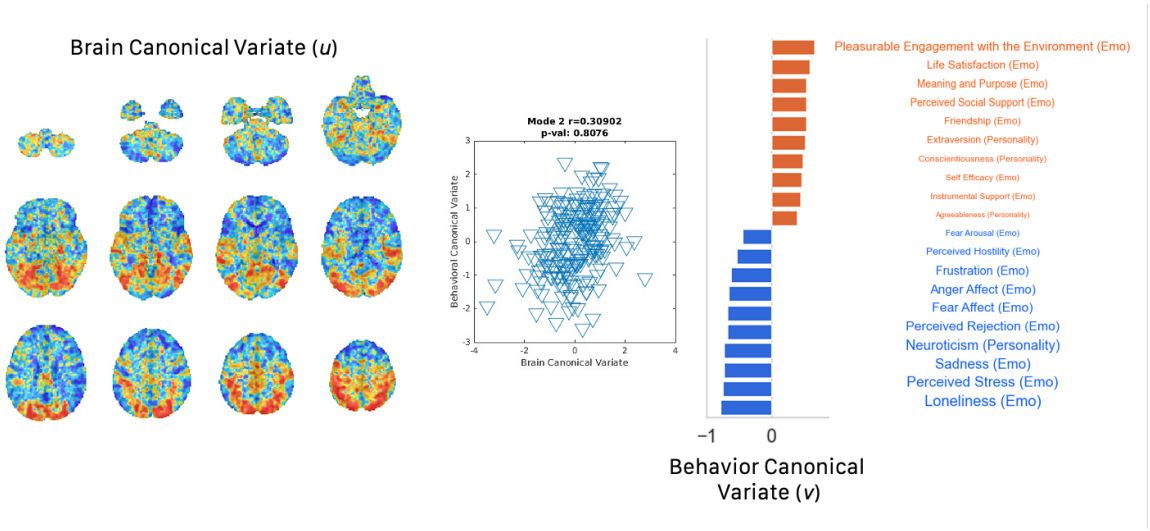


Figure S3.2: **BOLD Physiological Patterns 2nd CCA Mode**. Panel (a) demonstrates the second CCA mode. The mode represents a composite pattern that captures the shared information and covariance between the BOLD physiological patterns and behavioral measures. The brain weights are projected onto the 3D space, providing a spatial representation of the physiological patterns associated with the analyzed behaviors. The behavioral weights show the contribution of different behavioral factors to the observed physiological patterns. Panel (b) demonstrates the weights that are associated with the for 14 networks from Findlab atlas and the 90-ROI subdivisions of these networks.

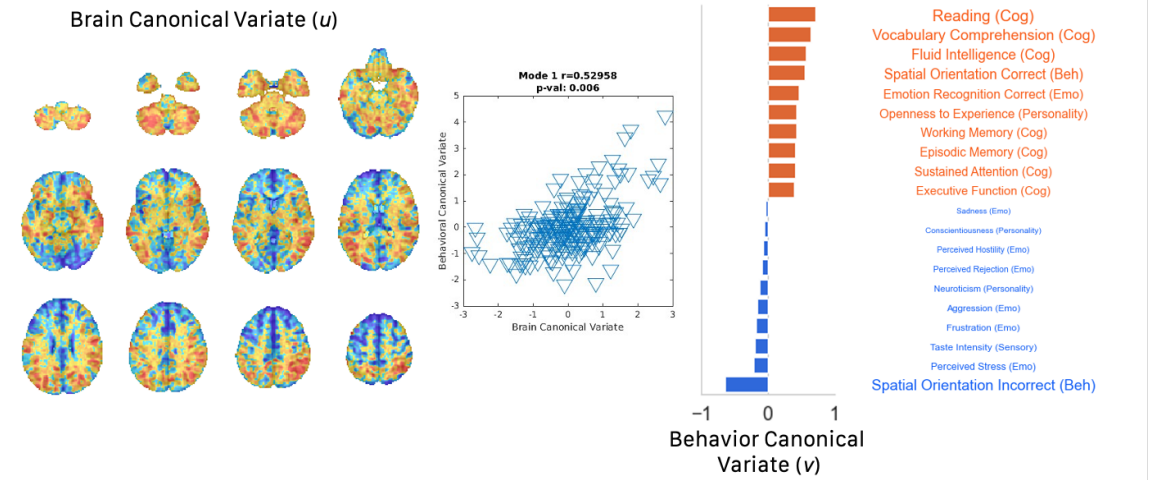


Figure S3.3: **Deconfounded fALFF 1st CCA Mode**. Panel (a) demonstrates the first CCA mode. The mode represents a composite pattern that captures the shared information and covariance between the deconfounded fALFF patterns and behavioral measures. The brain weights are projected onto the 3D space, providing a spatial representation of the fALFF patterns associated with the analyzed behaviors. The behavioral weights show the contribution of different behavioral factors to the observed physiological patterns. Panel (b) demonstrates the weights that are associated with the for 14 networks from Findlab atlas and the 90-ROI subdivisions of these networks.

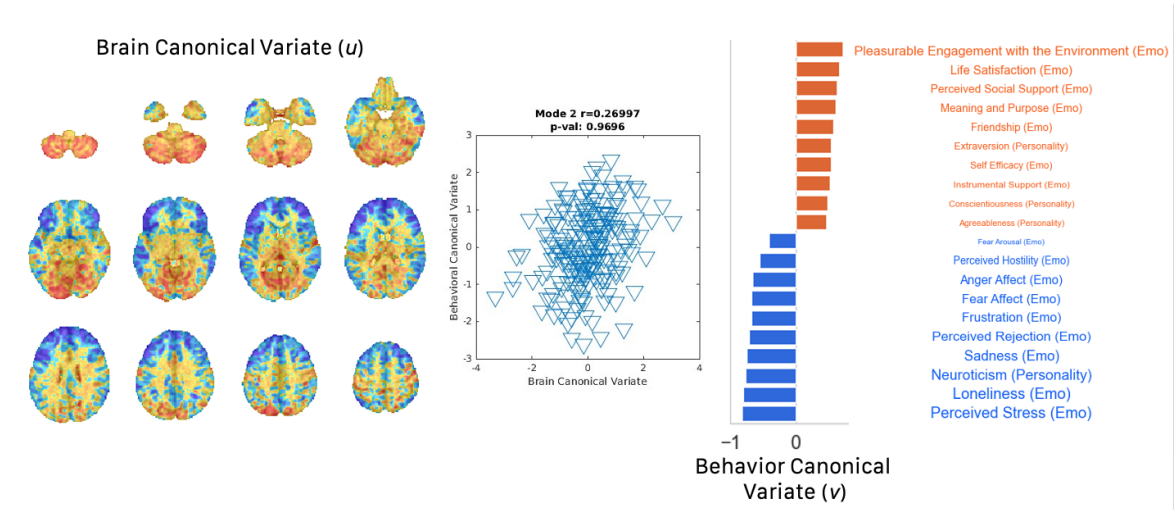


Figure S3.4: **Deconfounded fALFF 2nd CCA Mode**. Panel (a) demonstrates the second CCA mode. The mode represents a composite pattern that captures the shared information and covariance between the fALFF patterns and behavioral measures. The brain weights are projected onto the 3D space, providing a spatial representation of the fALFF patterns associated with the analyzed behaviors. The behavioral weights show the contribution of different behavioral factors to the observed physiological patterns. Panel (b) demonstrates the weights that are associated with the for 14 networks from Findlab atlas and the 90-ROI subdivisions of these networks.

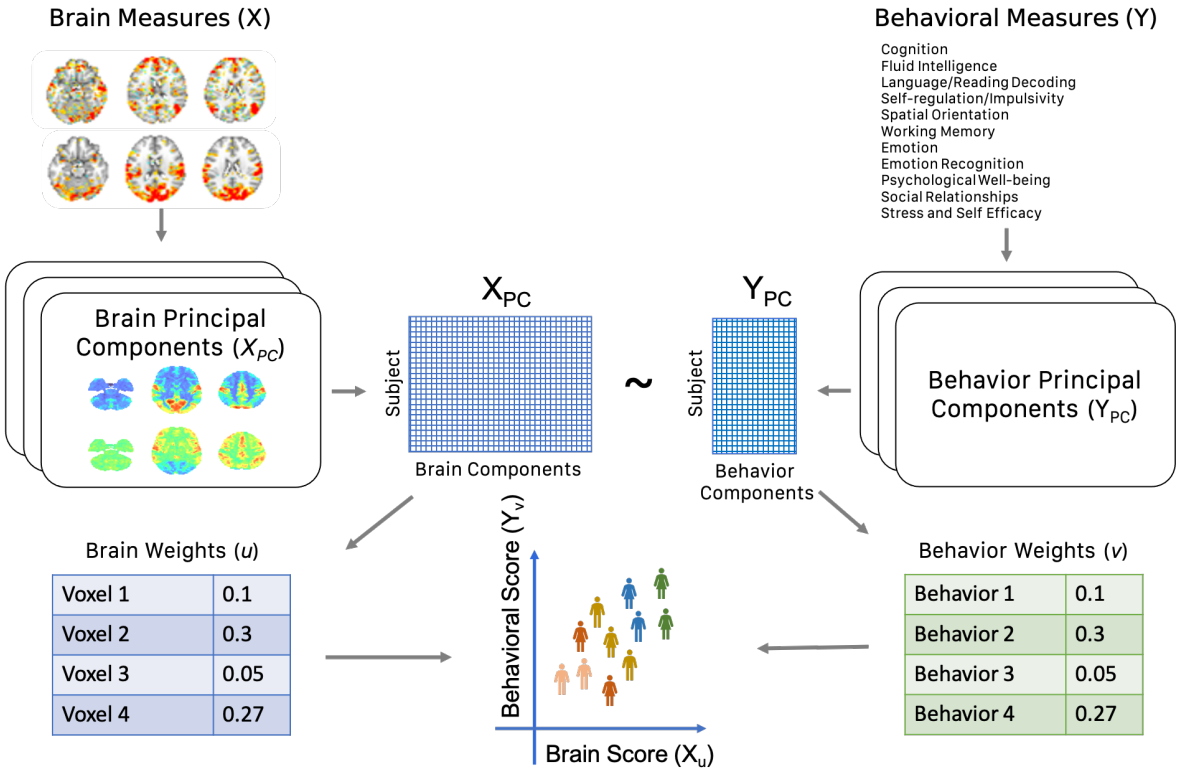


Figure S3.5: **Canonical Correlation Analysis** is performed to identify the most correlated linear combinations of brain and behavioral variables, revealing patterns of brain activity associated with specific cognitive or behavioral processes. This method provides valuable insights into the relationship between brain function and behavior, enhancing our understanding of neural mechanisms underlying various cognitive tasks and psychological constructs.

CHAPTER 4

Interactively Constructing Functional Brain Parcellations

A prominent goal of neuroimaging studies is mapping the human brain, in order to identify and delineate functionally-meaningful regions and elucidate their roles in cognitive behaviors. These brain regions are typically represented by atlases that capture general trends over large populations. Despite being indispensable to neuroimaging experts, population-level atlases do not capture individual differences in functional organization. In this work, we present an interactive visualization method, PRAGMA, that allows domain experts to derive scan-specific parcellations from established atlases. PRAGMA features a user-driven, hierarchical clustering scheme for defining temporally correlated parcels in varying granularity. The visualization design supports the user in making decisions on how to perform clustering, namely when to expand, collapse, or merge parcels. This is accomplished through a set of linked and coordinated views for understanding the user's current hierarchy, assessing intra-cluster variation, and relating parcellations to an established atlas. We assess the effectiveness of PRAGMA through a user study with four neuroimaging domain experts, where our results show that PRAGMA shows the potential to enable exploration of individualized and state-specific brain parcellations and to offer interesting insights into functional brain networks.

1 Introduction

Brain mapping research aspires to capture the complexity of brain organization and the functional roles of brain regions and networks. Atlases are often used by neuroimaging researchers to partition the brain into functionally meaningful units, typically referred to as *parcellations*. Atlases extracted using functional magnetic resonance imaging (fMRI) data are widely used to define nodes for network and functional connectivity (FC) analyses, to simplify a large set of voxels into a small set of regions that are easy to comprehend, and to boost signal-to-noise ratio by averaging across voxels that share temporally homogeneous activity. When derived at the population level, atlas regions are designed to correspond across subjects, thus capturing functional brain patterns that are shared in common across individuals (e.g. we can say that the hippocampus reflects memory-related activity for a large number of individuals). However, there can often be considerable variability in the functional subdivisions of different individuals (Fornito et al., 2013; Preti et al., 2017; Gonzalez-Castillo and Bandettini, 2018). Recent research suggests that obtaining neurobiologically meaningful functional regions at the level of the individual might serve as a fingerprint of human cognition or behavior and allow for the investigation of individualized structure-function relationships, psychological traits, or genetic variations (Finn

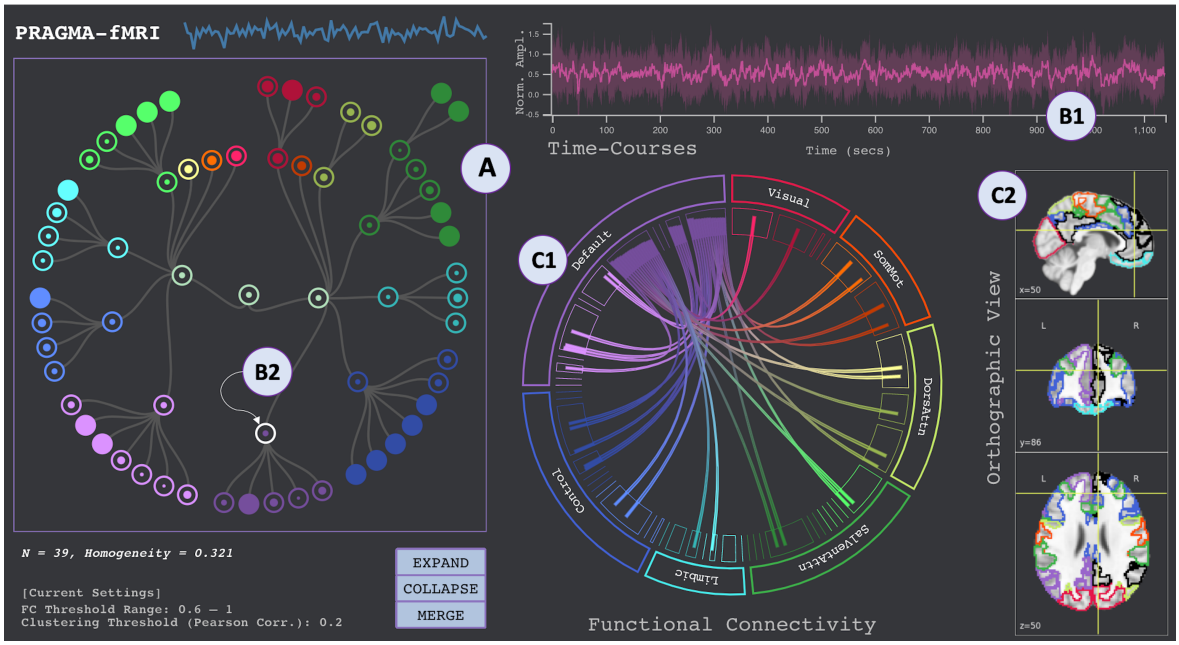


Figure 4.1: PRAGMA is an interactive tool for constructing scan-specific brain parcellations from mainstream atlases. Its interface incorporates five complementary visuals: (A) a node-link diagram as an abstract view of hierarchical structures in the brain, (B1) a line plot with confidence intervals representing similarity of fMRI signal time-courses, (B2) homogeneity glyphs within nodes, (C1) a chord diagram depicting functional connectivity, and (C2) an orthographic template showing anatomical locations of parcels.

et al., 2015; Wang et al., 2015; Salehi et al., 2018; Bijsterbosch et al., 2018; Kong et al., 2019).

A range of automated fMRI parcellation techniques have been developed for deriving individual parcellations (Wang et al., 2015; Gordon et al., 2016, 2017; Chong et al., 2017; Kong et al., 2019). However, by returning only the end-result of the parcellation, valuable information about relationships between brain regions, and how they are embedded into larger functional networks, is not readily accessible to the user. The ability to interactively parcellate a dataset, while engaging with local and long-range information at multiple spatial granularities, would provide users with richer information about individual differences in functional organization and allow for tailoring the parcellation to the specific research goals. Because the goal behind parcellating fMRI data changes between neuroimaging studies (i.e. clustering voxels based on *some* shared temporal behavior), there is no one formalism for clustering. To this end, we introduce PRAGMA, a visual analytics tool for interactively deriving parcellations of brain regions from fMRI data. PRAGMA depicts fMRI time-courses, and correlations between these signals, to support the user in interactively constructing brain parcellations. PRAGMA is inspired by studies that analyze submodularity of established atlases (Wang et al., 2015; Gordon et al., 2016, 2017), while following the conceptual framework proposed by Sacha et al.(Sacha et al., 2017b) for model steering.

However, reliably knowing what patterns to detect in order to make decisions on clustering is a non-trivial problem. To address this challenge, PRAGMA contextualizes the user's clustering with respect to an established atlas, in order to ensure their clustering decisions lead to anatomically and functionally consistent parcellations (see Figure 4.1). PRAGMA initializes the parcellation scheme from such an atlas, based on a user-specified Pearson correlation distance, and visually encodes the parcellation as a hierarchical node-link diagram (A). The user can modify their parcellation through this view by expanding, merging, and collapsing nodes, where the decisions that they make are supported by a set of linked and coordinated views that support: time-course data analysis for understanding the variability of time-series data (B1); homogeneity of parcels (B2); a functional connectivity chord diagram to investigate inter-parcel relations (C1); and an orthographic, slice-based view of their current parcellation to locate unique parcels (C2).

To evaluate PRAGMA, we performed a qualitative assessment with four domain experts. Our results demonstrate that PRAGMA could potentially leverage subject-specific analyses by allowing experts interactively identify imprecise regions and reconstruct unique functional brain parcellations.

2 Related Work

In this section, we review visualization methods designed for interactive clustering, and visual analysis methods for neuroimaging.

2.1 Interactive Clustering

There is rich literature on visual exploration of clustering (Kwon et al., 2017; L'Yi et al., 2015; Schreck et al., 2009; Yang et al., 2020; Das et al., 2020) and methods that leverage unsupervised clustering (Bruneau et al., 2015; Kern et al., 2017; Sacha et al., 2017a; Cavallo and Demiralp, 2018). More recent methods demonstrate the advantage of providing a feedback loop to the domain expert, making clustering human-centered. Hierarchical Clustering Explorer (Seo and Shneiderman, 2002) is a pioneering bio-informatics system for the interactive discovery of patterns. Since then, many approaches have been proposed for augmenting interactive clustering, by the means of finding the right clustering algorithm and parameters, and supporting exploratory clustering with visual and statistical analysis (L'Yi et al., 2015; Kwon et al., 2017; Cavallo and Demiralp, 2018; Ruta et al., 2019).

Akin to PRAGMA, in the aforementioned work, views and computational techniques are combined to help users interactively reach satisfactory clustering results. In order to use the aforementioned methods, experts need to translate their analytic goals into supported analysis techniques. In contrast, PRAGMA offers the ability to build upon an existing atlas and is equipped with domain-focused cluster analysis techniques that

inform the iterative process of parcellating fMRI data into functionally consistent regions. The visualization design supports the user in making decisions on how to perform parcellation, through a set of linked and coordinated views.

2.2 Visual Analytics in Neuroimaging

Similarly, a wealth of systems have been proposed to visualize and analyse anatomical and functional relationships between brain networks. Network-based neuroscience visualization tools have demonstrated how to gain insight from connectome datasets, some at the level of functional modules, e.g. NeuroCave (Keiriz et al., 2018) and TempoCave (Xu et al., 2019), and some at the level of individual neurons, e.g. BrainTrawler (Ganglberger et al., 2019). A different approach to analyzing functional modules using NodeTrix (Henry et al., 2007) representations is proposed by Yang et al. (Yang et al., 2016). VisualNeuro(Jönsson et al., 2020) offers a tool for hypothesis formation and reasoning about cohort study data. Brain Modulyzer (Murugesan et al., 2016) is an interactive visual exploration tool for investigating regional correlation and hierarchical network structures.

PRAGMA takes inspiration from the above visual analytics methods, but differs in task. While many methods focus on analysis of pre-defined nodes, PRAGMA focuses on redefining nodes and helping domain experts make parcellation decisions.

3 Design Objectives and Tasks

Functional magnetic resonance imaging (fMRI) detects time-varying changes associated with blood flow and captures volumetric time-courses of brain activity. As fMRI data is often noisy and of high spatial and temporal resolution, deriving individual parcellations that reflect meaningful functional information is challenging for domain experts. Automatic clustering methods, e.g. k-means, might yield results that are inconsistent with their prior knowledge, while manually creating parcellations can be a tedious process. Thus, the main objective of PRAGMA is to support domain experts in the analysis, and creation, of scan-specific parcellations.

To derive design objectives for PRAGMA, we initially conducted multiple group discussions with three experts from the fields of neuroimaging, network and cognitive neuroscience. We started our discussions by asking them about the current practices, needs, and challenges of delineating functional regions for their work. The general direction of the questions were: Do you currently use any established atlases or voxel clustering methods in your analyses? What are the pitfalls and challenges of these processes? How do you evaluate the parcels obtained by clustering? Based on responses from the experts, we identified a set of high-level design objectives, and refined them as we iteratively received feedback during different stages of development. Our discussions led to the following objectives:

(DO1) Leverage established atlases in the creation of scan-specific parcellations: Due to the difficulties of clustering fMRI data, the use of established, meaningful atlases can help domain experts avoid spurious clustering decisions that the clustering algorithm can impose on the data.

(DO2) Provide an intuitive approach for modifying the granularity of parcels: Domain experts should be able to decide the granularity of parcels based on the need of the study, and apply their domain knowledge for merging similar regions together and splitting inhomogeneous regions further into smaller regions.

(DO3) Provide useful within- and between- parcel information that supports parcellation decisions: The design should provide access to information about the current parcel properties, so that modifications are justifiable.

From these objectives, we list the following tasks that our visualization design aims to support:

(T1) Steer parcellations using a population-based atlas (DO1) The design should support the use of established atlases, both for initializing the clustering, and as a guide for steering clustering.

(T2) Support inter-parcel comparisons (DO3) The design should support the user in understanding the relationship between two parcels in their current parcellation.

(T3) Support intra-parcel comparisons (DO3) Users should be able to assess the homogeneity of a given parcel, e.g. how similar are a set of time-courses that belong to a single parcel.

(T4) Steering parcellation (DO2) Our design should allow the user to interactively modify the parcellation, where we identify three basic operations for steering: (1) splitting a parcel, (2) collapsing a set of parcels into one, and (3) merging two parcels.

We emphasize that tasks *T1-3* are intended to provide guidance for the user in making decisions on steering *(T4)*. Namely, a parcel should be split if the user does not consider it homogeneous, two parcels should be merged if they are sufficiently similar, and likewise, a set of parcels should be collapsed into one if they are all considered similar.

4 Visualization Design

In this section, we describe the visual encodings of the PRAGMA interface and their interactions. The interface is composed of a Hierarchical Node-Link Diagram (A), two Parcel Specific Views (B), and two Current Parcellation Views (C) (see Figure 4.1).

4.1 Hierarchical Node-Link Diagram (A)

The diagram represents the full hierarchy of the parcellation, and it is the main functional visual in the tool (see Figure 4.1). The parcels are grouped by functional networks in the left and right hemispheres. Each network is

represented by a unique color, and this coloring is consistent across the node-link diagram, chord diagram, and orthographic views. Furthermore, lighter and darker hues represent networks in the left and right hemispheres respectively. While the root node represents the whole brain, each leaf node in the hierarchy represents a parcel from the current parcellation.

The user selects a distance threshold to initialize the hierarchical node-link diagram. Pearson's correlation is used to form a precomputed distance matrix based on this threshold. The initial node-link diagram is formed by applying complete-linkage agglomerative clustering to this precomputed matrix, done on a network-by-network basis. Afterwards, the user may explore parcels by selecting nodes, which updates the remaining views with parcel-specific information (see Figure 4.2). Selecting a node will return the aggregated time-course data encoded as mean +/- standard error, functional connectivity of the selected parcel(s) to the other parcels, and anatomical locations of these regions in the orthographic view.

In case two or more clusters are found to be similar, we support a *merge* action *(T4)* (see Figure 4.3). To investigate such regions, a user can double click on a leaf node of interest. This will lock the parcel-specific information, and it will persist in view unless the node is released. Users can then continue exploring other parcels in reference to this selected parcel by single click selecting other nodes. Once the user finds another parcel that they want to merge into the initially selected node, they select it with a double click on this second leaf node. The user may then click on the *MERGE* button to perform the merging of these nodes, and parcel-specific views are subsequently updated to show the newly merged region.

While *merge* only allows two regions to form into one, *collapse* can combine all the leaf nodes into a non-leaf node. A user can first select a non-leaf node and then click on the *COLLAPSE* button to demonstrate that this is the intended action. The parcels can not be collapsed into the root, hemisphere, or network nodes, in order to keep the hierarchy intact. However, parcels can be identified and merged into different networks.

If users deem that within-parcel similarity is low, they can further split this parcel into smaller regions. Subsequent clustering requires a more constricting distance threshold. After this threshold is selected, users click on the *EXPAND* button to demonstrate that this is the intended action.

4.2 Parcel-Specific Views (B)

This portion of the visualization is designed to provide real-time analysis of within-parcel properties, as a means of evaluating parcels and supporting parcellation decisions.

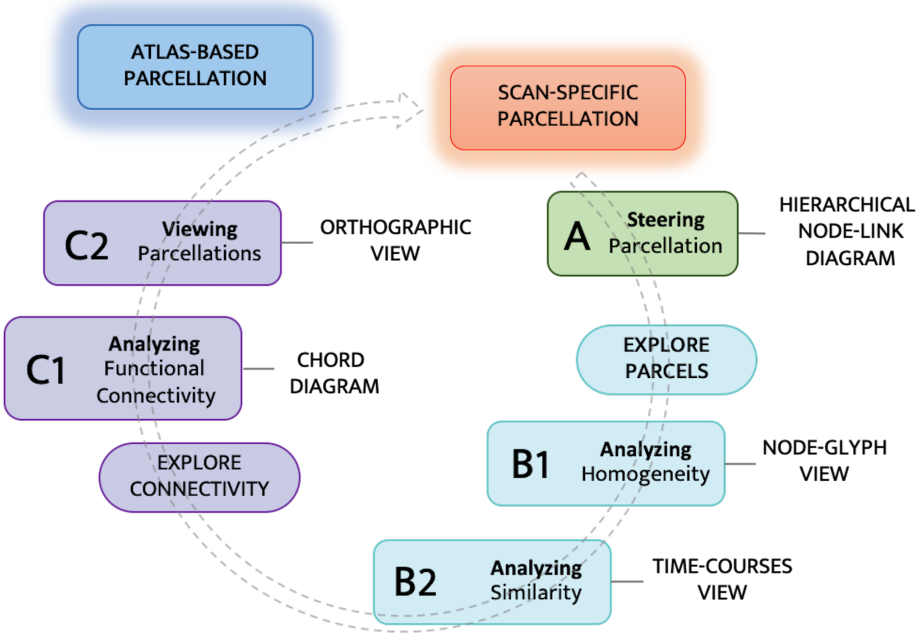


Figure 4.2: Iterative process of creating scan-specific parcellations from mainstream atlases. Each view provides some useful information to support and facilitate the decision on parcellation.

4.2.1 Intra-Parcel Time-Series Similarity (B1).

We designed a mean time-course line plot with a standard error confidence interval. This visual encoding demonstrates the similarity of the aggregated time-courses of all super-voxels grouped in a selected parcel. The main use of this analysis is to evaluate if a parcel should be further expanded. Additionally, the user can overlay the time-courses of two parcels for comparison. Brush selection is also supported to allow the user to zoom in on the time axis.

4.2.2 Parcellation Homogeneity (B2).

Homogeneity is defined as average pairwise temporal correlation, calculated using Pearson's correlation. This quantitative information is useful for making decisions about whether a parcel needs to be further divided into finer parcels. We carry this information within the node-link diagram, encoding homogeneity within the inner circle of each node. The homogeneity is normalized to the radius of the circle, where a more homogeneous parcel is represented by a bigger circle. Homogeneity is calculated for the entire hierarchy, allowing the user to compare each parcel in relation to the parcels it is linked to.

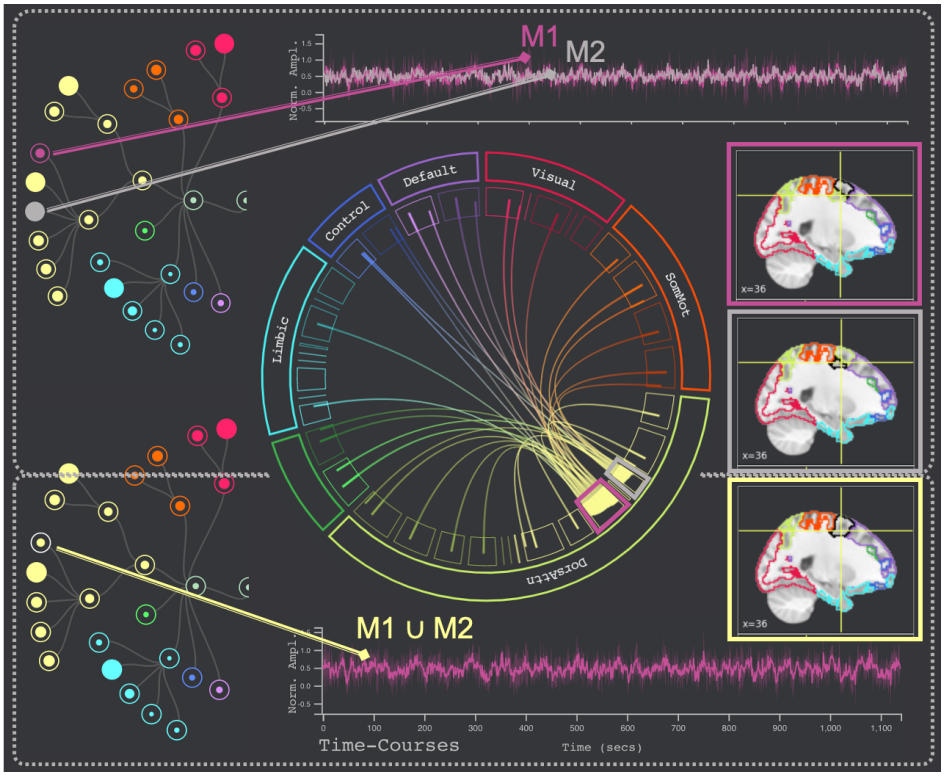


Figure 4.3: To merge two parcels together, a user selects node (M1) followed by node (M2). The top time-courses plot shows the two signals overlaid on top of each other for comparison. After the *merge* is applied, the super-voxels from the second node (M2) are merged with the first node (M1). The time-courses plot is updated to reflect the merge, along with the orthographic view.

4.3 Current-Parcellation Views (C)

This portion of the visualization is designed to support analysis of inter-parcel properties for the current set of non-leaf nodes.

4.3.1 Functional Connectivity (C1).

The chord diagram communicates functional connectivity (FC) patterns between the current set of leaf nodes. Functional connectivity describes co-activity of discrete regions in the brain and is calculated as the correlation between every pair of regions in the current parcellation. The inner arcs represent the parcels, and the chords encode the presence of connectivity between parcels. Each inner arc is accompanied by a bar plot encoding the strength of connectivity between its corresponding parcel and the other parcels. The connectivity strength is normalized to the height of each arc, where a stronger FC is represented by a higher bar. The width of the arc represents how connected a parcel is to other parcels in the current scheme. In the chord diagram, the parcels are nested under the functional networks, which are labelled in the outer arcs. When a node is selected from

within the node-link diagram or an inner arc is hovered over, the chord diagram illuminates the connections of just the corresponding parcel. To reduce visual clutter arising from the large number of connections, we allow the user to filter out chords within a prescribed range of FC strength.

4.3.2 Orthographic Parcellation View (C2).

The orthographic view shows the current parcellation scheme mapped onto a template image and highlights the corresponding physical locations of these parcels (see Figure 4.1). Every parcel is outlined by the color representing its functional network. The provided sliders allow users to view the current functional parcellation at different sagittal (yz), coronal (xz), and axial (xy) planes. When a parcel is selected in the node-link diagram, the corresponding parcel contour in the orthographic view is outlined in black.

5 User Study

A user study was performed to investigate the effectiveness of the tool. Four domain experts (two faculty, two post-docs) who have expertise in the understanding of functional networks evaluated the tool. Prior to the study, at different stages of development, three of these domain experts provided feedback that was used in the iterative process to improve the tool and one is a co-author in the paper. The participants received no compensation for taking part in the study.

Data. We used a single subject resting-state (rs-fMRI) scan from the Human Connectome Project (Van Essen et al., 2013) and Schaefer cortical brain atlas derived from rs-fMRI data (Schaefer et al., 2018). The rs-fMRI scan has been pre-processed with the "FIX" de-noising pipeline(Griffanti et al., 2014) and has undergone the HCP "minimal preprocessing pipeline" (Glasser et al., 2013). A Schaefer atlas is applied to this subject scan to initialize super-voxels by extracting averaged time courses within parcels. The latter process reduces the spatial dimension from 200K voxels to 400 super-voxels.

Implementation Details. The front-end of our tool was developed as an Observable notebook, enabling ease of access for our study, alongside a Python server that supports real-time analyses and computation. We have made PRAGMA publicly available at: https://github.com/neurdylab/PRAGMA.

Zoom-proceedings. The domain experts were invited to individual Zoom-meetings. They had been provided a user guide that included (1) instructions to install PRAGMA, (2) pre-/post- study surveys, and (3) written/animated instructions on how to use the tool. First, they were asked to fill the pre-study survey. Then the experts were explained the basic actions that they can take to use the interface (brush, hover over, etc.), as well as the visual encodings and their interactions. They were provided the pre-processed scan and asked to explore the data. The exploration of the tool was recorded via screen-record over Zoom. During this meeting

they were informed that they could ask design/functionality related questions to the visual analytics researchers. There were no restrictions on how they could use the tool, nor any time restrictions. For the group of experts who were more familiar with fMRI data parcellation, this freeform exploration took about an hour; for the other group, less than half an hour. Lastly, they were asked to fill a post-study survey to finalize their evaluation.

Findings. The findings are categorized into three main sections: design choices, intuitiveness, and scope. *Design choices*: Four experts rated the design choices (linked highlighting, spatial arrangements, etc.) as 7.5/10 on average. One expert commented that the arrangement of the parcels in the chord diagram (grouped by network) being different than the node-link diagram" (grouped by hemisphere), increased the cognitive load. On another note, the users observed that being able to compare two parcels on multiple views at the same time and interact with them (i.e. brush to zoom) was helpful. *Intuitiveness*: The average rating of all four experts for ease of use is 6.5/10. Two of the experts commented that they initially struggled with how to engage efficiently with the tool but they quickly picked up even though "there are a lot of features" to support their analyses. While the chord diagram and 2D orthographic views were found to be the most useful, the users suggested that rather than using a slider, they would prefer to be able to click on the image to navigate it. *Scope*: Two post-docs rated 8/10 that they would use PRAGMA to support subject-specific analyses. On the other hand, one expert suggested that a search for valid cognitive relations through single-scan analysis was likely to result in inconsistencies, mainly because fMRI data is very noisy. One of the experts suggested that a tool like PRAGMA could also be used for population-level exploration and clustering. Two of the experts who explored the tool used a similar sequence of actions: they located a region that they are familiar with in the orthographic view, found this region on the node-link diagram, and then compared the time-series signal and functional connectivity of the corresponding parcel to nearby parcels. The set of actions taken by these users suggested that the prior atlas was helpful.

6 Discussion and Conclusion

Forming precise and meaningful definitions of brain regions is an open and pressing research question. Our approach, PRAGMA, addresses these issues by supporting users in the customization of existing atlases to create individualized parcellations based on their domain needs. Our parcel-by-parcel analysis allows unique parcellations, while ensuring the results remain coherent with established atlases. The user study has shown that PRAGMA offers an intuitive way to understand individual-level depictions of brain activity that are not possible to resolve in group-level atlases. For future work, we plan on addressing issues related to common sources of noise in fMRI data (Liu et al., 2017; Zhang et al., 2020b) through visually conveying uncertainty as part of our design. Furthermore, we plan on making our system more extensible, handling various atlases,

clustering algorithms, and similarity measures, in order to better address demands from domain experts.

CHAPTER 5

Learning subject-specific functional parcellations from cortical surface measures

Functional cortical parcellations that are tailored to individual subjects have been shown to improve predictions of behavior and provide useful information about brain function and dysfunction. A gradient-infused multi-session hierarchical Bayesian (gMS-HBM) model using resting-state fMRI (rs-fMRI) to delineate individualized, spatially localized functional parcels is currently the state-of-the-art. However, rs-fMRI acquisitions are not routine in clinical practice and may not always be available. Based on the extensive body of literature demonstrating a robust association between brain function and cortical sheet characteristics, we hypothesize that individual-specific functional parcellations may be inferred using these features from more commonly acquired T1- and T2-weighted structural MRI scans. Here, we investigate this hypothesis by employing spherical convolutional neural networks (sCNN) to infer individualized functional parcellations from cortical thickness, sulcal depth, curvature, and myelination features. We show that versatile sCNN models can achieve comparable parcellation accuracy against rs-fMRI derived ground truth labels, with a mean Dice score of 0.77, while baseline FreeSurfer models fail to capture this relationship. We showed that reconstructed individual-level parcellations improve areal functional homogeneity over widely used group parcellations. We envision the use of this framework for predicting the spatially contiguous areal labels at the level of individuals when rs-fMRI is not available.

1 Introduction

Complex human cognition is thought to be enabled through communication between functionally synchronized distributed networks that are comprised of computational units across the brain. To study human behavior and brain function, there has been a significant interest in mapping these localized functional regions using non-invasive brain imaging techniques such as magnetic resonance imaging (MRI). These delineations of distinct functional units across the brain, also called parcellations, are widely used to facilitate investigations into cognition and disease (Arslan et al., 2018; Eickhoff et al., 2018).

Earlier efforts focused on defining brain areas through histological examination of post-mortem brains (Brodmann, 1909). With the advent of MRI, architectonic and topographical features could be examined non-invasively (e.g., Harvard-Oxford, Juelich atlases). In addition to structural and anatomic features, functional MRI has also been used to find temporally coherent regions across the brain (Shen et al., 2013; Schaefer et al., 2018). Computational atlases, both structural and functional, are typically summarized at the population level

(Schaefer et al., 2018; Thomas Yeo et al., 2011) and are designed to correspond across subjects, thus capturing stable units that are shared in common across individuals.

However, there is considerable variability across people that the group-level mappings fail to capture. Indeed, it has been shown that the neuroscientific validity of statistical inferences may improve greatly when using individualized parcellations (Bijsterbosch et al., 2018; Wang et al., 2015; Salehi et al., 2020; Kong et al., 2021). Subject-specific spatial features (i.e. spatial boundaries or parcel size) are shown to associate with behavior (Bijsterbosch et al., 2018) and functional state (Salehi et al., 2020). Individualized mapping of functional architecture is also a fundamental requirement of clinical procedures such as surgical planning and brain stimulation therapies (Lang et al., 2014). A range of automated techniques have been developed for deriving individual parcellations (Wang et al., 2015; Gordon et al., 2016; Chong et al., 2017), including a gradient-infused multi-session hierarchical Bayesian model (gMS-HBM) derived from resting-state fMRI (rs-fMRI) proposed by (Kong et al., 2021). This latter approach is currently the state-of-the-art (Kong et al., 2023) tool for delineating individualized, spatially localized functional parcellations and has shown to improve behavioral predictions from functional connectivity, demonstrating its utility in determining individualized functional parcels. However, application of this approach broadly relies on the availability and quality of rs-fMRI data. Of note, rs-fMRI acquisitions are not routine in clinical practice, and when available they may be brief in duration. The reproducibility and test-retest reliability of connectivity estimates have been shown to be compromised by the length of the rs-fMRI scan (Birn et al., 2013) rendering short rs-fMRI acquisitions ($<$10min (Gonzalez-Castillo et al., 2014)) of limited utility.

Over the last two decades, neuroimaging has provided the opportunity to study brain macrostructure non-invasively, in-vivo on diverse sample of individuals, resulting in a plethora of studies investigating the brain structure and its relation to function. The convergence of findings have shown that features extracted from structural MRI (T1- and T2-weighted images) can provide useful information regarding the functional specialization of the cortex. Across multiple modalities, structural heterogeneity is found to be linked to functional organization (Margulies et al., 2016; Suárez et al., 2020), and the organizing principle of cortical hierarchy (Llinares-Benadero and Borrell, 2019). For example, myelin maps are shown to capture areal variation in cytoarchitecture, cell-type distributions, and synaptic properties (Demirtaş et al., 2019), and a dominant neural axis of transcriptional variation in gene expression mapping, suggesting a common hierarchical organization for functional specialization (Burt et al., 2018). (Kurzawski et al., 2022) has shown relationships between BOLD response magnitude and cortical thickness, curvature, depth across subjects and datasets. Given the magnitude of the BOLD signal is often used as a surrogate for neural activity, this finding suggest that variation in BOLD response magnitudes across cortical locations reflects, in part, differences in anatomy. In (Fukutomi et al.,

2018), the authors show that cortical features do not only relate to functional, behavioral but tissue microstructure and geometry of fiber tract orientations. Cortical thickness and myelin maps (macrostructure) with the orientation dispersion index (ODI) and the neurite orientation dispersion (NDI) maps (microstructure) respectively exhibit striking similarities in their spatial patterns and distributions. The similarities between micro- and macro-structure can be attributed to the interdependencies within the brain to orchestrate function. In addition to the functional relevance of the structural features, T1/T2-weighted imaging from which the these measures are extracted, are performed more widely in clinical brain scans compared to resting-state fMRI. Therefore, to generate individualized parcellations in the absence of rs-fMRI data (or of sufficiently long rs-fMRI scans), we hypothesize that they could be inferred from more commonly acquired structural MRI scans.

In this work, we propose a novel way to infer individualized functional parcellations using only cortical surface measures reconstructed from commonly acquired T1-and T2-weighted MRI scans, by employing spherical convolutional neural networks. We train our models using cortical surface measures and evaluate the resulting functional parcels using rs-fMRI data. As ground truth labels, we use individualized functional parcellations from (Kong et al., 2021), which were derived from rs-fMRI using a multi-session Hierarchical Bayesian modeling approach. We evaluate the generalizability of our parcellations to out-of-sample subjects in terms of segmentation accuracy using Dice correlation coefficient. We also compare our proposed approach with a group-level parcellation (Schaefer et al., 2018) and with the previously defined rsfMRI-based individual-level parcellations (Kong et al., 2021) in terms of areal homogeneity (temporal similarity of vertices within a given parcel). And finally, we conduct a multivariate brain-behavior analysis using the predicted parcels. An earlier version of this work using only a subset of the data is presented in PRIME Workshop during MICCAI 2022. Now we extended this work utilizing the entire labeled dataset and explored multiple parcellation granularities, from coarsest or finest labeling schemes from the same individual labeling protocol (Kong et al., 2021).

The remainder of the paper is organized as follows. Section 2 briefly explains the dataset, preprocessing, model training and evaluation. In Section 3, the predicted labels are compared against population-based and ground truth individualized parcellations (derived from rs-fMRI) both qualitatively and through functional homogeneity. In Section 4, we discuss the implications and limitations of this work.

2 Methods

2.1 Dataset

The Human Connectome Project (HCP) is a large-scale data collection effort. The S1200 release consists of structural MRI (sMRI), resting-state fMRI (rs-fMRI) and behavioral measures (Van Essen et al., 2013). T1-weighted images (T1w) were acquired using a 3D-magnetization-prepared rapid acquisition with gradient echo

(MPRAGE) sequence and T2-weighted (T2w) images were acquired using a 3D T2-sampling perfection with application-optimized contrasts by using flip angle evolution (SPACE) sequence, both with spatial resolution of 0.7 mm isotropic (Van Essen et al., 2012). Resting-state functional MRI (rs-fMRI) images were obtained by multiband gradient echo-planar imaging with the following parameters: temporal resolution (TR) of 0.72 s, duration of 1200 frames per run (14.4 min), and spatial resolution of 2 mm isotropic (Van Essen et al., 2012). In this project, we used all 1029 subject cortical labels published by (Kong et al., 2021). for training. 729 subjects were used for training and 300 subjects were set aside for testing and subsequent analyses.

2.1.1 Preprocessing

Surface data was already preprocessed according to (Glasser et al., 2013) using both T1w and T2w images. In brief, the structural images were preprocessed (corrected for gradient nonlinearity, readout, and bias field; aligned to the "native" space and averaged when multiple runs were available; then registered to MNI 152 space using FSL's FNIRT). The native space images were used to generate individual white and pial surfaces (Glasser et al., 2013) using the Freesurfer and HCP pipelines (Fischl, 2012). Post FreeSurfer, native-mesh surfaces for individual subjects were registered using a multimodal surface matching (MSM) algorithm (Robinson et al., 2014) with MSMSulc to the Conte69 folding-based template (Van Essen et al., 2012). Cortical thickness, sulcal depth and curvature were calculated in FreeSurfer. Myelin content was estimated by dividing the T1w image by the T2w image, and mapped onto the cortical surface and corrected for the bias field (Glasser and Van Essen, 2011; Glasser et al., 2013). Finally, all surface data was mapped to a standard left/right symmetrical 32k reference mesh. Rs-fMRI was preprocessed according to (Smith et al., 2013). In brief, each fMRI scan is corrected for spatial distortions caused by gradient nonlinearity, head motion, and B0 distortion. Then, the data are registered to the T1-weighted structural image and transformed into 2mm MNI 152 space using a concatenation of various transforms and a structural-to-MNI nonlinear warp field. Global intensity normalization is applied to the entire dataset, and non-brain voxels are masked out. The resulting volume timeseries is then mapped onto the cortical surface using a ribbon-constrained approach that excludes noisy voxels. The timeseries are resampled onto a lower resolution registered standard 32k reference mesh with 2 mm average vertex spacing and smoothed with a 2 mm full-width at half-maximum (FWHM) surface smoothing. Subcortical gray matter voxels are also resampled using a 2 mm FWHM Gaussian neighborhood from individual FreeSurfer-derived subcortical parcels to a standard atlas set of subcortical parcels with 2 mm voxels. In addition to the original preprocessing, following (Kong et al., 2021), to reduce potential sources of artifact in fMRI data, we regressed out the global signal and head motion parameters, along with their temporal derivatives.

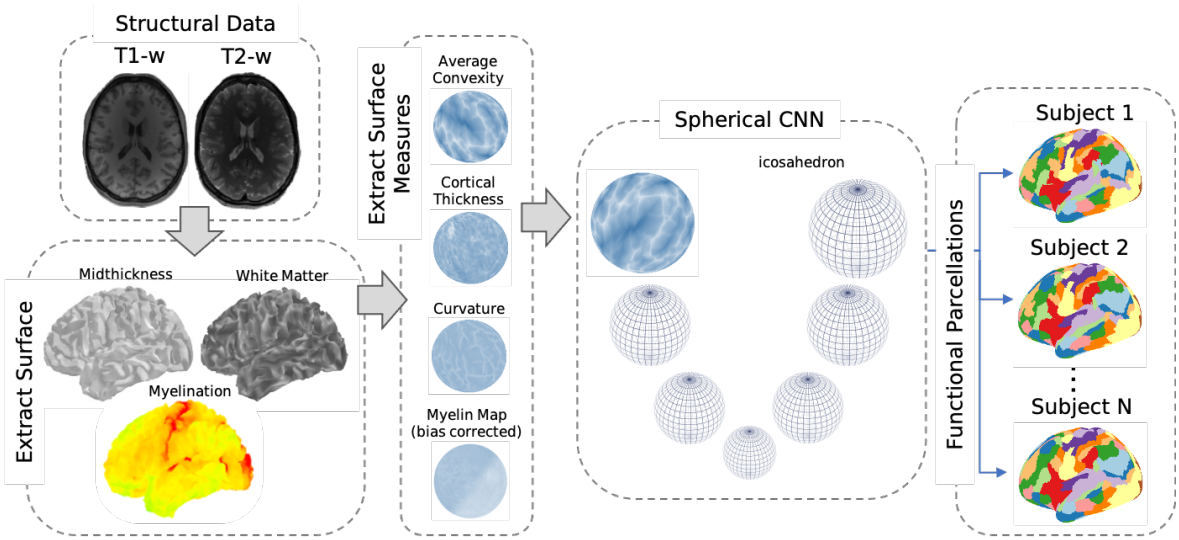


Figure 1: Learning framework

Figure 5.1: The pipeline demonstrates the process of obtaining individual-specific functional parcellations from structural data. Four surface measures, average convexity, curvature, cortical thickness, and bias corrected myelin map, are provided to the spherical convolutional neural network (CNN) as its input. The CNN then generates the desired output, which represents the functional parcellations specific to each individual.

2.1.2 Cortical surface features

For the cortical surface features, we used sulc (average convexity), curv (mean curvature), cortical thickness, and bias corrected cortical myelin maps. Curvature, in relation to the cortical surface, pertains to the geometric characteristics of its folding pattern. Negative values denote gyri (convexities), whereas positive values indicate sulci (concavities). Curvature is quantified as 1/r, with r representing the radius of an inscribed circle measured in millimeters. Thickness, measured in millimeters, represents the distance between the outermost boundary, near the cerebrospinal fluid, and the innermost boundary, near the white matter, of the gray matter. To estimate depth, six cortical surfaces were generated, evenly distributed between 10% and 90% of the distance from the pial surface to the boundary separating gray and white matter. Myelin maps were generated by dividing the T1w image by the T2w image.

2.2 Model 1

In the proof-of-context paper, we employed data augmentation to generate extra surface features through geometric transformations. Following (Parvathaneni et al., 2019), we computed decomposable deformation trajectories (ranging from rigid body alignment to more local non-rigid deformation) encoded by spherical harmonics coefficients (Lyu et al., 2019) and generated the intermediate deformation of each cortical feature as augmented data. For the segmentation task, the spherical CNN architecture by (Jiang et al., 2019; Parvathaneni et al., 2019)

was employed.

In this framework, briefly four cortical shape measures are provided as input, left hemisphere functional parcellations are predicted. The individual-specific parcellation from (Kong et al., 2021), consisting of 400 regions of interest (left, right total), were used as ground truth labels. Within the set of 150 (300 held-out test subjects) training subjects, 5-fold cross validation was used for model training, using a 60/20/20 split respectively for training, validation and testing by rotating partitions. We used cross-entropy loss and the ADAM optimizer, and networks were allowed to train up to 15 epochs. All networks were trained with learning rate 0.01 with decay of factor 0.1 and patience 2. The best models were saved based on the validation loss. A batch size of 2 was used at icosahedral subdivision level 6. The experiments were performed on an NVIDIA RTX 2080 GPU. Programs were implemented with Python using the Pytorch deep learning library.

2.3 Model 2

In this current version, we use spherical harmonics-based convolutional neural network (SPHARM-Net) (Ha and Lyu, 2022). This end-to-end model incorporates a constrained spherical convolutional (CSC) filter that can accommodate an infinite set of spectral components, eliminating the requirement for an intermediary rigid data augmentation process.

Here, differently from Model 1, two sets of parcellations (corresponding to individualized parcels across the cortex, described next) are predicted. The individual-specific parcellation from (Kong et al., 2021), consisting of 2 parcellation schemes i.e. 100 and 400 regions of interest, were used as ground truth labels. Within the set of 729 (1029 minus the 300 held-out test subjects) training subjects, 5-fold cross validation was used for model training, using a 60/20/20 split respectively for training, validation and testing by rotating partitions.

2.4 Generating and evaluating predicted parcellations on the test set

To evaluate the performance of our models, we utilized the set of 300 subjects that were withheld from the training phase. For each subject in this test set, 5 sets of segmentations (individualized parcellations) are predicted, corresponding to the 5 models (one from each fold) of the aforementioned 5-fold CV (that had been performed on the training set of 150 subjects). These segmentations are combined using majority vote to form one final segmentation. Segmentation accuracy is calculated using Dice coefficient, accuracy and Intersection-Over-Union (IoU, Jaccard Index):

$$Dice(A,B) = \frac{2|A \cdot B|}{|A| + |B|} \tag{5.1}$$

$$Accuracy(A,B) = \frac{|A \cdot B|}{|A|} \tag{5.2}$$

$$IoU(A,B) = \frac{|A \cdot B|}{|\max(A,B)|} = \frac{|A \cdot B|}{|A| + |B| - |A \cdot B|} \tag{5.3}$$

where A is ground truth parcel and B is predicted parcel. These three metrics were each averaged across the set of 200 parcels, resulting in a mean Dice, Accuracy, and IoU value per scan.

2.5 Assessment of homogeneity

To further evaluate the utility of the predicted parcellations in investigations of brain function, we assess the homogeneity (similarity) of time courses within a given parcel, based on the assumption that parcellations that follow an individual's functional boundaries ought to contain voxels with highly similar time courses. Therefore, higher homogeneity indicates better parcellation quality. For each subject, we computed homogeneity for all four runs separately (two fMRI sessions on two consecutive days). Vertex time-courses are extracted using a set of labels (i.e. Schaefer, Kong, or predicted parcellations). For each parcel, within-parcel homogeneity is calculated by averaging the pairwise Pearson's correlations between rs-fMRI time courses of all vertices within that parcel. An overall homogeneity value is then constructed by averaging across the within-parcel homogeneities of all parcels belonging to a parcellation scheme, adjusted for parcel size (Kong et al., 2021):

$$Homogeneity = \frac{\sum_{l=1}^{L} \rho_l |l|}{\sum_{l=1}^{L} |l|} \tag{5.4}$$

where ρ_l is the homogeneity of parcel l and $|l|$ is the number of vertices for parcel l.

3 Results

3.1 Model 1

Evaluation of the two proposed frameworks indicates that functional parcellations can be approximated solely from features derived from structural imaging data. Individualized parcellations derived from cortical surface measures on a held-out test cohort showed moderate agreement against multi-session hierarchical Bayesian-based ground truth labels derived from rs-fMRI data, with a mean Dice score of 0.740.

Figure 5.2a shows two example subject predictions, along with their ground truth labels, for subjects in the held-out test set. Figure 5.2b demonstrates the performance on the held-out partitions of 5-fold CV (within the training set of 150 subjects). Each data point in Figure 5.2b represents the average similarity score across all

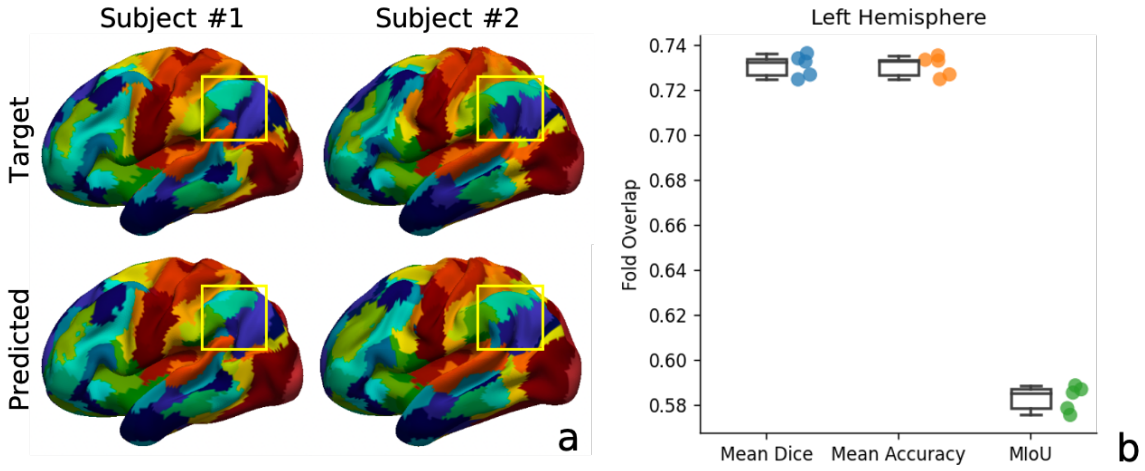


Figure 5.2: **Results from Model 1**. (a) Two example subjects were randomly selected for visual inspection. Predicted and ground truth labels are projected onto their respective inflated surfaces. The yellow box highlights that indeed the model was able to capture individual differences. (b) During training, 5-fold cross-validation is performed. Three separate scores, capturing the similarity of the predicted labels against ground truth, are reported separately on the test split for each fold. Both the mean Dice and mean accuracy are 0.730±0.004 across 5 folds.

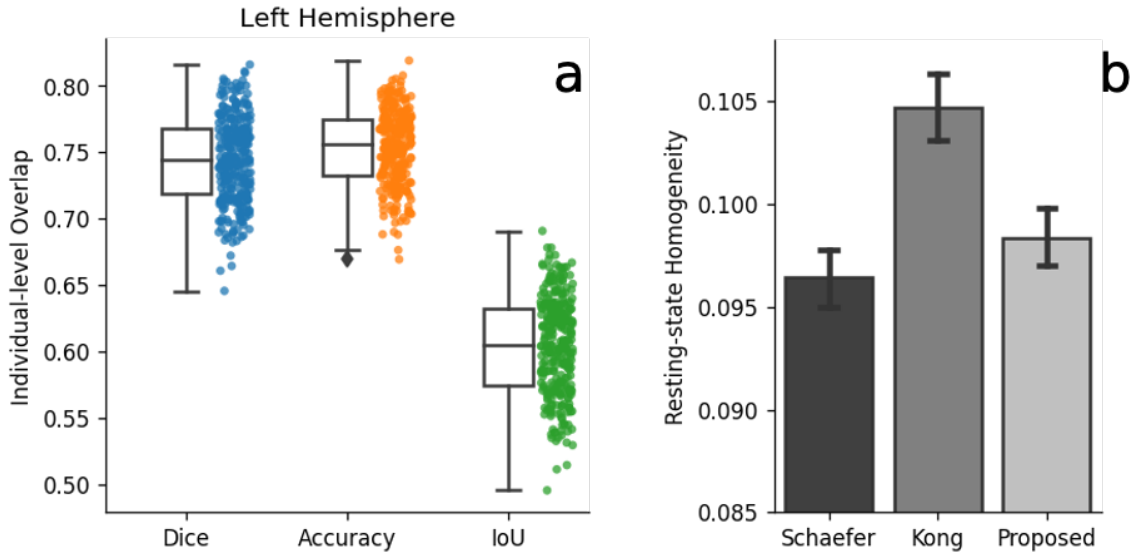


Figure 5.3: **Results from Model 1**. For each subject in the held-out test set: (a) similarity with ground truth parcellations was calculated using the same 2 metrics as in Figure 5.2. Scores recorded consisted of mean Dice = 0.740±0.032 and accuracy = 0.753±0.028; (b) 5 sets of parcellations are predicted, corresponding to the 5 models of the 5-fold CV (one from each fold). These parcellations were then applied to all four rs-fMRI runs. Resting-state homogeneity is calculated for Schaefer group-level, Kong individual-level (ground truth) and the predicted individual-level parcellation from the proposed model. Proposed method outperformed the group-level atlas, showing higher parcel homogeneity.

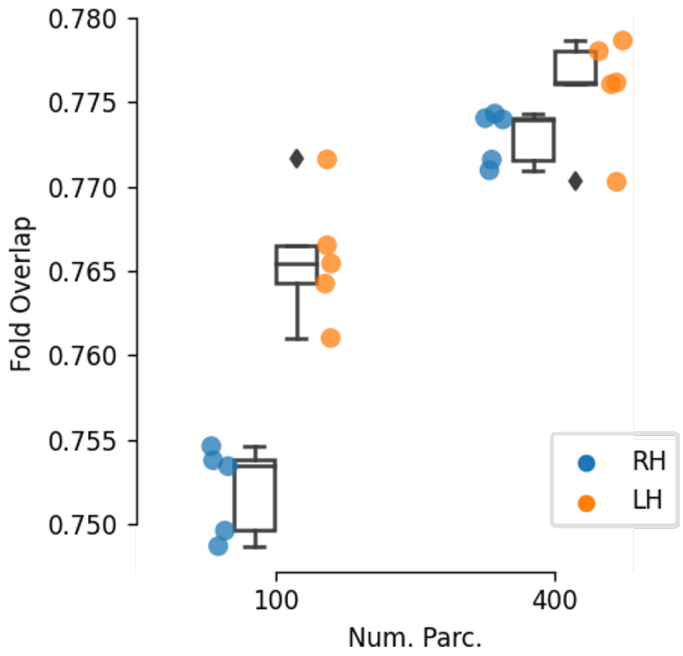


Figure 5.4: **Results from Model 2**. Similarly to the first experiments when training Model 2, 5-fold cross-validation is performed. DICE scores, capturing the similarity of the predicted labels against ground truth, are reported separately on the test split for each fold. Here we show, two parcellation schemes and models trained separately on right (RH) and left (LH) hemispheres.

parcels (and all subjects in a given test fold), using three different similarity metrics.

Results on a separate set of held-out test data are shown in Figure 5.3. As described above, for each subject in the test set (300 subjects) that was held out from training, 5 sets of parcellations are predicted, corresponding to the 5 models (one from each fold) of the aforementioned 5-fold CV that had been performed on the training set of 150 subjects. These segmentations are combined using majority vote to form one final segmentation. Figure 5.3a shows the similarity between predicted and ground-truth parcellations, quantified with the same three metrics, for each of these 300 test subjects. Figure 5.3b shows the temporal homogeneity of the parcellations, comparing across (i) the group-level parcellation (Schaefer, which uses the same parcels for all subjects), (ii) the individualized parcellations from rs-fMRI data (Kong; here used as ground truth), and (iii) the parcellation resulting from the proposed method. The proposed method outperformed the group-level parcellation, indicating that it is able to learn information about individualized functional boundaries.

3.2 Model 2

4 Discussion and Conclusion

We demonstrate a novel way of inferring individual-specific functional parcellations from cortical features derived from structural (T1w, T2w) MRI. Functional parcellations are most commonly derived using rs-fMRI data, as coherent activity across rs-fMRI time courses has been shown to correspond with subject-specific functional networks and behavioral measures more closely than those defined from traditional anatomic subdivisions (Dadi et al., 2020). By contrast, the present work indicates that structural brain images contain information about these functional boundaries, suggesting that functional parcellations could be inferred without the need to measure brain activity time courses with a resting-state fMRI scan.

Indeed, compared with the population-based (Schaefer et al., 2018) atlas parcellation, the proposed approach estimated more homogeneous regions during resting state, in agreement with the claims of (Kong et al., 2021). Of note, however apparent, the aim of this approach was to achieve individual precision (i.e., to outperform the population-based atlas) rather than to exceed the quality of the individualized rs-fMRI parcellations that were used as ground truth for training. Therefore, any improvement over a group-level parcellation (that is derived from fMRI data) using solely structural features exceeds the current expectations of the neuroimaging community.

Importantly, the ability to delineate functional parcellations solely from structural MRI has important implications both in clinical applications and research. In preoperative planning and brain stimulation therapies, the ability to precisely localize functional boundaries in an individual will enable individualized treatment and can potentially improve patient outcome. In addition, an increasing number of studies suggest that individual differences in the size or shape of functional parcels can be related to cognition and behavior (Bijsterbosch et al., 2018; Salehi et al., 2020), and may therefore present informative non-invasive biomarkers. Together with the aforementioned studies, our ability to extract functionally relevant units from anatomic images underscores a prominent link between structure and function.

Recent work by (Salehi et al., 2020; Boukhdhir et al., 2021) has suggested that the brain parcellations reconfigure dynamically and across different cognitive states. Accordingly, identifying parcellations within their dynamic states would improve reproducibility and quality over ”static” parcellations. Similar to the argument of Kong et al. (2021), our study does not challenge this notion. It is possible that computationally stable units across the brain could potentially reconfigure to form dynamic networks that change moment-to-moment, and the degree to which cortical shape features can predict parcellations corresponding to different brain states is an area of future investigation. Our study, and that of Kong et al. (2021)., may be delineating parcellations that are most stable – on average – in an individual.

5 Future Work

One limitation of this study is that when working with parcellations of varying numbers, the models were not able to learn from missing labels as effectively. A potential future direction could involve developing an agnostic model that accommodates different number of parcels for each subject. Considering the diversity of the biological substrate this approach could have the potential to yield more plausible parcellations. As this study used only resting-state data from healthy young adult subjects, another potential avenue of investigation could be to examine the generalizability and impact of this approach on task fMRI data and in-patient populations.

CHAPTER 6

TractEM: Evaluation of protocols for deterministic tractography white matter atlas

Reproducible identification of white matter pathways across subjects is essential for the study of structural connectivity of the human brain. One of the key challenges is anatomical differences between subjects and human rater subjectivity in labeling. Labeling white matter regions of interest presents many challenges due to the need to integrate both local and global information. Clearly communicating the manual processes to capture this information is cumbersome, yet essential to lay a solid foundation for comprehensive atlases. Segmentation protocols must be designed so the interpretation of the requested tasks as well as locating structural landmarks is anatomically accurate, intuitive and reproducible. In this work, we quantified the reproducibility of a first iteration of an open/public multi-bundle segmentation protocol. This allowed us to establish a baseline for its reproducibility as well as to identify the limitations for future iterations. The protocol was tested/evaluated on both typical 3T research acquisition Baltimore Longitudinal Study of Aging (BLSA) and high-acquisition quality Human Connectome Project (HCP) datasets. The results show that a rudimentary protocol can produce acceptable intra-rater and inter-rater reproducibility. However, this work highlights the difficulty in generalizing reproducible results and the importance of reaching consensus on anatomical description of white matter pathways. The protocol has been made available in open source to improve generalizability and reliability in collaboration. The goal is to improve upon the first iteration and initiate a discussion on the anatomical validity (or lack thereof) of some bundle definitions and the importance of reproducibility of tractography segmentation.

1 Introduction

1.1 Three-dimensional atlases (GM/WM)

Human brain atlases have had a transformative role in the study of neuroscience. Evolving neuroimaging, image processing, and analysis techniques are framing the way that we understand brain anatomy. Modern brain atlases (Amunts et al., 2013; Glasser et al., 2016; Hawrylycz et al., 2012) combine population atlases with multi-modal structural and functional information. These atlases provide a detailed representation of the anatomy and are a fundamental component of the field of image analysis. Although, most commonly used modern approaches tend to treat white matter (WM) as essentially homogeneous and each region is labeled as a single WM structure. Most current atlases are incomplete and do not contain detailed information about the full spatial extent of WM structures. A whole-brain WM atlas, labeled on a series of subjects using equivalent methodologies, tools and protocols, would be a fundamental resource for integrative and comprehensive mod-

ern brain atlases. The effort to create atlases is not new. The importance of white matter atlases arises from the information that can be inferred from them. White matter connectivity stores substantial information about functional neuroanatomy (Greicius et al., 2009; Skudlarski et al., 2008; Sporns et al., 2005; Van Den Heuvel et al., 2008), longitudinal changes (Resnick et al., 2003), brain morphology (Lawes et al., 2008) abnormalities and their cognitive correlation (Gunning-Dixon and Raz, 2000), critical periods for neural development (Anderson et al., 2011), recovery and plasticity (Dayan and Cohen, 2011; Jiang et al., 2006), and cognitive functioning (for Research into Nervous et al., 1998). Nonetheless, existing white matter atlases are limited to population-averaged from extremely time-consuming manual delineation or they lack comprehensive coordinate information about white matter structures and specific details about finding anatomical landmarks. They are also often limited to a 3D representation where each position (voxel) represents only one label/class, which is an oversimplification of the true architecture of the brain (Adluru et al., 2016; Besseling et al., 2012; Mori et al., 2008; Wakana et al., 2007).

1.2 Four-dimensional atlases (WM)

We know from ex-vivo dissections that white matter pathways cross each other and merge and/or fan in a complex manner. This challenges the usage of a 3D atlas limited to one label per voxel (Hansen et al., 2021) which greatly increases the complexity to manually label the pathways. For this reason, manual labeling of pathways from diffusion MRI tractography using virtual dissection (Catani et al., 2002; Mori et al., 1999) is a useful approach to visualize the spatial relationship between anatomical landmarks. By using inclusion and exclusion regions of interest (ROIs) to isolate streamlines belonging to known WM pathways, a "4D" atlas can be made, where any voxels can be associated to any number of labels (Catani and De Schotten, 2008; Hansen et al., 2021; de Schotten et al., 2020; Yeh et al., 2018). While there are expert-evaluated white matter atlases (Catani and De Schotten, 2008; Essayed et al., 2017), they lack protocols detailing how to delineate white matter structures (Chenot et al., 2019; Mandonnet et al., 2018). Additionally, most well-known 4D WM atlases were generated using diffusion tensor imaging, which has been shown to have limitations in defining crossing fiber and complex regions (Jones and Cercignani, 2010; Oouchi et al., 2007), whereas a number of more advanced techniques can successfully overcome these limitations (Dell'Acqua and Tournier, 2019; Jeurissen et al., 2014; Tournier et al., 2019). More often than not, WM atlases are intended to be used as a tool, not a guide to perform a delineation on a new dataset. Atlases created with the intentions of being used with automated methods often rely on manual delineations or opaque (if existing at all) anatomical definitions (Garyfallidis et al., 2018; O'Donnell and Westin, 2007; Wasserthal et al., 2018; Yeh et al., 2018; Yendiki et al., 2011; Yoo et al., 2015). The lack of explicit and detailed manual labeling protocols that are robust to human

subjectivity must be addressed.

1.3 Reproducibility of protocols

Atlases are crucial tools to investigate WM pathways at the scale of a population. However, an atlas cannot be constructed without delineating subject-specific WM pathways. While WM atlases are a worthwhile end goal, we must first take a step back and investigate the basic building block of any WM atlas: subject-specific segmentation protocols. Subject-specific delineation can be useful in multiple situations: 1) pathologies and lesions can make the mapping of subjects to the atlas difficult (De Schotten et al., 2011) if a specific population is studied and their anatomy is somewhat different from the population used for the atlas (Avants et al., 2010; Schmidt et al., 2018) or 3) simply to provide a framework to classify WM structure (Mandonnet et al., 2018; Panesar and Fernandez-Miranda, 2019). Two of the assumptions of a WM labeling is that all labels are anatomically valid and the same label in two datasets are in fact representing the same underlying structure. Despite often being taken for granted, these assumptions can be challenged. The potential implications of these assumptions being false would diminish the usability of the atlas in clinical application, for example. If anatomical validity is an important criterion in theory, reproducibility of its execution is an important criterion in practice. This work focuses solely on the second assumption, identical labels should represent the same structures. An anatomically valid and well-detailed protocol is great, but if it cannot be executed accurately by a rater, if each execution is too variable or if each rater interprets it differently, the utility of the protocol will be diminished or even curtailed completely.

1.3.1 Intra-rater

The first layer of reproducibility to be analyzed is intra-rater: Can a rater following the same instructions twice on the same datasets achieve similar results? This layer is crucial because we can hypothesize that the interpretation of the instructions should not vary within a single individual. Thus, this quantifies the variability in the execution of the instructions.

1.3.2 Inter-rater

The second layer of reproducibility to be analyzed is inter-rater: Can two (or more) raters following the same instructions on the same datasets achieve similar results? This layer will quantify the variability in interpretation of the instructions. The observed inter-rater reproducibility score is lower than intra-rater (since it encompasses both the variability within and across raters).

1.3.3 Inter-subject

This last layer is not actually due to the protocols or raters directly. However, as mentioned earlier it encompasses both the intra and inter-rater variability. Analyzing this level will provide information related to biological differences in the underlying anatomical structures. If the protocol's variability observed is too high at the intra or inter (or both) rater stage, this measurement error will drastically reduce the ability to interpret changes observed across subjects. This is the crucial point of this work: not only must a consensus be reached on anatomical definition, but the interpretation and execution of this consensus must be quantified in order to assess how useful a segmentation protocol is in practice.

1.4 Framework for protocol creation and evaluation

As shown in (Boccardi et al., 2011; Rheault et al., 2020; Schilling et al., 2021a), simply creating a protocol that can be followed "closely" is a difficult task. Segmentation protocols are more common in 3D medical imaging or even 2D natural images, but tractography is a much more complex 3D representation. A lot of work is necessary to reach a set of do's and don'ts to clearly design, evaluate and use bundle segmentation protocols in clinical settings. The logic behind our investigation is fourfold: 1) Can instructions from a protocol be followed by raters? 2) How can we quantify how reproducible a protocol is? 3) Can we identify portions of the protocol that increase or decrease reproducibility? 4) What are the next steps to take to improve the quality of a protocol or to create a better protocol in the future? In this work we approach and quantify the first two and provide insights into the last two. The framework used to obtain bundles of interest (TractEM) is a first iteration of an open/public protocol; the reproducibility analysis presented in this work should be seen as a baseline for future improvements to TractEM. This work should not be perceived as an attempt to demonstrate the superiority or inferiority of the protocol, nor as claiming to be a valid, definitive or even accepted/agreed upon set of anatomical definitions. This work is rather, in our opinion, a necessary step to generate open discussion about the anatomical definitions of white matter pathways, how to segment them using a protocol, and how to evaluate results.

2 Methods

2.1 MRI acquisition

The protocol was developed on datasets obtained from the Baltimore Longitudinal Study of Aging (BLSA) (Ferrucci et al., 2008) and Human Connectome Project (HCP) (Van Essen et al., 2013). The BLSA dataset used in this study contains 10 subjects with ages ranging between 57-77 years old. The data were acquired after written informed consent and institutional review board approval and accessed in de-identified form.

Each session included a T1-weighted structural MP-RAGE (number of slices = 170, voxel size = 1.0x1.0x1.2 mm3, reconstruction matrix = 256x256, flip angle = 8 degrees and TR/TE = 6.5ms/3.1 ms). Diffusion data were acquired with a 3D spin-echo diffusion-weighted EPI sequence (TR/TE = 7454/75 ms). Each acquisition consisted of an initial b0 and 32 diffusion weighted volumes all with the same b-value of 700 s/mm2 (number of slices = 170, voxel size = 0.81x0.81x2.2 mm3). The HCP dataset used in this study contains 10 healthy subjects with no known history of neuropathological and psychiatric diseases, ages ranging between 26-36 years old. Diffusion data were acquired with a 3D spin-echo diffusion weighted EPI sequence (TR/TE = 5520/89.5 ms). Each diffusion acquisition consisted of six b0 and 90 diffusion weighted volumes all with 3 shells of b=1000, 2000, and 3000 s/mm2 interspersed with an approximately equal number of acquisitions on each shell within each run (number of slices = 111, voxel size = 1.25x1.25x1.25 mm3). Data were pre-processed to correct the effects of gradient nonlinearities on the b-values and b-vectors for each voxel (Glasser et al., 2013). The choice of two different datasets was to provide insight for the protocol's sensitivity to input data (SNR, spatial and angular resolution, reconstruction methods, etc.) and robustness to age related changes in the WM structures to observe how these variables affect reproducibility.

2.2 MRI preprocessing

Susceptibility correction (Andersson et al., 2003), eddy current correction techniques (Andersson and Sotiropoulos, 2016), and b0 signal normalization were applied to the diffusion data as a preprocessing step. T1 weighted images were co-registered to b0. DWI and T1w were moved to Talairach space (Jenkinson and Smith, 2001) using an affine transformation from ICBM152-space and the technique from (Lancaster et al., 2007). The affine transformation was applied to the DWI and T1w, and the appropriate corresponding adjustment was made to the b-vectors. DTI metrics such as FA and RGB were computed using DSI-Studio (http://dsi-studio.labsolver.org). For the orientation distribution function, two high angular resolution diffusion imaging (HARDI) models were selected; for characterizing multiple directions per voxel and providing enhanced ability to resolve crossing fibers (Descoteaux et al., 2007; Tuch et al., 2002). For the BLSA dataset, we fit 4th order spherical harmonic Q-ball. QBI uses a single shell (single b-value) to estimate the dODF to represent the fiber orientation distribution. DSI-Studio adopts the spherical harmonics based QBI reconstruction method. For multi-shell HCP data, generalized Q-sampling (GQI) reconstruction was used (Yeh et al., 2010). GQI in addition to being capable of resolving crossing fiber issues, supports multi-shell (multi-b-value) data and works best with fairly dense sampling of directions to similarly estimate the dODF.

2.3 Study design

For intra-rater, 5 raters were recruited to perform segmentation of all bundles on the same dataset, twice. For association and projection pathways, segmentations were performed only on the left hemisphere. In addition, the inter-rater variability was evaluated with an additional 10 raters and every dataset was traced at least twice by different raters (see Figure 6.1). Raters have no prior experience with bundle segmentation from tractography data (virtual dissection) to assess how the TractEM protocols can be learned from instruction only. Evaluation of the protocol was performed using 20 subjects (10 HCP, 10 BLSA) with a wide age range to evaluate robustness to heterogeneity in WM structures. The pre-processed data was provided so it could be directly input into DSI-Studio.

2.3.1 Protocol description

The TractEM project is a tractography-based whole-brain protocol (https://my.vanderbilt.edu/tractem/); in its current version it is informed by the EVE atlas (see Figure 6.2). The goal was to create a first release and by soliciting feedback on both label quality and protocol definitions, the protocol could be iteratively refined.

For example, many pathways utilize seed and/or inclusion ROIs using the orientation and locations visualized directly in the EVE atlas (see Figure 6.2), with additional inclusion/exclusion regions used to capture the bundle as completely as possible while also removing false positive streamlines as some in Figure 6.3. A total of 53 unique pathways and 8 subject-specific lobar regions (defined by a multi-atlas segmentation algorithm (Huo et al., 2016) were selected such that most of the brain was accounted for. Accounting for bilaterality, there are 7 commissural, 11 association, 7 projection pathways and 5 within the brainstem. All pathway descriptions were directly inspired from the EVE atlas.

In this manuscript, the complexity of the protocols is referred to as simple and complex tracts. If a protocol contains only a single seed region, e.g., genu of the corpus callosum (GCC), we refer to it as a simple bundle. If a bundle's definition includes multiple seed or ROIs in combination, this is referred as a complex bundle, e.g., inferior longitudinal fasciculus (ILF) (see Figure 6.3).

The current version of the TractEM protocol or the description of some segmentations can be debated. However, the logic behind our investigation remains the same: 1) Can the protocol's instructions be followed by raters? 2) How can we rigorously assess how reproducible a protocol is? 3) Can we identify specific aspects of the protocol that increase or decrease reproducibility? 4) What are the next steps to take to improve protocol

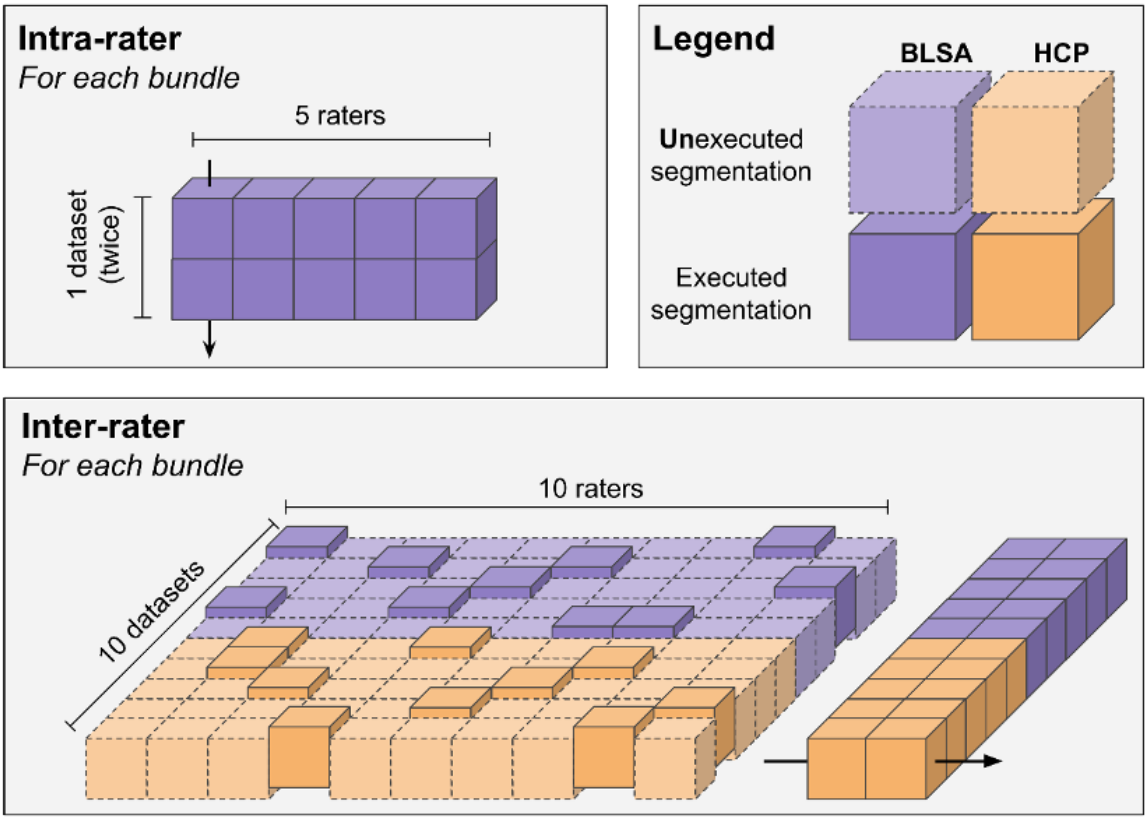


Figure 6.1: Visual representation of our study design. For intra-rater, all 5 raters had to segment all bundles on the duplicated BLSA dataset that was provided. For inter-rater, 10 raters had to execute only a subset of all combinations of datasets/bundles. The above schematic represents a perfectly balanced workload for all raters and between BLSA and HCP. However, in reality some raters had more segmentation to perform (either on more datasets or more bundles or both). This planned division of labor ultimately leads to each bundle being segmented twice for each dataset.

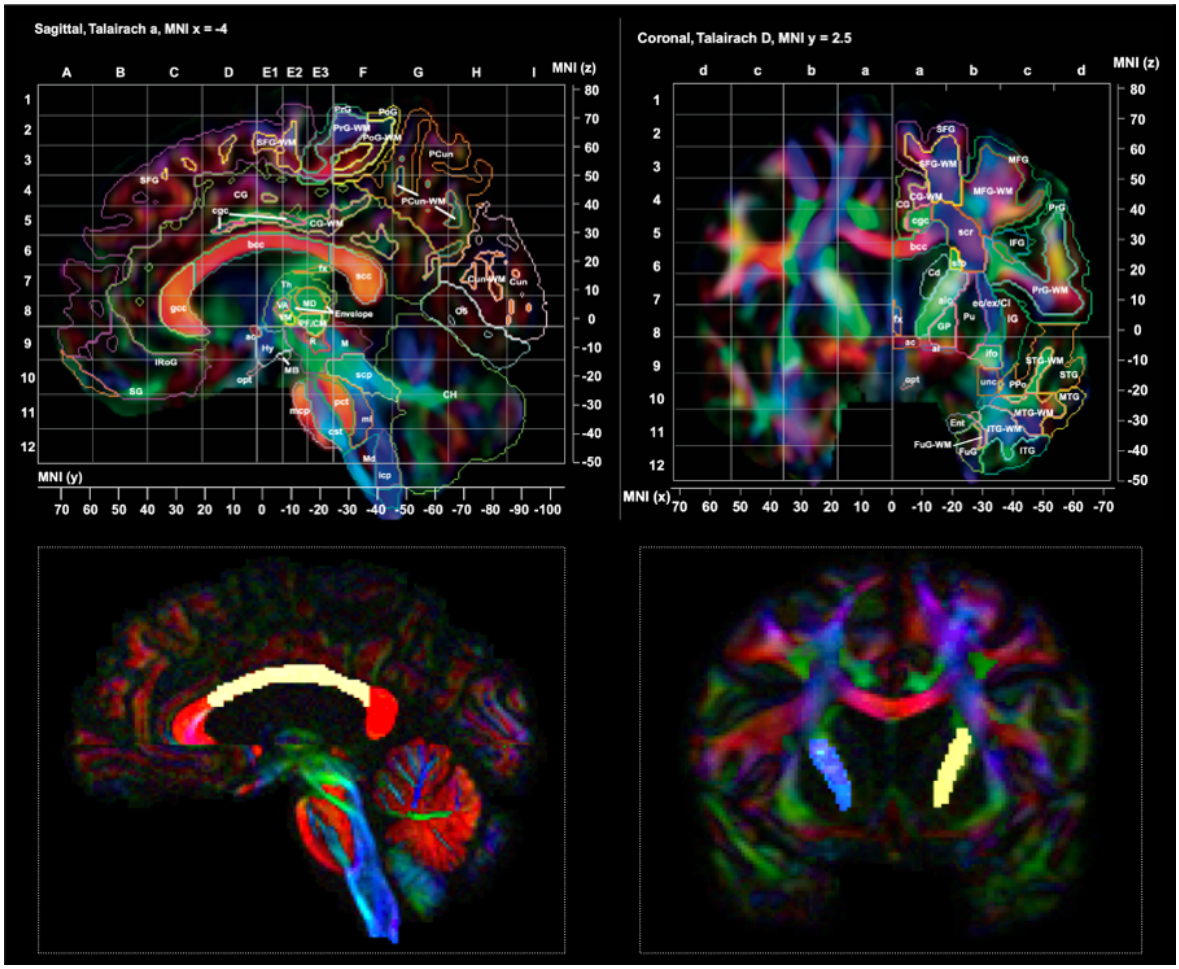


Figure 6.2: Most pathways were defined using the EVE atlas as a reference, with the aim to create a protocol that non-experts in anatomy and tractography can follow intuitively. We again stress that we are not suggesting these are optimal protocols for a given pathway, but rather a rudimentary guide of how one may delineate pathways given knowledge of EVE-based regions.

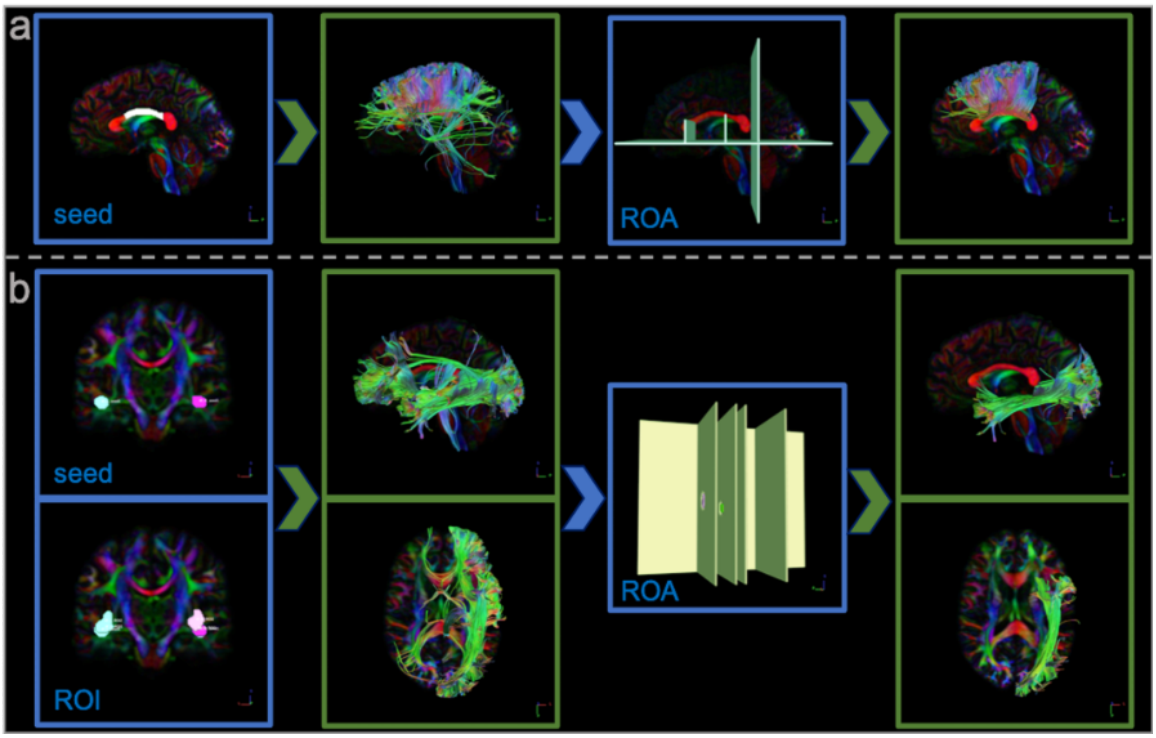


Figure 6.3: Difference between simple or complex bundle as defined in this work. Simple bundles (a) do not use regions of interest to guide their path. Only seed and a region of avoidance are necessary. Complex bundles (b) require seeding region and one, or more, region of interest. General direction is enforced by the regions of interest.

quality or alternatively, to create a better protocol in the future? The current work acts as a baseline, the first iteration, of an open and incremental protocol for multiple bundles' segmentation.

2.3.2 Task execution in DSI-Studio

DSI Studio is a tractography software that reconstructs WM pathways. It is a bundle-specific, ROI-based deterministic tractography approach (Yeh, 2020) to reconstruct, making it a tool of choice at the time of the study design. However, we acknowledge that this decision limits the scope of the baseline results presented in this work. Tracking parameters were chosen carefully to reduce the number of false positive fibers and to ensure that the tracking does not suffer from a premature termination, these parameters apply to all bundles in the TractEM protocol. Termination criterion is normalized quantitative anisotropy with a threshold of 0.1 and is used to ensure the maximum number of fibers are produced. Smoothing was set to 1, seed initialization limited to 100K for each bundle, and streamline length constraints were 30-300 mm for the simple and complex bundles. Angular and quantitative anisotropy (QA) thresholds are the determining parameters of how the fibers are tracked from the region of interest. The QA threshold was set to default and the angular threshold was set to zero, therefore stopping criterion is not determined by this parameter. The tracking algorithm is streamline (Euler), seed orientation is primary, and the direction interpolation is set to trilinear. The random seed generator was set to a constant to ensure tracking results are identical and reproducible.

2.4 Reproducibility analysis

2.4.1 Data representation

To assess the reproducibility of the TractEM protocol, Dice coefficient, weighted Dice coefficient and bundle adjacency distance were utilized. Reproducibility analysis is essential to validate robustness. The reader should note that the concept of ground truth does not apply to this study, instead each pair was compared with each. Streamlines were converted to a binary image representation (each voxel crossed by a streamline is set to 1) or a density image (a voxel crossed by N streamlines is set to N).

2.4.2 Reproducibility score / variability metrics

Dice coefficient (Dice, 1945) is a quantification of overlap between two sets, the resulting value is between 0 and 1. When applied to a binary image, it quantifies morphological similarity and overall volume agreement. The weighted Dice coefficient (Cousineau et al., 2017) is a measure of agreement taking into account streamline density. This way outliers and spurious streamlines will not have a large impact on the measure. The bundle adjacency (Garyfallidis et al., 2015) is another measure of spatial agreement. Bundle adjacency does not take into account streamline density. To reduce the impact of outliers, spurious streamlines or streamlines offset by

a few millimeters a simple distance (in mm) is used. Between pairs of segmentation, each voxel (occupied by a streamline) in the first segmentation is mapped to their closest voxel in the second segmentation and vice-versa. This way overlapping voxels are considered zeros, while the non-overlapping voxels are assigned the distance (in millimeter, mm). With this measure, values closer to 0 mm are desired. These three similarity metrics are used to quantify intra-rater and inter-rater reproducibility scores for each bundle of the TractEM protocol. While individually useful, it is best to look at the three simultaneously. The easiest scenario to understand occurs when the Dice (and weighted Dice) coefficient are high with a low bundle adjacency. This indicates similar spatial overlap, similar spatial distribution of density with rare outliers. However, a relatively high Dice with a very high bundle adjacency could indicate an overall good volume overlap with a few spurious streamlines extending far beyond the "cores". Another example would be two bundles with a high Dice and a low weighted Dice, this scenario would indicate a good spatial overlap, but their "cores" (where density is highest) are not overlapping.

3 Results

From the 10 raters, we received close to 2600 manually identified bundles. On average, an entire protocol took 6 hours to complete. This includes learning how to use DSI-studio, familiarization with the instructions for each bundle of interest and the segmentation itself (drawing the ROIs). Most raters did not have to perform the segmentation of all bundles for a given subject, instead they performed a few (random) bundle segmentations on a few subjects. Raters were allowed to use any version of DSI-Studio. After all the data were gathered, to reduce the variability due to DSI-Studio version differences and to ensure that the data was free of human errors (naming convention, saving issues, etc.), we performed the tracking for all subjects using the provided manual labels. If the tracking did not result in the same streamline submitted, this specific task was given to another rater and manually relabeled to replace the missing/inconsistent data. Some errors due to human fallibility were left unfixed in the dataset. Non-exhaustive list of issues: 1) some raters sought placement guidance from other raters, resulting in very similar labels, 2) some raters used the same labels for different subjects, 3) some raters did not closely follow the protocols, 4) some raters delineated the same-subject bundle more than once by accident.

Quality assessment is a crucial step to ensure validity of the submitted bundles. An automated PDF report was generated for each bundle. It includes streamline density visualization overlaid on the subject's FA images for all three views: sagittal, coronal, axial, and a 3D representation of the streamline. This tool served as a quick visual sanity check to see if the assembled bundle appeared consistent with the expected morphology. Its main use was to check if the rater interpreted the protocol correctly (obvious mistakes, saving issues, etc.).

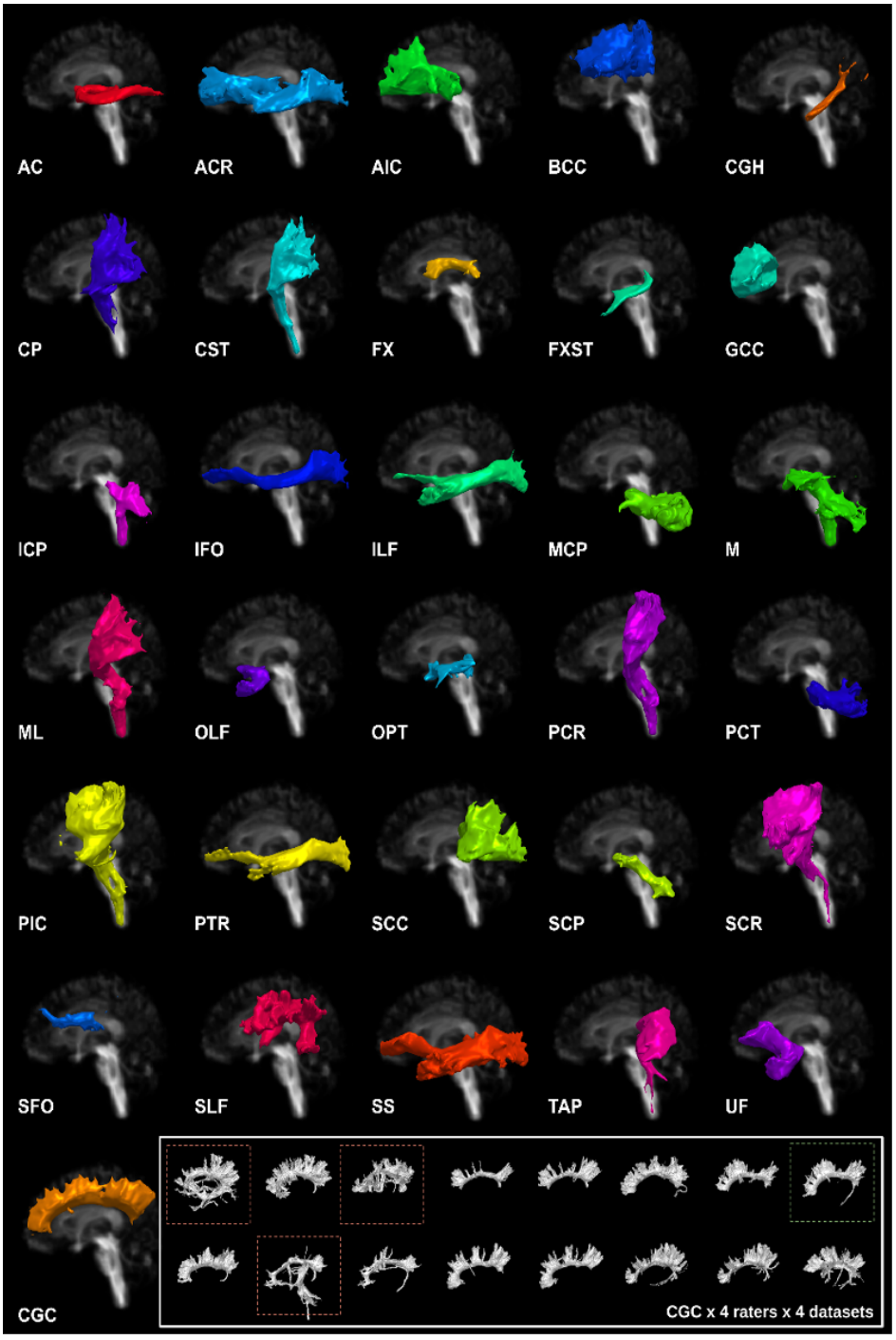


Figure 6.4: Average shape and position of all bundles that are part of the TractEM protocol (association and projection pathways are shown only for the left hemisphere). The average was computed using segmentation from 4 raters and 4 datasets (BLSA) using a majority vote. This highlights the general agreement on the shape, but as seen in the lower right vignette the variability of each individual segmentation can be quite extreme.

3.1 Intra-rater reproducibility

Five Raters were assigned to test all protocols (except lobar bundles) on a given BLSA subject. The results have shown acceptable homogeneity of results on duplicate datasets. However, while the median is an 'acceptable' level, the interquartile range is high. As seen in Figure 6.5 the small sample makes the distribution sensitive to extreme scores. No single rater was responsible for very high or very low reproducibility scores. For each tract, medians and standard deviations were plotted for both datasets. HCP has shown Dice > 0.65, Weighted Dice > 0.85 and Bundle Adjacency < 2 mm reproducibility on most bundles. BLSA has shown Dice > 0.6, Weighted Dice > 0.8 and Bundle Adjacency < 2 mm. We defined the pathways during the TractEM development stage as lobar regions, simple and complex bundles. Lobar regions exhibit very high reproducibility across both datasets, all lobar bundles have Dice above 0.9 (above 0.99 Weighted Dice). Simple bundles do not show increased reproducibility over complex bundles. Pathways were categorized per type: lobar, brainstem, commissures, associations, projections. It is important to mention that in our nomenclature, the brainstem category encompasses bundles that are within the brainstem regions and the projection bundles encompass all long-ranging bundles with a general up-down orientation. Except for the lobar category, which has a high reproducibility score, there is no association between the pathways' types and reproducibility.

3.2 Inter-rater reproducibility

The scores were calculated by averaging similarity across all subjects for a corresponding bundle. For each tract, medians and standard deviations were plotted for both datasets. HCP has shown Dice > 0.65, Weighted Dice > 0.85 and Bundle Adjacency < 2 mm reproducibility on most bundles. BLSA has shown Dice > 0.6, Weighted Dice > 0.8 and Bundle Adjacency < 2 mm reproducibility on most bundles. It is important to mention that median scores do not accurately represent the performance of the raters, as seen in Figure 2.6-8 the interquartile ranges show that segmentation tasks are prone to major interpretation/execution discrepancy.

3.3 Reproducibility score and data quality

The difference in data quality influences pathway reconstruction. Our hypothesis was that better acquisition, such as HCP, would positively impact the reproducibility score. However, the difference between both datasets was not statistically significant for most bundles (see Table 6.1). 24 out of 35 bundles did not show increased

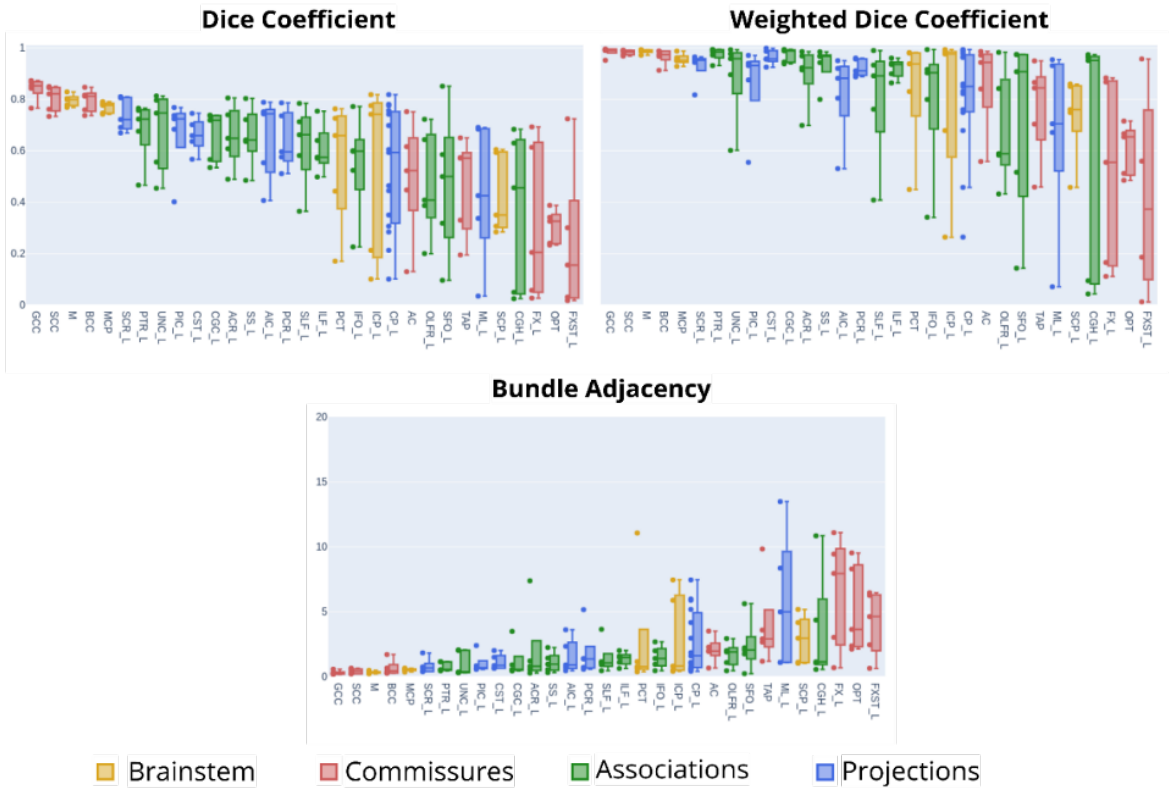


Figure 6.5: Intra-rater reproducibility metrics across all pathways. The pathways are ordered from the highest Dice Coefficient to the lowest. Coloring represents categories (brainstem, commissures, etc.) described earlier. No particular type of bundle obtained a higher or lower score. Due to the small sample size, outliers are difficult to identify and lead to a high interquartile range of the boxplots.

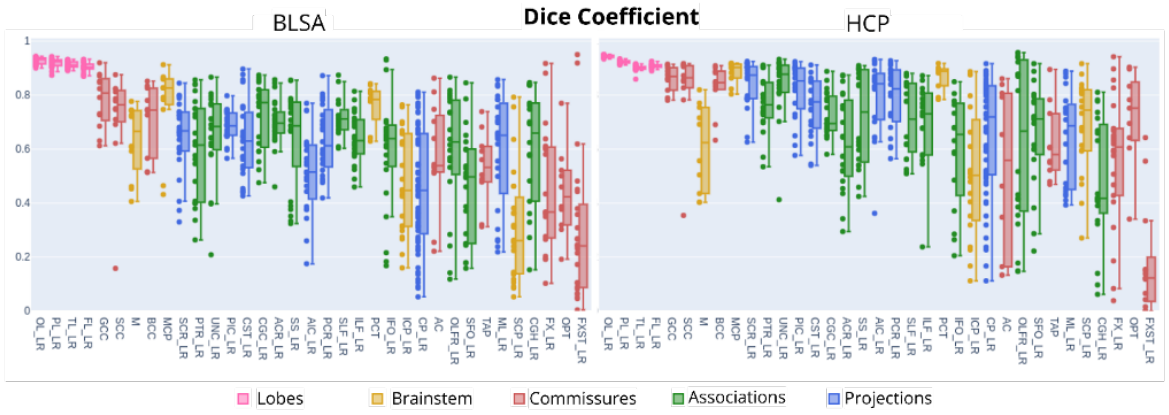


Figure 6.6: Dice Coefficient across all pathways for both datasets. All bundles with identical left and right descriptions were fused for the analysis. Pathway ordering and coloring are the same as in Figure 5. Low average or large interquartile ranges expose the lack of consistency in results (in terms of overall volume).

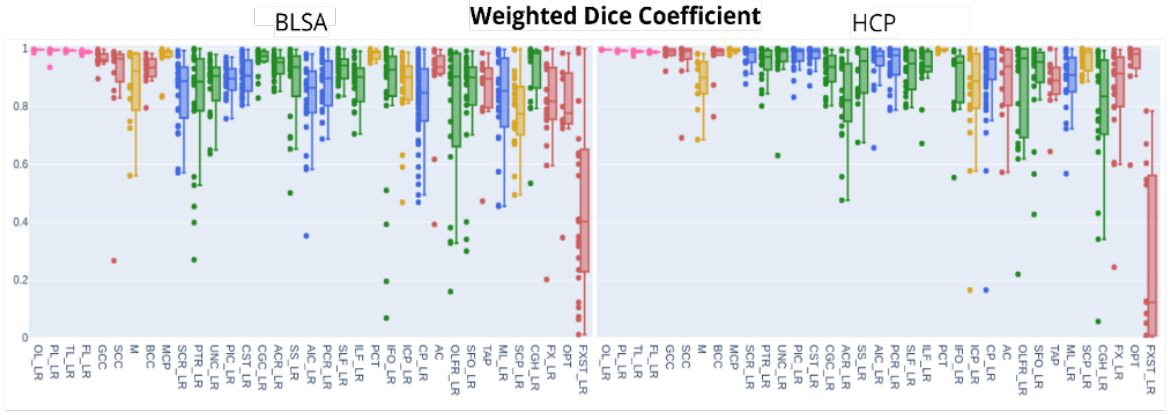


Figure 6.7: Weighted Dice Coefficient across all pathways for both datasets. Pathway ordering and coloring are the same as in Figure 5. High average and smaller interquartile range show that bundles' cores (most dense portion of a pathway) are more consistently obtained.

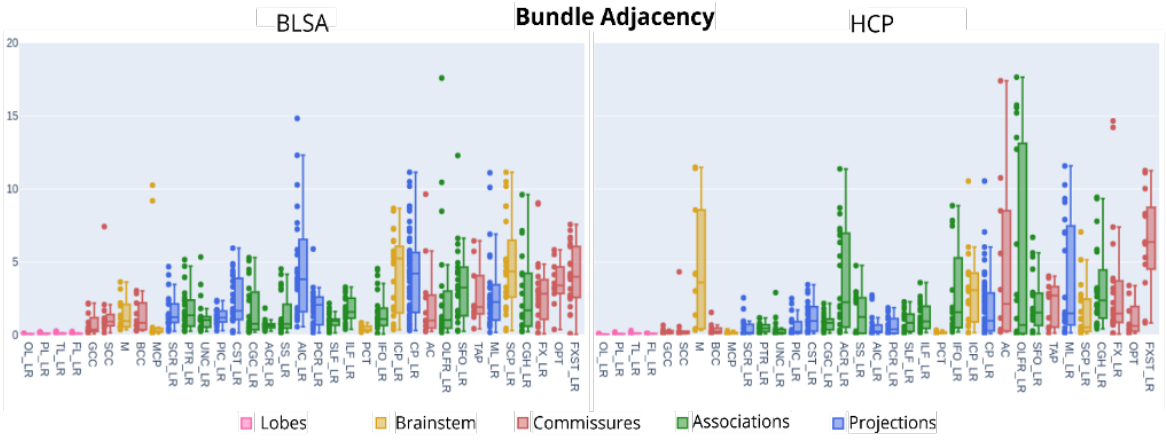


Figure 6.8: Bundle adjacency across all pathways for both datasets. Pathway ordering and coloring are the same as in Figure 6.5. Medium average and smaller interquartile range shows that segmentations were producing similar results, but all with various levels of outliers or spurious streamlines.

reproducibility between BLSA and HCP. Only 11 bundles showed statistical differences between BLSA and HCP (see Table 6.2).

4 Discussion

Qualitative and quantitative identification of fiber pathways are crucial connectivity studies. Reproducible tractography can provide a backbone to the group-wise and longitudinal testing studies for which a comprehensive modern atlas is essential. Despite the importance, there are a limited number of reproducibility studies for manual bundle segmentation, particularly those with models other than DTI. Those that exist often investigate a limited number of fiber pathways due to the time-consuming nature of the tasks (Besseling et al., 2012; Ciccarelli et al., 2003; Heiervang et al., 2006; Rheault et al., 2020; Veenith et al., 2013; Wakana et al., 2007). Using a large number of WM pathways with state-of-the-art local model and tractography we have shown that most bundles achieve an acceptable level of reproducibility when it comes to the core (most dense regions) of the bundles, but that spurious streamlines and outliers remain problematic. The inclusion of older adults, from the BLSA database, and reproducibility scores being similar to younger adults, from the HCP database, demonstrate that the segmentation protocol is robust to WM structures heterogeneity related to aging. The main goal of this study was not only to design and quantify segmentation protocols that are clinically relevant, but to establish a baseline quality assessment. The current project identified strengths and weaknesses of our segmentation protocols and what can be improved in future iterations of the protocols.

4.1 Tasks complexities

By using an intra-rater and inter-rater project design with the TractEM protocol to obtain 35 (61 counting bilateral) individual bundles of interest, we quantified the ability of raters to follow simple/complex instructions. Our raters had no prior experiences with tractography to mimic how neurosciences students or clinicians without expertise in diffusion MRI could learn the TractEM protocols from instruction only. As shown in (Rheault et al., 2020), expertise in neuroanatomy or tractography had little impact on the quality/reproducibility of the segmentation executed by raters. The most reproducible bundles, when taking into account all metrics, were the subdivision of the corpus callosum (CC), pontine crossing tract (PCT), cingulum (CG), middle cerebral peduncle (MCP), corticospinal tract (CST), superior longitudinal fasciculus (SLF), posterior limb of the internal capsule (PIC), and uncinate fasciculus (UNC) on both datasets. In contrast, FX, fornix stria-terminalis (FXST), superior cerebellar peduncle (SCP), superior fronto-occipital fasciculus (SFO) exhibited low reproducibility. Some of the most reproducible bundles were categorized as simple bundles and some of the least reproducible were categorized as complex bundles. Even though HCP showed a more distinguishable division of repro-

ducibility between simple and complex bundles, the BLSA dataset shows the same trend. However, in most cases this categorization was not a good predictor of scores. In more extreme cases, it is possible that as the complexity of the manual protocol increases, the rater reproducibility decreases. However, the 'complexity' of a protocol is hard to define. Variables such as the number of ROIs, size of ROIs, shape of ROIs, number of planes (axial, coronal, sagittal) the ROIs are drawn on and even the presentation itself (figures, text disposition, ordering, etc.) could all increase or decrease complexity. It is possible that in our protocol, the number of ROIs was a driver of complexity. We hypothesize that the best way to increase the reproducibility score of the protocol itself would be to reduce the number of smaller exclusion ROIs required for complex bundles and to focus on using more clearly defined inclusion ROIs (point of passage, bottleneck, etc.).

4.2 Study design and protocol limitations

4.2.1 Intra-rater

Intra-rater reproducibility was evaluated only on the BLSA dataset due to the workload generated by the large number of pathways to segment. The effect of intra-rater variation will always reduce the observed reproducibility for any datasets (Gwet, 2014; Liao et al., 2010). The exact causes of the origin of this variation are still difficult to identify; we initially hypothesized that raters' interpretations do not change, only the execution. The exact level of reproducibility will vary due to multiple factors such as experience with anatomy or tractography, academic background of raters or familiarity with the software. However, it was observed that for virtually all bundles the intra-rater reproducibility scores were extremely close to their inter-rater counterparts. This indicates that, for this version of the protocol, whether one rater performs two segmentations or two raters perform one segmentation each (on the same dataset) the reproducibility scores will be similar.

4.2.2 Inter-rater

From the results, across metrics, across bundles for BLSA and HCP, it is difficult to reach simple conclusions. Overall, the Dice Coefficient for most bundles (in both database/acquisition) shows poor agreement. It was initially hypothesized that the HCP database would achieve higher scores due to its quality. We believed that quality of acquisition would lead to better bundle reconstruction and so easier, more intuitive, segmentation. However, only a minority of bundles reached our threshold for statistical significance between BLSA and HCP datasets; even though significant, many differences were quite small. The weighted Dice coefficient shows a more interesting pattern, the higher scores indicate that raters were able to segment the regions where density was the highest. While the average scores seem acceptable, the standard deviation is very large. Such a range of interpretation/execution across raters is in itself a problem. Bundle adjacency distance indicates that most

bundles (for both acquisitions/database) contained similar amounts of outliers and spurious streamlines. It is important to mention that bundles with high Dice and high bundle adjacency represent cases where extreme outliers were present. The reason why distant/isolated streamlines have such an impact on the bundle adjacency metric and not on the Dice coefficient is due to the fact that the distance to the closest neighbor (in millimeter) is part of the bundle adjacency computation. This means that a few streamlines extending far beyond the core of the bundle have a low impact on the total volume, but their cumulative distances to the core will rapidly increase. Scenarios with low Dice and high bundle adjacency can represent multiple (more complex) situations that cannot be easily (or intuitively) interpreted. The better reproducibility scores are, the easier they are to interpret and/or disentangle. While intra-rater reproducibility is absolutely crucial for a protocol, one can imagine a protocol that can be easily done again and again but is prone to misinterpretation. This would create a scenario where a, wrong, interpretation is reproducible for a single rater, but every rater submits entirely different looking bundles. The lower reproducibility across raters is expected since it merges the variability in execution and in interpretation. The similarity between intra-rater and inter-rater reproducibility score likely comes from the fact that our protocol is difficult to interpret and difficult to execute.

4.2.3 Intra-subject and Inter-subjects Since the reproducibility scores for most bundles of interest have either a low median reproducibility score, or a wide range of results (interquartile range); we would not recommend interpreting anatomical variability from these bundles if obtained from the TractEM protocol in its current version. Intra-subject (test-retest) reliability of DSI-Studio was evaluated for another protocol and showed high reproducibility (Yeh, 2020). The fact that DSI-Studio uses deterministic tractography, predetermined tracking parameters and that the segmentation protocol used was defined on a template rather than being subject-specific explains why bundle reproducibility is higher than other tools/algorithms (Cousineau et al., 2017; Zhang et al., 2019). Since our intra-rater reproducibility was evaluated on identical datasets (rather than test-retest), the intra-subject reliability can only be equal to or lower than our reported results. When a segmentation protocol has high intra-rater reproducibility scanner effects will dominate the intra-subject reproducibility. However, it was shown that when it comes to protocols with low reproducibility scanner effects will not significantly decrease reproducibility and the segmentation protocol will dominate intra-subject reproducibility (Schilling et al., 2021b). Using the results of the current work, it would be unwise to try to quantify variability at the individual or the population level (i.e., structural differences). Improvements to the TractEM protocol are needed before a more exhaustive quantification of the intra-subject reliability and inter-subjects comparison is performed. At the moment, the reproducibility scores range of most bundles is too large to accurately interpret the intra-subject reliability or the inter-subjects comparison. This highlights the importance of not only having a clearly defined and open protocol for bundle segmentation, but also a rigorous quantification of

its reproducibility.

4.3 Future work on the protocol

4.3.1 Generalization to other tools

TractEM allows us to obtain bundle segmentation from non-experts. At its initial developmental stage, it has shown encouraging results. Using the protocol, reconstructing a full dataset with 35 bundles with their bilateral counterparts is feasible in less than 6 hours per subject (including the software's learning curve and tractography itself). However, we must acknowledge the limitations of the protocol we used in this work. Some decisions were made at the time due to technological limitations or simple lack of familiarity with the cutting-edge processing tools available at the design stage. For example, the TractEM protocol is specific to deterministic tractography and embedded with DSI-Studio, which limits the generalization of anatomical definitions. Future works include modification to at least accept any deterministic tractograms from tools such as Dipy or MRtrix (Contributors et al., 2014; Tournier et al., 2019). Major modifications would be required to create a robust protocol for probabilistic tractography, but this work provides a framework as well as a baseline to quantify modification to the TractEM protocol. An important consideration, if the TractEM protocols become independent of the tractography reconstruction process, would be to evaluate the impact of the tractography method used. In its current state TractEM protocols only apply to deterministic tractography, but not all deterministic tractography algorithms are equal. For the current project, we used the same configuration (tracking parameters), but an approach of parameters exploration (Yeh, 2020) combining randomized parameters and a high streamline count to saturate bundle coverage could, hypothetically, lead to different results. Investigations of optimal tracking parameters and spurious streamlines pruning threshold (within DSI-Studio) could be investigated separately to determine which approach leads to higher reproducibility. However, it is likely impossible to generalize an existing segmentation protocol to new tracking algorithms or new bundles of interest (Rheault et al., 2020). Small deviations from the framework in which the protocol was evaluated can be acceptable, but major modifications such as generating streamlines using MRtrix (Tournier et al., 2019) that would impact both the local reconstruction as well as the tracking itself are difficult to predict. Furthermore, the choice of software in which the segmentation protocol is performed could lead to major differences in reproducibility scores (Rheault et al., 2020). Even a minor/major user-interface update or a change in the underlying delineation behavior in DSI-Studio could be enough to throw off previously quantified protocols. However, improvements to the TractEM protocol are needed before a more exhaustive quantification of the impact of more minor variables.

4.3.2 Fantastic tracts and how to define them

This first iteration of the protocol was inspired by the EVE atlas, initially the intent was to define pathways using simple regions described in the original work. However, the original work was a 3D atlas of the WM when in fact WM pathways often (always) overlap each other. This means that the way bundles are defined in TractEM is not optimal considering the structure of what it attempts to segment/reconstruct. In order to respect the bundles as defined in the EVE atlas, the TractEM protocol used minimal seeds/ROIs regions with numerous ROAs. Furthermore, the number of ROAs and their size was left to the rater with little to no description, this not only made the segmentation tasks difficult to follow, but likely drove the variability. From Figure 6.4, we can observe that some pathways are nearly indistinguishable from one another. This is due to multiple factors: 1) Major pathways are easier to reconstruct than others, bundles with close trajectory will often simply merge into a single entity. 2) The raters likely did not know what was the expected shape or the instructions were not clear enough to disentangle close bundles. 3) The EVE atlas does not define pathways globally; it rather simply defines local regions. This leads to problems when global pathways share local regions and this disconnection between the EVE WM 3D atlas and tractography highlight the need to challenge classical WM delineation. Another area of improvement would be to increase the details on anatomical landmarks surrounding ROIs and how to identify the right slice, to identify ROIs' size and border. Using better nomenclature for the identification of landmarks would also provide more context for participants familiar with neuroanatomy while being a learning tool for those not familiar with it. In DSI-Studio, the data is co-registered to an atlas and in a similar fashion to the EVE atlas the approximate position of ROIs/ROAs can be easily suggested to the raters. However, this iteration of the protocol provided no anatomical landmark, no general description of the ROIs/ROAs and limited figures. This often resulted in the rater picking the exact suggested slices and blindly drawing a region in the general neighborhood of what was expected. Again, this not only makes the protocol difficult to follow, but increases the variability. This also makes the current protocol entirely dependent on DSI-Studio and its internal co-registration routine. In the current version all bundles are considered independent, no ROIs/ROAs are reusable. Furthermore, for each bundle the ordering of regions to draw is often arbitrary, switching from axial to coronal to sagittal or going back and forth from inclusion to exclusion regions. This drastically increases software interaction (and so the chance for error/mistake), leading to some suboptimality. It would be more efficient to design a protocol where some regions can be reused. For example, all association bundles required a mid-sagittal plane to cut off commissural pathways. This region of exclusion was redrawn each time, which is time-consuming and generates variability (e.g., choosing a different slice each time). This would require the protocol to cross reference ROIs, but the advantages for the raters and reproducibility are non-negligible. Another important factor is the consistency of the text. Instructions should be similarly phrased, similarly

ordered and have a similar level of detail. This would facilitate the process of developing a routine for raters, which tends to accelerate the tasks, decrease the occurrence of mistakes and increase reproducibility. All seeds/ROIs/ROAs should be accounted for and described, leaving the total number of regions unclear or to the raters' interpretation is a source of confusion. Optional regions should be clearly identified. Overall, we believe all regions must be justified, their goals/reasons be explicitly described and their impact (if drawn right) on the global pathway trajectory detailed. Finally, a future iteration would require more visual description. Each region should be defined using at least one image. When surrounding landmarks need to be found, they should be pointed to, zoomed in and shown on more than one contrast, at the moment 100% of the protocol relies on the RGB (ColorFA). It would also be useful to provide a multi-orientation view of the bundles in 3D or a population average probability map to inform the raters of what is expected and help to detect errors/mistakes more easily. All figures should clearly indicate, plane (sagittal, coronal, axial), orientation (R/L, A/P, S/I) and zoomed figures should show the entire slice in a vignette.

4.3.3 Theoretical anatomical definitions

Finally, a major update would be required to separate the conceptual anatomical definition (theory) vs its execution (practice). For example, if the definition of the corticospinal tract requires it to terminate in the brainstem and in the precentral and postcentral gyrus as well as passing through the internal capsule (Chenot et al., 2019), that would be an agreed upon theoretical anatomical definition. However, the execution can vary to respect these criteria. In a specific protocol these rules could be achieved using planar ROIs limiting the posterior and anterior region at the cortex (Rheault et al., 2020) or in the lower brainstem, but in another version, it could require the usage of an atlas, e.g., Freesurfer (Desikan et al., 2006) to obtain well-delineated 3D ROIs. These variations in execution would likely give different reproducibility scores for each bundle while nothing changed in the theoretical anatomical definition. Defining as a community a set of agreed upon theoretical anatomical definitions is crucial, then defining an agreed upon execution of this definition is also crucial and the variability must be quantified each time it is modified (or each deviation from the original one). As of right now, TractEM is only an execution protocol that was not created from a consensus in the field, but it still provided a first iteration, a baseline, to identify limitations, pave the way forward for the next iteration and initiate discussion on how to define WM pathways in humans.

4.3.4 Usefulness for automatic algorithms

Different automated protocols are capable of producing streamlines, highly reproducible (small and large) bundles covering the whole brain (Garyfallidis et al., 2018; Rheault et al., 2018; Wasserthal et al., 2018; Zhang

et al., 2020a).These methods often rely on manual delineations as their a priori. This highlights the importance of reproducibility studies at the manual level. Subject-specific delineations can provide a backbone to automated methods. As mentioned earlier, there is a distinction between an anatomical definition in theory and in practice. Automatic methods can be seen as yet one more version to execute a theoretical anatomical definition. Consensus must be obtained; automatic methods would greatly benefit from agreed upon definition from the field of neuroanatomy method.

5 Conclusion

In this work we investigated the reproducibility of a first iteration of the TractEM protocol, an open and public bundle segmentation protocol based on the EVE WM atlas. We have shown that most bundles achieve good reproducibility in the densest regions (core) of the bundles. Spurious streamlines and outliers remain challenging to avoid and affect negatively the reproducibility scores for most bundles. The complexity of some segmentation or the lack of clarity in the instructions sometimes leads to bundles of questionable anatomical validity. Identifying flaws and limitations of the protocol will lead to improvements. The experience obtained from creating a whole brain bundle segmentation protocol as well as evaluating its performance provided useful insight for future iterations.

6 Acknowledgements

This work was supported by the National Institutes of Health under award numbers R01EB017230, T32EB001628, and in part by ViSE/VICTR VR3029 and the National Center for Research Resources, Grant UL1 RR024975-01. This research was conducted with the support from Intramural Research Program, National Institute on Aging, NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

7 Data Availability Statement

TractEM is an open-source project, in terms of data and labeling guidelines. Comments and discussion are open for each bundle. The frameworks enable versioning of bundle definitions as consensus definitions evolve. By crowd sourcing our results, we aim to obtain expert validation for each bundle. The raw input data, Talairach-aligned data, DSI-Studio-ready input data, and TractEM results are made available on our website, https://my.vanderbilt.edu/tractem.

Complete Name	Acronym
Anterior Commissure	AC
Anterior Corona Radiata	ACR
Anterior Limb of the Internal Capsule	AIC
Body of Corpus Callosum	BCC
Cerebral Peduncle	CP
Cingulum (Cingulate Gyrus part)	CGC
Cingulum (Hippocampal part)	CGH
Corticospinal Tract	CST
Fornix	FX
Fornix / Stria Terminalis	FXST
Genu of the Corpus Callosum	GCC
Inferior Cerebellar Peduncle	ICP
Inferior Fronto Occipital Fasciculus	IFO
Inferior Longitudinal Fasciculus	ILF
Medial Lemniscus	ML
Midbrain	M
Middle Cerebellar Peduncle	MCP
Olfactory Radiation	OLFR
Optic Tract	OPT
Posterior Corona Radiata	PCR
Pontine Crossing Tract	PCT
Posterior Limb of the Internal Capsule	PIC
Posterior Thalamic Radiation	PTR
Sagittal Stratum	SS
Splenium of the Corpus Callosum	SCC
Superior Cerebellar Peduncle	SCP
Superior Corona Radiata	SCR
Superior Fronto Occipital Fasciculus	SFO
Superior Longitudinal Fasciculus	SLF
Tapetum of the Corpus Callosum	TAP
Uncinate Fasciculus	UNC
Frontal Lobe	FL
Parietal Lobe	PL
Occipital Lobe	OL
Temporal Lobe	TL

Table 6.1: Complete list of acronyms for the TractEM pathways.

	Dice		Weighted Dice		Bundle Adjacency	
	BLSA	**HCP**	**BLSA**	**HCP**	**BLSA**	**HCP**
OL_LR	0.93 (0.01)	0.94 (0.00)	1.00 (0.00)	1.00 (0.00)	0.09 (0.02)	0.07 (0.01)
PT_LR	0.92 (0.02)	0.92 (0.01)	0.99 (0.01)	0.99 (0.00)	0.12 (0.04)	0.10 (0.03)
TL_LR	0.91 (0.01)	0.90 (0.01)	0.99 (0.00)	0.99 (0.00)	0.13 (0.05)	0.13 (0.05)
FL_LR	0.90 (0.02)	0.91 (0.01)	0.99 (0.00)	0.99 (0.00)	0.12 (0.04)	0.10 (0.01)
GCC	0.78 (0.10)	0.86 (0.05)	0.96 (0.03)	0.98 (0.02)	0.76 (0.73)	0.25 (0.17)
SCC	0.71 (0.18)	0.82 (0.16)	0.89 (0.19)	0.95 (0.09)	1.51 (1.85)	0.64 (1.24)
M	0.63 (0.13)	**0.61 (0.15)**	0.87 (0.13)	**0.89 (0.09)**	1.42 (1.06)	**4.55 (4.20)**
BCC	0.71 (0.12)	0.83 (0.09)	0.92 (0.05)	0.96 (0.07)	1.29 (1.00)	0.42 (0.42)
MCP	0.77 (0.15)	0.88 (0.04)	0.96 (0.06)	0.99 (0.01)	1.85 (3.54)	0.15 (0.08)
SCR_LR	0.65 (0.13)	0.84 (0.09)	0.85 (0.12)	0.97 (0.03)	1.58 (1.12)	0.51 (0.64)
PTR_LR	0.60 (0.18)	0.77 (0.10)	0.81 (0.21)	0.95 (0.06)	1.70 (1.37)	0.51 (0.31)
UNC_LR	0.65 (0.16)	0.84 (0.11)	0.87 (0.10)	0.96 (0.08)	1.32 (1.20)	0.42 (0.64)
PIC_LR	0.68 (0.05)	**0.82 (0.10)**	0.89 (0.05)	**0.97 (0.04)**	1.27 (0.49)	**0.60 (0.71)**
CST_LR	0.63 (0.14)	**0.76 (0.11)**	0.90 (0.06)	**0.97 (0.03)**	2.30 (1.62)	**1.28 (1.03)**
CGC_LR	0.72 (0.12)	**0.73 (0.08)**	0.96 (0.04)	**0.92 (0.06)**	1.73 (1.75)	**0.85 (0.54)**
ACR_LR	0.70 (0.08)	0.62 (0.17)	0.94 (0.04)	0.83 (0.14)	0.68 (0.31)	3.87 (3.45)
SS_LR	0.64 (0.17)	0.72 (0.17)	0.88 (0.13)	0.92 (0.09)	1.38 (1.36)	1.48 (1.41)
AIC_LR	0.51 (0.14)	**0.79 (0.14)**	0.81 (0.15)	**0.95 (0.08)**	4.74 (3.70)	**0.66 (0.86)**
PCR_LR	0.64 (0.12)	0.79 (0.12)	0.88 (0.09)	0.94 (0.07)	1.91 (1.26)	0.65 (0.58)
SLF_LR	0.72 (0.07)	**0.71 (0.14)**	0.93 (0.04)	**0.93 (0.07)**	0.96 (0.50)	**0.97 (0.72)**
ILF_LR	0.64 (0.09)	0.69 (0.15)	0.89 (0.07)	0.93 (0.08)	1.71 (0.87)	1.26 (0.94)
PCT	0.76 (0.07)	**0.87 (0.04)**	0.98 (0.02)	**1.00 (0.01)**	0.45 (0.21)	**0.18 (0.08)**
IFO_LR	0.60 (0.20)	0.59 (0.20)	0.82 (0.25)	0.90 (0.11)	1.58 (1.34)	2.87 (3.05)
ICP_LR	0.48 (0.18)	0.51 (0.23)	0.86 (0.13)	0.85 (0.19)	4.27 (2.65)	3.15 (2.43)
CP_LR	0.45 (0.21)	0.66 (0.21)	0.83 (0.12)	0.92 (0.13)	4.10 (2.76)	1.88 (2.07)
AC	0.57 (0.19)	0.52 (0.29)	0.88 (0.18)	0.89 (0.13)	2.24 (2.70)	4.78 (5.54)
OLFR_LR	0.59 (0.22)	0.62 (0.30)	0.78 (0.26)	0.86 (0.20)	2.95 (4.32)	5.21 (6.67)
SFO_LR	0.46 (0.20)	**0.66 (0.19)**	0.82 (0.20)	**0.89 (0.16)**	3.54 (2.66)	**2.16 (1.92)**
TAP	0.53 (0.13)	0.62 (0.13)	0.85 (0.14)	0.88 (0.09)	2.70 (1.92)	2.19 (1.40)
ML_LR	0.59 (0.21)	**0.64 (0.17)**	0.82 (0.16)	0.89 (0.11)	3.09 (2.80)	3.97 (3.92)
SCP_LR	0.31 (0.20)	0.69 (0.17)	0.78 (0.12)	0.97 (0.04)	4.82 (3.11)	1.54 (1.84)
CGH_LR	0.60 (0.21)	**0.48 (0.23)**	0.92 (0.12)	**0.77 (0.24)**	2.87 (2.77)	**3.29 (2.77)**
FX_LR	0.45 (0.22)	**0.56 (0.24)**	0.80 (0.17)	0.85 (0.17)	3.04 (2.43)	**3.18 (4.02)**
OPT	0.44 (0.15)	**0.72 (0.17)**	0.78 (0.17)	**0.93 (0.12)**	3.56 (1.53)	**1.15 (1.16)**
FXST_LR	0.29 (0.25)	0.18 (0.20)	0.46 (0.30)	0.27 (0.29)	3.90 (2.10)	6.42 (3.26)

Table 6.2: Reported average (standard deviation) values for all bundles for both datasets. Cells in bold show statistical significance ($p < 0.01$, one-sided) for our hypothesis that HCP would yield higher reproducibility scores than BLSA. When statistical significance is reached, all metrics reach the threshold (2 out of 3 for FXLR). The bundle names in italics have Dice coefficient below 0.5, this curtails any potential analysis due to the variability in shape being too high. The bundles with the suffix 'LR' represent bilateral pathways (Left/Right).

CHAPTER 7

Conclusions

1 Summary

This dissertation is focused on the developing models with the aim of advancing the individual precision in neuroimaging analysis.

Inter-individual differences are a fascinating aspect of human beings. It encompasses the diverse range of characteristics, traits, and behaviors that make each person unique. From physical attributes to cognitive abilities, from personality traits to emotional responses, the variations among individuals contribute to the rich tapestry of human existence. However, in neuroimaging research inter-individual variability poses a significant challenge that needs to be addressed in order to unravel the factors that make each person unique and to elucidate commonalities and differences in brain function in health and disease. In this dissertation, I have developed computational methods to tackle some of the current challenges associated with individualized mapping of the human brain.

To begin with, we investigated how each person's peripheral physiology imprints a unique signature on the brain. We devised learning frameworks that enable the inference of heart rate and respiration variation directly from fMRI data (Chapter II), and subsequently explored the individual-specific spatial patterns imposed by these physiological processes on these brain activity proxies (Chapter III). While tackling this at the voxel-level using traditional machine learning approaches, we quickly realized that most advanced machine learning models cannot handle the size of fMRI data and require reliable dimensionality reduction techniques. We proposed parcellation methods to devise individual-specific functional parcellations. One approach involved the creation of a pragmatic and interactive tool that facilitates the projection of group parcellations onto subject-specific parcellations (Chapter IV). Additionally, we proposed a technique that leverages commonly acquired structural magnetic resonance imaging (sMRI) data, such as T1 and T2 scans, to delineate functional regions at the individual level (Chapter V). Lastly, towards the advancement of reproducible tractography protocols, we created manual delineation guidelines for identification of white matter pathways (Chapter VI). Overall, these studies presented novel computational methods to address the challenges of individualized mapping of the human brain and provided new insights into the inter-individual variability of brain function. The subsequent sections present a concise overview of the primary findings and contributions made in this research, followed by suggestions for potential future research directions.

2 Summary of Contributions and Future Work

2.1 Peripheral Physiology and BOLD fMRI

2.1.1 Summary

The ability to image the brain in real-time, non-invasively has opened up exciting opportunities for research. However, it is important to acknowledge that in techniques such as blood oxygenation level-dependent functional magnetic resonance imaging (BOLD fMRI), we are measuring a proxy for neural activity rather than directly observing the neural signals themselves. This means that our measures contain both confounds and sources of information that need to be carefully considered. One significant factor that influences BOLD signals is peripheral physiological signals. These signals, which include measures of heart rate and respiration could be recorded during fMRI scans and have been observed to co-vary with what we expect as neural components. Historically, these signals were regarded as mere confounds and were removed from the BOLD data. However, it is now evident that the relationship between peripheral physiological signals and neural activity is more complex than initially thought (Chen et al., 2020; Bright et al., 2020; Yoo et al., 2022; Nashiro et al., 2022).

In this dissertation, we have shown that measures of peripheral physiology are not only useful in studying the constituents of BOLD fMRI signals but also exhibit unique BOLD physiological patterns that possess heritable characteristics. The heritability of physiological signals underscores their genetic basis, suggesting shared genetic influences on both brain function and physiological mechanisms. In line with the existing literature [cite], we defend using physiological measure to improve the validity and reproducibility of fMRi analysis.

However, the field is fast growing yet still niche, only a small group of research labs and initiatives recognize the importance of these peripheral windows into the brain. There are many existing datasets lack these measures (ADNI (Jack Jr et al., 2015); UK Biobank (Bycroft et al., 2018); HCP 7T Release (WU-Minn, 2017)). Collecting these recordings also present challenges, even when it is acquired, it is not always of sufficient quality. One of our major contribution is modeling two key physiological signals, heart rate and respiration variation from fMRI data alone. Using advanced recurrent neural network (RNN) models, we were able to model both of these signals simultaneously. Resulting models are shown to be generalizable without further training within fMRI paradigm. Thereby, enabling researchers to include this rich source of complementary information to their studies.

2.1.2 Technical Innovations

I We demonstrated that fluctuations in respiration volume can be extracted from fMRI alone.

II We developed generalizable frameworks to jointly reconstruct the heart rate and respiration variation. By doing so, we enabled enrichment of datasets that lack physiological measurements collected concurrently with fMRI.

III We characterized the BOLD physiological patterns and showed the first evidence that these patterns may be heritable.

2.1.3 Future Directions

Our current approach for inferring physiological signals is a supervised learning technique, which learns from labeled data known as "ground truth". However, the effectiveness of the models heavily relies on the quality of these labels. Developing models that can distinguish between good and bad physiological waveforms has the potential to enhance the utility of supervised models. Another direction potentially could expand the scope of the inference models and include small and extreme variability samples in breathing and heart rate. This could help capture the full spectrum of physiological variations leading to a more accurate and nuanced representation of physiological mechanisms. Additionally, exploring models for aging populations and patient cohorts is certainly of interest in the field. By extending the study of physiological signals to aging and patient cohorts, we have the opportunity to identify novel biomarkers that are specific to these populations, enabling early detection, monitoring, and targeted interventions. Finally, characterization of the physiological signal models, studying when they work, when they fail is certainly of interest in future iterations. Aging and patients cohorts that often exhibit distinct brain patterns compared to younger and healthy individuals. Further characterization of BOLD physiological patterns in these cohorts can provide valuable insights into age-related changes in physiology and progression of mental disorders. One future data collection effort could involve a longitudinal study assessing health status, disease progression, and treatment response.

2.2 Functional Parcellations

2.2.1 Summary

The shift in the neuroimaging community from group-level studies to individual-level inference has sparked interest in developing individual-specific functional parcellations. To facilitate this process, we proposed PRAGMA, an interactive visualization tool that maps group-level functional regions to individual-specific ones. PRAGMA offers multiple data views to enable users in making informed decisions when customizing the parcels for each individual. This framework has garnered positive feedback from both the visualization and neuroimaging communities.

However, beyond the realm of fMRI, where structural MRI scans are more readily available, such methods

for deriving functional boundaries are lacking. This data gap presents a significant challenge in various applications. To fill this gap, we proposed an approach to learn functional boundaries of cortical surface regions from structural MRI measures. Our proposed method utilizes a spherical CNN, which offers distinct advantages over voxel-level cortical segmentation approaches. In preoperative planning and brain stimulation therapies, the ability to precisely localize functional boundaries in an individual will enable individualized treatment and can potentially improve patient outcome.

2.2.2 Technical Innovation

1. We developed the first model to predict functional parcellations from structural data.

2. We developed the first interactive analytics parcellation approach to tailor group level functional parcels to individual-specific parcels.

2.2.3 Future Directions

PRAGMA in its beta stage can only be modified through manual intervention. This renders the tool tedious and time consuming. To address this limitation, a potential direction is to incorporate a semi-automated individualization workflow into the framework. This addition would enable an initial automated pass to tailor the parcels to individuals, reducing the manual effort required and streamlining the process. Regarding the limitations of the cortical labeling study, it was observed that finer parcellation schemes had missing labels, resulting in a variable number of parcels across subjects in the training set. This variability made it challenging to effectively learn these parcels when working with different numbers of parcellations. To overcome this limitation, a future direction could involve developing an agnostic model capable of accommodating varying numbers of parcels for each subject. Such a model would allow for more flexible and adaptable parcellations, taking into account the diversity of the underlying biological substrate and potentially leading to more plausible results.

2.3 Manual Tractography

2.3.1 Summary

Tractography in neuroimaging presents a complex and challenging task. While automated methods may seem like an easy solution, they often rely on manual delineations as their a priori. However, existing manual protocols are tedious, time consuming and limited to a handful of fiber pathways. Manual delineations can provide a backbone to automated methods. In this work, Tractography with Empirical MRI (TractEM) project, we created a first iteration execution protocol. While evaluating the TractEM protocol, we discovered manual pipelines introduce notable inter-rater variability. Even with the application of well-defined protocols to min-

imize subjectivity, slight variations in interpretation can still lead to inconsistent results. The consequences of inter-rater variability would extend to subsequent models, research studies, and algorithm development, potentially impeding progress in the field.

To tackle this challenge, it is essential to undertake standardization efforts, establish clear guidelines, and implement quality control measures. These measures are crucial in reducing variability and enhancing the reliability of manual delineation of fiber pathways. At its current stage, TractEM serves as an execution protocol that was not developed based on a consensus within the field. However, it provides a starting point, allowing us to identify limitations, lay the foundation for subsequent iterations, and initiate discussions on how to precisely define white matter (WM) pathways in humans.

2.3.2 Technical Innovation and Impact on Scientific Research

I We introduced an execution protocol aimed at addressing the limitations of existing manual protocols in neuroimaging.

II We identified several sources of variability in manual delineation protocols.

III We prescribed community consensus when bringing theoretical anatomical definitions and execution of definitions.

2.3.3 Future Directions

Manual delineations of fiber pathways not only useful at the individual subject level but also provide a backbone to automated methods. It is important to curate reproducible manual tractography protocols that can guide wide range of tractography algorithms and tracking parameters. In future iterations of this work, building a community consensus will be the most important first step. That is, as a community defining a set of agreed upon theoretical anatomical definitions and an agreed upon execution of these definitions. In each iteration, for a new set of definitions, rigorous quantification of its reproducibility is advised. The variability should be quantified to assess how useful a delineation protocol is in practice, strengthening the foundation for subsequent iterations.

3 Data and Code Availability

The data and code used in this PhD thesis are available to promote transparency, reproducibility, and further research. By sharing our code, we hope to foster a culture of open science, accelerate the progress of research, and enable others to build upon our work. We are committed to supporting the scientific community in their efforts to reproduce, validate, and extend our research findings. We encourage researchers and developers to

engage with our code, contribute improvements, and report any issues or bugs they encounter. Feedback from the community is highly valuable in refining and expanding the codebase.

The GitHub repository for this thesis can be found at https://github.com/rgbayrak. The analysis, modeling, and visualization code used in this thesis, along with any related code, are publicly available in the repository. The datasets used in this thesis, along with any additional datasets employed, are publicly available. For datasets that were not public, they have been made available through a specific data access process. All data and code in the repository are released under the MIT License, granting users the freedom to use, modify, and distribute them. Please refer to the accompanying documentation and README file for detailed instructions on accessing data and utilizing the code.

References

Adluru, N., Destiche, D. J., Tromp, D. P., Davidson, R. J., Zhang, H., and Alexander, A. L. (2016). Evaluating consistency of deterministic streamline tractography in non-linearly warped dti data. *arXiv preprint arXiv:1602.02117*.

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-É., Bludau, S., Bazin, P.-L., Lewis, L. B., Oros-Peusquens, A.-M., et al. (2013). Bigbrain: an ultrahigh-resolution 3d human brain model. *Science*, 340(6139):1472–1475.

Anderson, K. M., Ge, T., Kong, R., Patrick, L. M., Spreng, R. N., Sabuncu, M. R., Yeo, B. T., and Holmes, A. J. (2021). Heritability of individualized cortical network topography. *Proceedings of the National Academy of Sciences*, 118(9):e2016271118.

Anderson, V., Spencer-Smith, M., and Wood, A. (2011). Do children really recover better? neurobehavioural plasticity after early brain insult. *Brain*, 134(8):2197–2221.

Andersson, J., Skare, S., and Ashburner, J. (2003). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *NeuroImage*, 20(2):870–888.

Andersson, J. L. and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging. *Neuroimage*, 125:1063–1078.

Arslan, S., Ktena, S. I., Makropoulos, A., Robinson, E. C., Rueckert, D., and Parisot, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *Neuroimage*, 170:5–30.

Ash, T., Suckling, J., Walter, M., Ooi, C., Tempelmann, C., Carpenter, A., and Williams, G. (2013). Detection of physiological noise in resting state fmri using machine learning. *Human brain mapping*, 34(4):985–998.

Aslan, S., Hocke, L., Schwarz, N., and Frederick, B. (2019). Extraction of the cardiac waveform from simultaneous multislice fmri data using slice sorted averaging and a deep learning reconstruction filter. *Neuroimage*, 198:303–316.

Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., and Gee, J. C. (2010). The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*, 49(3):2457–2466.

Azzalini, D., Rebollo, I., and Tallon-Baudry, C. (2019). Visceral signals shape brain dynamics and cognition. *Trends in cognitive sciences*, 23(6):488–509.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., and Hyde, J. S. (1992). Time course epi of human brain function during task activation. *Magnetic resonance in medicine*, 25(2):390–397.

Barber, A. D., Hegarty, C. E., Lindquist, M., and Karlsgodt, K. H. (2021). Heritability of functional connectivity in resting state: assessment of the dynamic mean, dynamic variance, and static connectivity across networks. *Cerebral Cortex*, 31(6):2834–2844.

Barrett, L. F. and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature reviews neuroscience*, 16(7):419–429.

Basser, P. J. and Pierpaoli, C. (2011). Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri. *Journal of magnetic resonance*, 213(2):560–570.

Bayrak, R. G., Hansen, C. B., Salas, J. A., Ahmed, N., Lyu, I., Huo, Y., and Chang, C. (2021). From brain to body: Learning low-frequency respiration and cardiac signals from fmri dynamics. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer.

Bayrak, R. G., Salas, J. A., Huo, Y., and Chang, C. (2020). A deep pattern recognition approach for inferring respiratory volume fluctuations from fmri data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–436. Springer.

Beall, E. B. and Lowe, M. J. (2007). Isolating physiologic noise sources with independently determined spatial measures. *Neuroimage*, 37(4):1286–1300.

Beissner, F., Meissner, K., Bär, K.-J., and Napadow, V. (2013). The autonomic brain: an activation likelihood estimation meta-analysis for central processing of autonomic function. *Journal of neuroscience*, 33(25):10503–10511.

Bernier, M., Cunnane, S. C., and Whittingstall, K. (2018). The morphology of the human cerebrovascular system. *Human brain mapping*, 39(12):4962–4975.

Besseling, R. M., Jansen, J. F., Overvliet, G. M., Vaessen, M. J., Braakman, H. M., Hofman, P. A., Aldenkamp, A. P., and Backes, W. H. (2012). Tract specific reproducibility of tractography based morphology and diffusion metrics. *PloS one*, 7(4):e34125.

Bhagwat, N., Barry, A., Dickie, E. W., Brown, S. T., Devenyi, G. A., Hatano, K., DuPre, E., Dagher, A., Chakravarty, M., Greenwood, C. M., et al. (2021). Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *GigaScience*, 10(1):giaa155.

Bijsterbosch, J. D., Woolrich, M. W., Glasser, M. F., Robinson, E. C., Beckmann, C. F., Van Essen, D. C., Harrison, S. J., and Smith, S. M. (2018). The relationship between spatial configuration and functional connectivity of brain regions. *Elife*, 7:e32992.

Birn, R. M., Cornejo, M. D., Molloy, E. K., Patriat, R., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., and Prabhakaran, V. (2014). The influence of physiological noise correction on test–retest reliability of resting-state functional connectivity. *Brain connectivity*, 4(7):511–522.

Birn, R. M., Diamond, J. B., Smith, M. A., and Bandettini, P. A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fmri. *Neuroimage*, 31(4):1536–1548.

Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., and Prabhakaran, V. (2013). The effect of scan length on the reliability of resting-state fmri connectivity estimates. *Neuroimage*, 83:550–558.

Birn, R. M., Smith, M. A., Jones, T. B., and Bandettini, P. A. (2008). The respiration response function: the temporal dynamics of fmri signal fluctuations related to changes in respiration. *Neuroimage*, 40(2):644–654.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J. G., De Leon, M. J., deToledo Morrell, L., et al. (2011). Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint eadc-adni harmonized protocol. *Journal of Alzheimer's disease*, 26(s3):61–75.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88.

Boukhdhir, A., Zhang, Y., Mignotte, M., and Bellec, P. (2021). Unraveling reproducible dynamic states of individual brain functional parcellation. *Network Neuroscience*, 5(1):28–55.

Bright, M. G., Whittaker, J. R., Driver, I. D., and Murphy, K. (2020). Vascular physiology drives functional brain networks. *NeuroImage*, page 116907.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth.

Bruneau, P., Pinheiro, P., Broeksema, B., and Otjacques, B. (2015). Cluster sculptor, an interactive visual clustering system. *Neurocomputing*, 150:627–644.

Burt, J. B., Demirtaş, M., Eckner, W. J., Navejar, N. M., Ji, J. L., Martin, W. J., Bernacchia, A., Anticevic, A., and Murray, J. D. (2018). Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nature neuroscience*, 21(9):1251–1259.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Catani, M. and De Schotten, M. T. (2008). A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *cortex*, 44(8):1105–1132.

Catani, M., Howard, R. J., Pajevic, S., and Jones, D. K. (2002). Virtual in vivo interactive dissection of white matter fasciculi in the human brain. *Neuroimage*, 17(1):77–94.

Cavallo, M. and Demiralp, Ç. (2018). Clustrophile 2: Guided visual clustering analysis. *IEEE transactions on visualization and computer graphics*, 25(1):267–276.

Chai, X. J., Castañón, A. N., Öngür, D., and Whitfield-Gabrieli, S. (2012). Anticorrelations in resting state networks without global signal regression. *Neuroimage*, 59(2):1420–1428.

Chang, C., Cunningham, J. P., and Glover, G. H. (2009). Influence of heart rate on the bold signal: the cardiac response function. *Neuroimage*, 44(3):857–869.

Chang, C. and Glover, G. H. (2009). Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. *Neuroimage*, 47(4):1448–1459.

Chen, J. E., Lewis, L. D., Chang, C., Tian, Q., Fultz, N. E., Ohringer, N. A., Rosen, B. R., and Polimeni, J. R. (2020). Resting-state "physiological networks". *NeuroImage*, 213:116707.

Chenot, Q., Tzourio-Mazoyer, N., Rheault, F., Descoteaux, M., Crivello, F., Zago, L., Mellet, E., Jobard, G., Joliot, M., Mazoyer, B., et al. (2019). A population-based atlas of the human pyramidal tract in 410 healthy participants. *Brain Structure and Function*, 224(2):599–612.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Chong, M., Bhushan, C., Joshi, A. A., Choi, S., Haldar, J. P., Shattuck, D. W., Spreng, R. N., and Leahy, R. M. (2017). Individual parcellation of resting fmri with a group functional connectivity prior. *NeuroImage*, 156:87–100.

Ciccarelli, O., Parker, G. J., Toosy, A. T., Wheeler-Kingshott, C., Barker, G. J., Boulby, P. A., Miller, D. H., and Thompson, A. J. (2003). From diffusion tractography to quantitative white matter tract measures: a reproducibility study. *Neuroimage*, 18(2):348–359.

Cinciute, S. (2019). Translating the hemodynamic response: why focused interdisciplinary integration should matter for the future of functional neuroimaging. *PeerJ*, 7:e6621.

Contributors, D., Garyfallidis, E., Brett, M., Amirbekian, B. B., Rokem, A., van der Walt, S., Descoteaux, M., and Nimmo-Smith, I. (2014). Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8.

Cousineau, M., Jodoin, P.-M., Garyfallidis, E., Côté, M.-A., Morency, F. C., Rozanski, V., Grand'Maison, M., Bedell, B. J., and Descoteaux, M. (2017). A test-retest study on parkinson's ppmi dataset yields statistically significant white matter fascicles. *NeuroImage: Clinical*, 16:222–233.

Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K. J., Wassermann, D., Thirion, B., and Mensch, A. (2020). Fine-grain atlases of functional modes for fmri analysis. *arXiv preprint arXiv:2003.05405*.

Das, S., Saket, B., Kwon, B. C., and Endert, A. (2020). Geono-cluster: Interactive visual cluster analysis for biologists. *IEEE Transactions on Visualization and Computer Graphics*.

Dayan, E. and Cohen, L. G. (2011). Neuroplasticity subserving motor skill learning. *Neuron*, 72(3):443–454.

De Schotten, M. T., Bizzi, A., Dell'Acqua, F., Allin, M., Walshe, M., Murray, R., Williams, S. C., Murphy, D. G., Catani, M., et al. (2011). Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with mr diffusion tractography. *Neuroimage*, 54(1):49–59.

de Schotten, M. T., Foulon, C., Nachev, P., et al. (2020). Brain disconnections link structural connectivity with function and behaviour. *Nature Communications*, 11(1):1–8.

Dell'Acqua, F. and Tournier, J.-D. (2019). Modelling white matter with spherical deconvolution: How and why? *NMR in Biomedicine*, 32(4):e3945.

Demirtaş, M., Burt, J. B., Helmer, M., Ji, J. L., Adkinson, B. D., Glasser, M. F., Van Essen, D. C., Sotiropoulos, S. N., Anticevic, A., and Murray, J. D. (2019). Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron*, 101(6):1181–1194.

Descoteaux, M., Angelino, E., Fitzgibbons, S., and Deriche, R. (2007). Regularized, fast, and robust analytical q-ball imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(3):497–510.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Dubois, J. and Adolphs, R. (2016). Building a science of individual differences from fmri. *Trends in cognitive sciences*, 20(6):425–443.

Edlow, B. L., Takahashi, E., Wu, O., Benner, T., Dai, G., Bu, L., Grant, P. E., Greer, D. M., Greenberg, S. M., Kinney, H. C., et al. (2012). Neuroanatomic connectivity of the human ascending arousal system critical to consciousness and its disorders. *Journal of Neuropathology & Experimental Neurology*, 71(6):531–546.

Eickhoff, S. B., Yeo, B. T., and Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686.

Essayed, W. I., Zhang, F., Unadkat, P., Cosgrove, G. R., Golby, A. J., and O'Donnell, L. J. (2017). White matter tractography for neurosurgical planning: A topography-based review of the current state of the art. *NeuroImage: Clinical*, 15:659–672.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019a). fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116.

Esteban, O., Wright, J., Markiewicz, C. J., Thompson, W. H., Goncalves, M., Ciric, R., Blair, R. W., Feingold, F., Rokem, A., Ghosh, S., et al. (2019b). Nipreps: enabling the division of labor in neuroimaging beyond fmriprep.

Ferreira, L. K. and Busatto, G. F. (2013). Resting-state functional connectivity in normal brain aging. *Neuroscience & Biobehavioral Reviews*, 37(3):384–400.

Ferrucci, L., Giallauria, F., and Guralnik, J. M. (2008). Epidemiology of aging. *Radiologic Clinics of North America*, 46(4):643–652.

Finn, E. S. (2021). Is it time to put rest to rest? *Trends in cognitive sciences*, 25(12):1021–1032.

Finn, E. S. and Constable, R. T. (2022). Individual variation in functional brain connectivity: implications for personalized approaches to psychiatric disease. *Dialogues in clinical neuroscience*.

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

Fischl, B. (2012). Freesurfer. *Neuroimage*, 62(2):774–781.

for Research into Nervous, A. A., Diseases, M., Llinás, R., Ribary, U., Contreras, D., and Pedroarena, C. (1998). The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1377):1841–1849.

Fornito, A., Zalesky, A., and Breakspear, M. (2013). Graph analysis of the human connectome: promise, progress, and pitfalls. *Neuroimage*, 80:426–444.

Fothergill, B. T., Knight, W., Stahl, B. C., and Ulnicane, I. (2019). Responsible data governance of neuroscience big data. *Frontiers in neuroinformatics*, 13:28.

Fukutomi, H., Glasser, M. F., Zhang, H., Autio, J. A., Coalson, T. S., Okada, T., Togashi, K., Van Essen, D. C., and Hayashi, T. (2018). Neurite imaging reveals microstructural variations in human cerebral cortical gray matter. *Neuroimage*, 182:488–499.

Ganglberger, F., Swoboda, N., Frauenstein, L., Kaczanowska, J., Haubensak, W., and Bühler, K. (2019). Braintrawler: A visual analytics framework for iterative exploration of heterogeneous big brain data. *Computers & Graphics*, 82:304–320.

Garyfallidis, E., Côté, M.-A., Rheault, F., Sidhu, J., Hau, J., Petit, L., Fortin, D., Cunanne, S., and Descoteaux, M. (2018). Recognition of white matter bundles using local and global streamline-based registration and clustering. *NeuroImage*, 170:283–295.

Garyfallidis, E., Ocegueda, O., Wassermann, D., and Descoteaux, M. (2015). Robust and efficient linear registration of white-matter fascicles in the space of streamlines. *NeuroImage*, 117:124–140.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

Ghorbani, A. and Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR.

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

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.

Glasser, M. F. and Van Essen, D. C. (2011). Mapping human cortical areas in vivo based on myelin content as revealed by t1-and t2-weighted mri. *Journal of neuroscience*, 31(32):11597–11616.

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22(2):133–139.

Gonzalez-Castillo, J. and Bandettini, P. A. (2018). Task-based dynamic functional connectivity: Recent findings and open questions. *Neuroimage*, 180:526–533.

Gonzalez-Castillo, J., Handwerker, D. A., Robinson, M. E., Hoy, C. W., Buchanan, L. C., Saad, Z. S., and Bandettini, P. A. (2014). The spatial structure of resting state connectivity stability on the scale of minutes. *Frontiers in neuroscience*, 8:138.

Goodale, S. E., Ahmed, N., Zhao, C., de Zwart, J. A., Özbay, P. S., Picchioni, D., Duyn, J., Englot, D. J., Morgan, V. L., and Chang, C. (2021). fmri-based detection of alertness predicts behavioral response variability. *Elife*, 10:e62376.

Goodman, Z. T., Bainter, S. A., Kornfeld, S., Chang, C., Nomi, J. S., and Uddin, L. Q. (2021). Whole-brain functional dynamics track depressive symptom severity. *Cerebral cortex*, 31(11):4867–4876.

Gordon, E. M., Laumann, T. O., Adeyemo, B., Gilmore, A. W., Nelson, S. M., Dosenbach, N. U., and Petersen, S. E. (2017). Individual-specific features of brain systems identified with resting state functional correlations. *NeuroImage*, 146:918–939.

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., and Petersen, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex*, 26(1):288–303.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9.

Greicius, M. D., Supekar, K., Menon, V., and Dougherty, R. F. (2009). Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cerebral cortex*, 19(1):72–78.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014). Ica-based artefact removal and accelerated fmri acquisition for improved resting state network imaging. *Neuroimage*, 95:232–247.

Gunning-Dixon, F. M. and Raz, N. (2000). The cognitive correlates of white matter abnormalities in normal aging: a quantitative review. *Neuropsychology*, 14(2):224.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Ha, S. and Lyu, I. (2022). Spharm-net: Spherical harmonics-based convolution for cortical parcellation. *IEEE Transactions on Medical Imaging*, 41(10):2739–2751.

Hamilton, J. P., Furman, D. J., Chang, C., Thomason, M. E., Dennis, E., and Gotlib, I. H. (2011). Default-mode and task-positive network activity in major depressive disorder: implications for adaptive and maladaptive rumination. *Biological psychiatry*, 70(4):327–333.

Hansen, C. B., Yang, Q., Lyu, I., Rheault, F., Kerley, C., Chandio, B. Q., Fadnavis, S., Williams, O., Shafer, A. T., Resnick, S. M., et al. (2021). Pandora: 4-d white matter bundle population-based atlases derived from diffusion mri fiber tractography. *Neuroinformatics*, 19(3):447–460.

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., Van De Lagemaat, L. N., Smith, K. A., Ebbert, A., Riley, Z. L., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399.

Heiervang, E., Behrens, T., Mackay, C. E., Robson, M. D., and Johansen-Berg, H. (2006). Between session reproducibility and between subject variability of diffusion mr and tractography measures. *Neuroimage*, 33(3):867–877.

Henry, N., Fekete, J.-D., and McGuffin, M. J. (2007). Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics*, 13(6):1302–1309.

Henson, R. N., Price, C. J., Rugg, M. D., Turner, R., and Friston, K. J. (2002). Detecting latency differences in event-related bold responses: application to words versus nonwords and initial versus repeated face presentations. *Neuroimage*, 15(1):83–97.

Hillman, E. M. (2014). Coupling mechanism and significance of the bold signal: a status report. *Annual review of neuroscience*, 37:161–181.

Hohenfeld, C., Werner, C. J., and Reetz, K. (2018). Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker? *NeuroImage: Clinical*, 18:849–870.

Huo, Y., Plassard, A. J., Carass, A., Resnick, S. M., Pham, D. L., Prince, J. L., and Landman, B. A. (2016). Consistent cortical reconstruction and multi-atlas brain segmentation. *NeuroImage*, 138:197–210.

Jack Jr, C. R., Barnes, J., Bernstein, M. A., Borowski, B. J., Brewer, J., Clegg, S., Dale, A. M., Carmichael, O., Ching, C., DeCarli, C., et al. (2015). Magnetic resonance imaging in alzheimer's disease neuroimaging initiative 2. *Alzheimer's & Dementia*, 11(7):740–756.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156.

Jeurissen, B., Tournier, J.-D., Dhollander, T., Connelly, A., and Sijbers, J. (2014). Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion mri data. *NeuroImage*, 103:411–426.

Jiang, C., Huang, J., Kashinath, K., Marcus, P., Niessner, M., et al. (2019). Spherical cnns on unstructured grids. *arXiv preprint arXiv:1901.02039*.

Jiang, Q., Zhang, Z. G., Ding, G. L., Silver, B., Zhang, L., Meng, H., Lu, M., Pourabdillah-Nejed-D, S., Wang, L., Savant-Bhonsale, S., et al. (2006). Mri detects white matter reorganization after neural progenitor cell treatment of stroke. *Neuroimage*, 32(3):1080–1089.

Jones, D. K. and Cercignani, M. (2010). Twenty-five pitfalls in the analysis of diffusion mri data. *NMR in Biomedicine*, 23(7):803–820.

Jönsson, D., Bergström, A., Forsell, C., Simon, R., Engström, M., Walter, S., Ynnerman, A., and Hotz, I. (2020). Visualneuro: A hypothesis formation and reasoning application for multi-variate brain cohort study data. In *Computer Graphics Forum*. Wiley Online Library.

Kassinopoulos, M. and Mitsis, G. D. (2019). Identification of physiological response functions to correct for fluctuations in resting-state fmri related to heart rate and respiration. *Neuroimage*, 202:116150.

Keiriz, J. J., Zhan, L., Ajilore, O., Leow, A. D., and Forbes, A. G. (2018). Neurocave: A web-based immersive visualization platform for exploring connectome datasets. *Network Neuroscience*, 2(3):344–361.

Kern, M., Lex, A., Gehlenborg, N., and Johnson, C. R. (2017). Interactive visual exploration and refinement of cluster assignments. *BMC bioinformatics*, 18(1):1–13.

Koban, L., Gianaros, P. J., Kober, H., and Wager, T. D. (2021). The self in context: brain systems linking mental and physical health. *Nature Reviews Neuroscience*, 22(5):309–322.

Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., et al. (2019). Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cerebral cortex*, 29(6):2533–2551.

Kong, R., Tan, Y. R., Wulan, N., Ooi, L. Q. R., Farahibozorg, S.-R., Harrison, S., Bijsterbosch, J. D., Bernhardt, B. C., Eickhoff, S., and Yeo, B. T. (2023). Comparison between gradients and parcellations for functional connectivity prediction of behavior. *NeuroImage*, 273:120044.

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X.-N., Spreng, N., Ge, T., Holmes, A., et al. (2021). Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral Cortex*, 31(10):4477–4500.

Kurzawski, J. W., Gulban, O. F., Jamison, K., Winawer, J., and Kay, K. (2022). Non-neural factors influencing bold response magnitudes within individual subjects. *Journal of Neuroscience*, 42(38):7256–7266.

Kwon, B. C., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W. F., and Perer, A. (2017). Clustervision: Visual supervision of unsupervised clustering. *IEEE transactions on visualization and computer graphics*, 24(1):142–151.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., and Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679.

Lancaster, J. L., Tordesillas-Gutiérrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., Mazziotta, J. C., and Fox, P. T. (2007). Bias between mni and talairach coordinates analyzed using the icbm-152 brain template. *Human brain mapping*, 28(11):1194–1205.

Lang, S., Duncan, N., and Northoff, G. (2014). Resting-state functional magnetic resonance imaging: review of neurosurgical applications. *Neurosurgery*, 74(5):453–465.

Lawes, I. N. C., Barrick, T. R., Murugam, V., Spierings, N., Evans, D. R., Song, M., and Clark, C. A. (2008). Atlas-based segmentation of white matter tracts of the human brain using diffusion tensor tractography and comparison with classical dissection. *Neuroimage*, 39(1):62–79.

Le Bihan, D., Poupon, C., Amadon, A., and Lethimonnier, F. (2006). Artifacts and pitfalls in diffusion mri. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 24(3):478–488.

Li, J., Bolt, T., Bzdok, D., Nomi, J. S., Yeo, B., Spreng, R. N., and Uddin, L. Q. (2019). Topography and behavioral relevance of the global signal in the human brain. *Scientific reports*, 9(1):1–10.

Liao, S. C., Hunt, E. A., and Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Annals Academy of Medicine Singapore*, 39(8):613.

Liu, T. T., Nalci, A., and Falahpour, M. (2017). The global signal in fmri: Nuisance or information? *NeuroImage*, 150:213–229.

Llinares-Benadero, C. and Borrell, V. (2019). Deconstructing cortical folding: genetic, cellular and mechanical determinants. *Nature Reviews Neuroscience*, 20(3):161–176.

L'Yi, S., Ko, B., Shin, D., Cho, Y.-J., Lee, J., Kim, B., and Seo, J. (2015). Xclusim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC bioinformatics*, 16(S11):S5.

Lynn, C. W. and Bassett, D. S. (2019). The physics of brain network structure, function and control. *Nature Reviews Physics*, 1(5):318–332.

Lyu, I., Kang, H., Woodward, N. D., Styner, M. A., and Landman, B. A. (2019). Hierarchical spherical deformation for cortical surface registration. *Medical image analysis*, 57:72–88.

Mandonnet, E., Sarubbo, S., and Petit, L. (2018). The nomenclature of human white matter association pathways: proposal for a systematic taxonomic anatomical classification. *Frontiers in neuroanatomy*, page 94.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., et al. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579.

Mckay, L. C., Evans, K. C., Frackowiak, R. S., and Corfield, D. R. (2003). Neural correlates of voluntary breathing in humans. *Journal of applied physiology*, 95(3):1170–1178.

Molle, L. and Coste, A. (2022). The respiratory modulation of interoception. *Journal of Neurophysiology*.

Mori, S., Crain, B. J., Chacko, V. P., and Van Zijl, P. C. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 45(2):265–269.

Mori, S., Oishi, K., Jiang, H., Jiang, L., Li, X., Akhter, K., Hua, K., Faria, A. V., Mahmood, A., Woods, R., et al. (2008). Stereotaxic white matter atlas based on diffusion tensor imaging in an icbm template. *Neuroimage*, 40(2):570–582.

Murphy, K., Birn, R. M., and Bandettini, P. A. (2013). Resting-state fmri confounds and cleanup. *Neuroimage*, 80:349–359.

Murugesan, S., Bouchard, K., Brown, J. A., Hamann, B., Seeley, W. W., Trujillo, A., and Weber, G. H. (2016). Brain modulyzer: interactive visual analysis of functional brain connectivity. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(4):805–818.

Nashiro, K., Min, J., Yoo, H. J., Cho, C., Bachman, S. L., Dutt, S., Thayer, J. F., Lehrer, P. M., Feng, T., Mercer, N., et al. (2022). Increasing coordination and responsivity of emotion-related brain regions with a heart rate variability biofeedback randomized trial. *Cognitive, Affective, & Behavioral Neuroscience*, pages 1–18.

Noble, S., Scheinost, D., and Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, 203:116157.

O'Donnell, L. J. and Westin, C.-F. (2007). Automatic tractography segmentation using a high-dimensional white matter atlas. *IEEE transactions on medical imaging*, 26(11):1562–1575.

Ogawa, S., Lee, T.-M., Kay, A. R., and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955.

Oouchi, H., Yamada, K., Sakai, K., Kizu, O., Kubota, T., Ito, H., and Nishimura, T. (2007). Diffusion anisotropy measurement of brain white matter is affected by voxel size: underestimation occurs in areas with crossing fibers. *American Journal of Neuroradiology*, 28(6):1102–1106.

Panesar, S. S. and Fernandez-Miranda, J. (2019). Commentary: the nomenclature of human white matter association pathways: proposal for a systematic taxonomic anatomical classification. *Frontiers in Neuroanatomy*, page 61.

Parvathaneni, P., Bao, S., Nath, V., Woodward, N. D., Claassen, D. O., Cascio, C. J., Zald, D. H., Huo, Y., Landman, B. A., and Lyu, I. (2019). Cortical surface parcellation using spherical convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 501–509. Springer.

Picchioni, D., Özbay, P. S., Mandelkow, H., de Zwart, J. A., Wang, Y., van Gelderen, P., and Duyn, J. H. (2022). Autonomic arousals contribute to brain fluid pulsations during sleep. *NeuroImage*, page 118888.

Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., Haselgrove, C., Helmer, K. G., Keator, D. B., Marcus, D. S., et al. (2012). Data sharing in neuroimaging research. *Frontiers in neuroinformatics*, 6:9.

Power, J. D., Lynch, C. J., Dubin, M. J., Silver, B. M., Martin, A., and Jones, R. M. (2020). Characteristics of respiratory measures in young adults scanned at rest, including systematic changes and "missed" deep breaths. *Neuroimage*, 204:116234.

Power, J. D., Lynch, C. J., Silver, B. M., Dubin, M. J., Martin, A., and Jones, R. M. (2019). Distinctions among real and apparent respiratory motions in human fmri data. *NeuroImage*, 201:116041.

Power, J. D., Plitt, M., Laumann, T. O., and Martin, A. (2017). Sources and implications of whole-brain fmri signals in humans. *Neuroimage*, 146:609–625.

Preti, M. G., Bolton, T. A., and Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage*, 160:41–54.

Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B., and Davatzikos, C. (2003). Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *Journal of Neuroscience*, 23(8):3295–3301.

Rheault, F., De Benedictis, A., Daducci, A., Maffei, C., Tax, C. M., Romascano, D., Caverzasi, E., Morency, F. C., Corrivetti, F., Pestilli, F., et al. (2020). Tractostorm: The what, why, and how of tractography dissection reproducibility. *Human Brain Mapping*, 41(7):1859–1874.

Rheault, F., St-Onge, E., Sidhu, J., Chenot, Q., Petit, L., and Descoteaux, M. (2018). Bundle-specific tractography. In *Computational Diffusion MRI*, pages 129–139. Springer.

Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J., Burgess, G. C., Harms, M. P., Smith, S. M., Van Essen, D. C., and Jenkinson, M. (2014). Msm: a new flexible framework for multimodal surface matching. *Neuroimage*, 100:414–426.

Ruta, N., Sawada, N., McKeough, K., Behrisch, M., and Beyer, J. (2019). Sax navigator: Time series exploration through hierarchical clustering. In *2019 IEEE Visualization Conference (VIS)*, pages 236–240. IEEE.

Sacha, D., Kraus, M., Bernard, J., Behrisch, M., Schreck, T., Asano, Y., and Keim, D. A. (2017a). Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE transactions on visualization and computer graphics*, 24(1):120–130.

Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., and Keim, D. A. (2017b). What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, 268:164–175.

Salas, J. A., Bayrak, R. G., Huo, Y., and Chang, C. (2021). Reconstruction of respiratory variation signals from fmri data. *NeuroImage*, 225:117459.

Salehi, M., Greene, A. S., Karbasi, A., Shen, X., Scheinost, D., and Constable, R. T. (2020). There is no single functional atlas even for a single individual: Functional parcel definitions change with task. *NeuroImage*, 208:116366.

Salehi, M., Karbasi, A., Shen, X., Scheinost, D., and Constable, R. T. (2018). An exemplar-based approach to individualized parcellation reveals the need for sex specific functional networks. *NeuroImage*, 170:54–67.

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

Schilling, K. G., Rheault, F., Petit, L., Hansen, C. B., Nath, V., Yeh, F.-C., Girard, G., Barakovic, M., Rafael-Patino, J., Yu, T., et al. (2021a). Tractography dissection variability: what happens when 42 groups dissect 14 white matter bundles on the same dataset? *NeuroImage*, 243:118502.

Schilling, K. G., Tax, C. M., Rheault, F., Hansen, C., Yang, Q., Yeh, F.-C., Cai, L., Anderson, A. W., and Landman, B. A. (2021b). Fiber tractography bundle segmentation depends on scanner effects, vendor effects, acquisition resolution, diffusion sampling scheme, diffusion sensitization, and bundle segmentation workflow. *NeuroImage*, 242:118451.

Schmidt, M. F., Storrs, J. M., Freeman, K. B., Jack Jr, C. R., Turner, S. T., Griswold, M. E., and Mosley Jr, T. H. (2018). A comparison of manual tracing and freesurfer for estimating hippocampal volume over the adult lifespan. *Human brain mapping*, 39(6):2500–2513.

Schreck, T., Bernard, J., Von Landesberger, T., and Kohlhammer, J. (2009). Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8(1):14–29.

Seo, J. and Shneiderman, B. (2002). Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86.

Sharma, P. and Singh, A. (2017). Era of deep neural networks: A review. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Shen, X., Tokoglu, F., Papademetris, X., and Constable, R. T. (2013). Groupwise whole-brain parcellation from resting-state fmri data for network node identification. *Neuroimage*, 82:403–415.

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral cortex*, 22(1):158–165.

Shmueli, K., van Gelderen, P., de Zwart, J. A., Horovitz, S. G., Fukunaga, M., Jansma, J. M., and Duyn, J. H. (2007). Low-frequency fluctuations in the cardiac rate as a source of variance in the resting-state fmri bold signal. *Neuroimage*, 38(2):306–320.

Shokri-Kojori, E., Tomasi, D., and Volkow, N. D. (2018). An autonomic network: synchrony between slow rhythms of pulse and brain resting state is associated with personality and emotions. *Cerebral Cortex*, 28(9):3356–3371.

Skudlarski, P., Jagannathan, K., Calhoun, V. D., Hampson, M., Skudlarska, B. A., and Pearlson, G. (2008). Measuring brain connectivity: diffusion tensor imaging validates resting state temporal correlations. *Neuroimage*, 43(3):554–561.

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., et al. (2013). Resting-state fmri in the human connectome project. *Neuroimage*, 80:144–168.

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., and Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565–1567.

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4):e42.

Suárez, L. E., Markello, R. D., Betzel, R. F., and Misic, B. (2020). Linking structure and function in macroscale brain networks. *Trends in Cognitive Sciences*, 24(4):302–315.

Thomas Yeo, B., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165.

Tian, Y., Margulies, D. S., Breakspear, M., and Zalesky, A. (2020). Hierarchical organization of the human subcortex unveiled with functional connectivity gradients. *bioRxiv*.

Tong, Y., Hocke, L. M., and Frederick, B. B. (2019). Low frequency systemic hemodynamic "noise" in resting state bold fmri: characteristics, causes, implications, mitigation strategies, and applications. *Frontiers in neuroscience*, page 787.

Tournier, J.-D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., Christiaens, D., Jeurissen, B., Yeh, C.-H., and Connelly, A. (2019). Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage*, 202:116137.

Tu, W. and Zhang, N. (2022). Neural underpinning of a respiration-associated resting-state fmri network. *ELife*, 11:e81555.

Tuch, D. S. et al. (2002). *Diffusion MRI of complex tissue structure*. PhD thesis, Massachusetts Institute of Technology.

Uddin, L. Q., Supekar, K., and Menon, V. (2010). Typical and atypical development of functional human brain networks: insights from resting-state fmri. *Frontiers in systems neuroscience*, 4:1447.

Van Den Heuvel, M., Mandl, R., Luigjes, J., and Pol, H. H. (2008). Microstructural organization of the cingulum tract and the level of default mode functional connectivity. *Journal of Neuroscience*, 28(43):10844–10851.

Van Den Heuvel, M. P. and Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., et al. (2012). The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231.

Van Houdt, G., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955.

Veenith, T. V., Carter, E., Grossac, J., Newcombe, V. F., Outtrim, J. G., Lupson, V., Williams, G. B., Menon, D. K., and Coles, J. P. (2013). Inter subject variability and reproducibility of diffusion tensor imaging within and between different imaging sessions. *PloS one*, 8(6):e65941.

Wakana, S., Caprihan, A., Panzenboeck, M. M., Fallon, J. H., Perry, M., Gollub, R. L., Hua, K., Zhang, J., Jiang, H., Dubey, P., et al. (2007). Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage*, 36(3):630–644.

Wang, D., Buckner, R. L., Fox, M. D., Holt, D. J., Holmes, A. J., Stoecklein, S., Langs, G., Pan, R., Qian, T., Li, K., et al. (2015). Parcellating cortical functional networks in individuals. *Nature neuroscience*, 18(12):1853.

Wasserthal, J., Neher, P., and Maier-Hein, K. H. (2018). Tractseg-fast and accurate white matter tract segmentation. *NeuroImage*, 183:239–253.

Wise, R. G., Ide, K., Poulin, M. J., and Tracey, I. (2004). Resting fluctuations in arterial carbon dioxide induce significant low frequency variations in bold signal. *Neuroimage*, 21(4):1652–1664.

WU-Minn, H. (2017). 1200 subjects data release reference manual. *URL https://www. humanconnectome. org*.

Xifra-Porxas, A., Kassinopoulos, M., and Mitsis, G. D. (2021). Physiological and motion signatures in static and time-varying functional connectivity and their subject identifiability. *Elife*, 10:e62324.

Xu, R., Thomas, M. M., Leow, A., Ajilore, O. A., and Forbes, A. G. (2019). Tempocave: Visualizing dynamic connectome datasets to support cognitive behavioral therapy. In *2019 IEEE Visualization Conference (VIS)*, pages 186–190. IEEE.

Yang, W., Wang, X., Lu, J., Dou, W., and Liu, S. (2020). Interactive steering of hierarchical clustering. *IEEE Transactions on Visualization and Computer Graphics*.

Yang, X., Shi, L., Daianu, M., Tong, H., Liu, Q., and Thompson, P. (2016). Blockwise human brain network visual comparison using nodetrix representation. *IEEE transactions on visualization and computer graphics*, 23(1):181–190.

Yeh, F.-C. (2020). Shape analysis of the human association pathways. *Neuroimage*, 223:117329.

Yeh, F.-C., Panesar, S., Fernandes, D., Meola, A., Yoshino, M., Fernandez-Miranda, J. C., Vettel, J. M., and Verstynen, T. (2018). Population-averaged atlas of the macroscale human structural connectome and its network topology. *Neuroimage*, 178:57–68.

Yeh, F.-C., Wedeen, V. J., and Tseng, W.-Y. I. (2010). Generalized q-sampling imaging. *IEEE transactions on medical imaging*, 29(9):1626–1635.

Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T., et al. (2011). Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Frontiers in neuroinformatics*, 5:23.

Yoo, H. J., Nashiro, K., Min, J., Cho, C., Bachman, S. L., Nasseri, P., Porat, S., Dutt, S., Grigoryan, V., Choi, P., et al. (2022). Heart rate variability (hrv) changes and cortical volume changes in a randomized trial of five weeks of daily hrv biofeedback in younger and older adults. *International Journal of Psychophysiology*, 181:50–63.

Yoo, S. W., Guevara, P., Jeong, Y., Yoo, K., Shin, J. S., Mangin, J.-F., and Seong, J.-K. (2015). An example-based multi-atlas approach to automatic labeling of white matter tracts. *PloS one*, 10(7):e0133337.

Zelano, C., Jiang, H., Zhou, G., Arora, N., Schuele, S., Rosenow, J., and Gottfried, J. A. (2016). Nasal respiration entrains human limbic oscillations and modulates cognitive function. *Journal of Neuroscience*, 36(49):12448–12467.

Zhang, F., Karayumak, S. C., Hoffmann, N., Rathi, Y., Golby, A. J., and O'Donnell, L. J. (2020a). Deep white matter analysis (deepwma): Fast and consistent tractography segmentation. *Medical image analysis*, 65:101761.

Zhang, F., Wu, Y., Norton, I., Rathi, Y., Golby, A. J., and O'Donnell, L. J. (2019). Test–retest reproducibility of white matter parcellation using diffusion mri tractography fiber clustering. *Human brain mapping*, 40(10):3041–3057.

Zhang, J., Kucyi, A., Raya, J., Nielsen, A. N., Nomi, J. S., Damoiseaux, J. S., Greene, D. J., Horovitz, S. G., Uddin, L. Q., and Whitfield-Gabrieli, S. (2021a). What have we really learned from functional connectivity in clinical populations? *NeuroImage*, 242:118466.

Zhang, R.-Y., Wei, X.-X., and Kay, K. (2020b). Understanding multivariate brain activity: evaluating the effect of voxelwise noise correlations on population codes in functional magnetic resonance imaging. *PLOS Computational Biology*, 16(8):e1008153.

Zhang, S., Spoletini, L. J., Gold, B. P., Morgan, V. L., Rogers, B. P., and Chang, C. (2021b). Interindividual signatures of fmri temporal fluctuations. *Cerebral Cortex*, 31(10):4450–4463.

Zhang, Y., Kim, J.-H., Brang, D., and Liu, Z. (2021c). Naturalistic stimuli: A paradigm for multiscale functional characterization of the human brain. *Current Opinion in Biomedical Engineering*, 19:100298.

Zou, Q.-H., Zhu, C.-Z., Yang, Y., Zuo, X.-N., Long, X.-Y., Cao, Q.-J., Wang, Y.-F., and Zang, Y.-F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (alff) for resting-state fmri: fractional alff. *Journal of neuroscience methods*, 172(1):137–141.