The Pennsylvania State University
The Graduate School

MACHINE-LEARNING-BASED FUNCTIONAL CONNECTIVITY ANALYSES: CHALLENGES AND OPPORTUNITIES

A Dissertation in
Neuroscience
by
Shlomit Gur

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2019

The dissertation of Shlomit Gur was reviewed and approved* by the following:

Vasant Gajanan Honavar
Professor and Edward Frymoyer Chair of Information Sciences and Technology
Dissertation Advisor, Chair of Committee

John Yen
Professor of Information Sciences and Technology

Michele Theresa Diaz
Associate Professor of Psychology and Linguistics

Michael Nelson Hallquist
Assistant Professor of Psychology

Kevin Douglas Alloway
Professor of Neural and Behavioral Sciences
Co-Chair, Intercollege Graduate Degree Program in Neuroscience

*Signatures are on file in the Graduate School.

Abstract

Over the past decade, substantial progress has been made in neuroscience in general and in neuroimaging in particular. This progress has been propelled by many factors, including technological advances, the development of theory, methodology, and tools, and making data sharing a practice in neuroimaging research. Consequently, neuroimaging data have increased in volume and complexity, which according to some researchers marks a new era in the field, a big-data era. Coincidentally, one of the more prominent directions to have been espoused by researchers in the field is that of functional connectivity analysis and interpretation, using machine learning techniques. This direction, or sub-field, represents the intersection between three independent fields: neuroscience, network science, and machine learning. To date, the application of machine learning techniques to functional connectivity data has been dominated primarily by neuroscientists with limited expertise in machine learning and network science, and computer scientists with limited understanding of the data's domain. However, we postulate that cross-talk between the fields is imperative for the contribution and progress of the sub-field. E.g., methods should be developed or adjusted to meet domain-specific needs, data in the domain should be curated to meet requirements of methods of interest, and proper use of both methods and data should be ensured. Against this background, this dissertation examines the current state of functional connectivity analyses, identifies challenges and opportunities, and addresses them with the application and development of domain-aware machine learning techniques. More specifically, the challenges and opportunities in the present dissertation pertain to: (i) population-condition interactions in static functional connectivity, (ii) multi-site static functional connectivity repositories, and (iii) evolving functional connectivity from different participants.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AA	Amino Acid
AAL	Automated Anatomical Labeling
ABIDE	Autism Brain Imaging Data Exchange
AND	Average Neighbor Degree
ASD	Autism Spectrum Disorder
BA	Barabasi-Albert, p. 74
BCM	Betweenness Centrality Measure
BOLD	Blood Oxygen Level-Dependent
C-PAC	Configurable Pipeline for the Analysis of Connectomes
CC	Clustering Coefficient
CT	Computed Tomography
CV	Cross-Validation
DM	Dorogovtsev-Mendes, p. 74
EC	Eigen Centrality
EEG	Electroencephalography
ER	Erdos-Renyi, p. 74
EZ	Eickhoff-Zilles

FC	Functional Connectivity, p. 1
FO	Frontal Operculum cortex
fMRI	Functional Magnetic Resonance Imaging
FP	Frontal Pole
GLM	General Linear Model
GNB	Gaussian Naïve Bayes, p. 44
GNBMM	Gaussian Naïve Bayes Mixture Model, p. 94
HO	Harvard-Oxford
HRF	Hemodynamic Response Function
HSIC	Hilbert-Schmidt Independence Criterion
KTST	Kernel Two-Sample Test, p. 46
LOO	Leave-One-Out
LOPO	Leave-One-Participant-Out
MEG	Magnetoencephalography
ML	Machine Learning, p. 3
MMD	Maximum Mean Discrepancy, p. 45
MNI	Montreal Neurological Institute
NB	Naïve Bayes
NSSD	Normalized Sum Squared Difference
OSN	Ordered Sequence of networks
PATENet	Pairwise Alignment of Temporaly-Evolving Networks, p. 63
PCP	Preprocessed Connectomes Project
RF	Random Forest
RF50	Random Forest with 50 trees

ROI	Region Of Interest
SD	Standard Deviation
SDCIT	Self-Discrepancy Conditional Independence Test
SNNSM	Signed Normalized Network Similarity Measure, p. 65
SP	Statistical Power
SW	Smith-Waterman, p. 65
TD	Typically Developing
TEN	Time-Evolving Network
TOT	Tip Of the Tongue
TT	Talaraich and Tournoux
UNNSM	Unsigned Normalized Network Similarity Measure, p. 65
WCC	Weighted Clustering Coefficient
WD	Weighted Degree

Acknowledgments

First, I would like to thank my PhD adviser, Dr. Vasant Honavar, for his guidance, support, constructive criticism, and patience during my PhD studies. His open-mindedness to take on a Neuroscience PhD student, and his pursuit and facilitation of inter-disciplinary collaborations are appreciable. I have learned a lot from him and from the tools, resources, and opportunities he put at my disposal, for which I am extremely grateful. Under his guidance, I was able to attain a good balance between thorough and focused research. Through it all, Dr. Honavar encouraged me to explore on my own and allowed me to make mistakes, and in the words of Oscar Wilde, *"Experience [...] was merely the name we gave to our mistakes"* (The Picture of Dorian Gray).

I would also like to thank my doctoral committee, Dr. John Yen, Dr. Michele Diaz, and Dr. Michael Hallquist. Their different, yet complementing, expertise have greatly contributed to the inter-disciplinary nature of my work. I am grateful for their time, support, feedback, and helpful suggestions, all of which I feel have improved the quality of my work in general and of this dissertation is particular. I would also like to express my gratitude to past and present lab-mates in the Artificial Intelligence Research Laboratory at Pennsylvania State University, for helpful discussions and suggestions.

Additionally, I would like to thank my family in Israel, who supported me throughout this journey and long before it. My parents, Baruch and Elizabeth, who encouraged my curiosity, provided me with opportunities, and learned to accept my over-independence. My sisters, Yaara, Netanela, and Moriya, without whom I probably would not be the person I am today. My niece, Naomi, and nephews, Alon, Daniel, and Jonathan, for being their amazing selves, and always putting a smile on their "roboaunt"'s face. I also want to thank my grandma, grandpa, oma and opa, who exemplified perseverance, morality, and humanity even under the most difficult of circumstances. I literally would not be here if it was not for them.

Finally, much of the research reported in this dissertation depended on pre-existing data. I would like to thank the researchers who generously shared with me

their carefully designed and curated data, invaluable knowledge, and precious time. In particular, I thank the staff and scientists at the Brain Imaging and Analysis Center, Duke University School of Medicine (where the data analyzed in chapter 3 were collected and preprocessed), and the staff and scientists at the Laboratory of Neurocognitive Development, Pittsburgh University (where the longitudinal fMRI data in adolescents and young adults were collected and preprocessed).

The research in this dissertation was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health through Grant UL1 TR000127 and TR002014, by the National Science Foundation, through Grant SHF 1518732, by the Center for Big Data Analytics and Discovery Informatics at Pennsylvania State University, by the National Institutes of Health's National Institute on Aging grant R01 AG034138, by the Edward Frymoyer Endowed Professorship in Information Sciences and Technology at Pennsylvania State University, and by the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science funded by the Pratiksha Trust at the Indian Institute of Science [both held by Vasant Honavar]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

Dedication

To my parents, sisters, niece, and nephews.

Chapter 1
Introduction

1.1 Neuroimaging

As we try to uncover and understand underlying mechanisms of different functions and disorders in the brain, in-vivo recordings are of great importance. Due to their invasive nature, traditional in-vivo recordings (i.e., in-vivo microelectrode recordings) have been collected primarily in nonhuman primates and animal models [1,2]. The emergence of noninvasive brain imaging techniques has allowed the scientific community to study the human brain in-vivo on a larger scale. Noninvasive techniques include the earlier electroencephalography (EEG; first human recording acquired in 1924 [3]), magnetoencephalography (MEG; introduced in the 1960s [4]), and computed tomography (CT) scans (introduced into medical practice in the 1970s [5]), as well as the more recent diffuse optical imaging (first used to image a human brain in 1995 [6]) and functional magnetic resonance imaging (fMRI; first explored in humans in 1991 [7]). The different methods provide different trade-offs between the temporal and spatial resolutions [8] (Fig. 1.1).

1.1.1 Functional Connectivity

Functional connectivity (FC) is the representation of fMRI data in the form of undirected graphs, where nodes typically represent anatomical, functional, or anatomical-functional hybrid brain regions, and edges describe a temporally-based relationship between these regions [9, 10]. FC networks are typically considered static per neu-

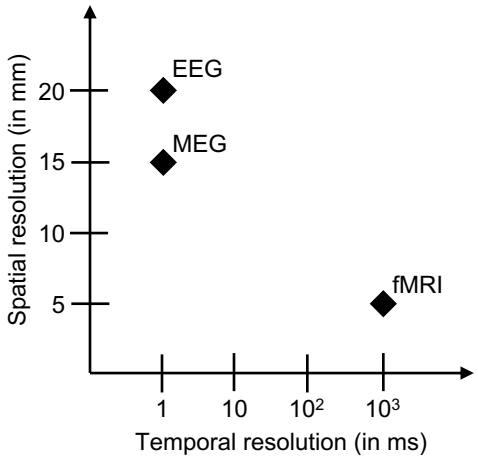

Figure 1.1. Trade-off between temporal and spatial resolution of three common noninvasive neuroimaging techniques. EEG = electroencephalography; MEG = magnetoencephalography; fMRI = functional magnetic resonance imaging.

roimaging session, due to the considerably low temporal-resolution of fMRI (Fig. 1.1). While some studies use dynamic network representation per session (e.g., [11]), we use the more traditional static network per session in the present work. FC has been shown to be a useful fMRI representation, as numerous studies revealed relationships between differences in FC and other observed differences, such as behavior, age, and disease (e.g., [12–14]). These relationships provide insight into potential underlying mechanisms of the observed differences.

1.1.1.1 Network Neuroscience

Over the last decade there have been many developments in the field of FC, such as increased temporal-resolution fMRI (e.g., [15]), a spike in the number of functional atlases and parcellation methods (e.g., [16, 17]), and automated meta-analysis (e.g., [18]), to name a few. Additionally, technological advances in general, and greater availability of computational resources and power in particular, have lead to an increase in volume and complexity of neural data, collected and shared. Data are now made available in repositories and shared not only in raw format, but also post-preprocessing using multiple pipelines (e.g., [19]). Consequently, advanced data analysis and modeling techniques are in order. Network neuroscience pertains to the

theoretical and computational advances in data analysis and modeling, specific to the domain of advanced neurobiological data, including neuroimaging [20].

The analysis and modeling in network neuroscience have different objectives. For example, one objective is to examine and understand the dynamics of different neural networks (e.g., FC; see section 2.2), and another is prediction (e.g., prediction of network dynamics or classification of networks; see section 2.1), which in itself can serve different purposes (e.g., identifying biomarkers and understanding the effect of external parameters such as age or intervention).

1.2 Machine Learning

Machine learning (ML) is a sub-field of artificial intelligence that includes techniques for statistics-based learning from data. With a variety of learning objectives, the most common categories of ML techniques include supervised learning (e.g., classification and regression), unsupervised learning (e.g., clustering and anomaly detection), and reinforcement learning [21]. In the present work we use primarily supervised learning techniques and more specifically, classification algorithms.

1.2.1 ML Applications in FC

Over the past decade ML has gained popularity in analysis and interpretation of fMRI data [22]. Much like other domains, fMRI has also taken on network representation [23,24], such as FC Networks (see section 1.1.1) and effective connectivity networks [25]. The combination of FC and ML techniques has been proven useful in providing insight into differences related to observable variables such as behavior, age, and disease (e.g., [12–14, 26, 27]). Noted for its contribution and further potential, the combination has been recently included as part of an emerging discipline, "network neuroscience" [20, 28] (see section 1.1.1.1).

Traditionally, ML techniques applied to FC were simple, often linear, classifiers such as linear support vector machine classifiers [22, 29]. This was in part due to prioritizing the interpretability of the models over their performance. Another reason was that FC data more often than not have far more features than samples, and

more complex models have more parameters, which in turn require more training instances as compared to simpler models, as well as more in-depth understanding of the models and theory and manual tuning of hyper-parameters. As data sharing has not been common-practice in the field of neuroscience, models were often applied by neuroscientists, who not only lacked ML expertise, but also enforced the prioritization of interpretability of the results. However, in recent years, with the emergence of open-access FC datasets (e.g., [30, 31]) and the revived popularity of deep learning methods, more researchers with ML expertise have become involved in FC analysis and have introduced often highly complex models (e.g., long short-term memory networks [32] and graph convolutional network [33]), prioritizing performance over interpretability. In the present work we focus on interpretable models.

1.2.1.1 Performance Evaluation

Due to the small sample sizes in FC data, researchers often use leave-one-out (LOO) cross-validation (CV) [22]. LOO CV minimizes bias at the expense of variance in performance, while keeping the results independent of each other, thereby allowing for meaningful mean and standard deviation (SD). However, from a ML stand point, k-fold CV with $k \in \{5, 10\}$, for example, is preferable to LOO CV, as it is more "challenging" for the model and therefore more indicative of the model's performance and fit for the data and task. Additionally, it is recommended to use stratified CV whenever possible.

The recent practice of fMRI data sharing and the emergence of multi-site fMRI repositories (e.g., [30, 31, 34–39]), have allowed researchers to examine the question of generalizability of results. While traditionally, a single FC sample was used per study, generalizability studies often compare the performance of similar or identical models on different samples of different sizes. Test sets in these FC samples are often fairly small and even more-so with k-fold CV (e.g., [40, 41]). Therefore, for the results per sample to be comparable, they should be corrected. The results can be corrected, for example, by applying Laplace "add-one" smoothing to the confusion matrices before using it to compute any performance measure.

Finally, there is an abundance of performance measurements, not all of which are suited for all scenarios [42]. For example, for unbalanced data with only 10% of the

observations belonging to the target class, accuracy may not be a good measure, as a non-informative model that classifies all observations as the non-target class would have an accuracy of 90%. FC studies commonly use accuracy to evaluate models, even if sensitivity and specificity are reported in addition. Notice that even for balanced data, accuracy gives equal weight to type I and type II errors, which might be desirable in some cases, but might not be in others. Sensitivity is the recall of the target class ($\frac{TP}{TP+FN}$) and is focused on keeping type II error to a minimum, while specificity is defined as the recall of the non-target class ($\frac{TN}{TN+FP}$) and is therefore focused on keeping type I error to a minimum. However, F_1-score of the target class ($\frac{2 \cdot TP}{2 \cdot TP+FN+FP}$) is more commonly used by ML experts, as it is a simple measurement, yet it gives a balance between precision ($\frac{TP}{TP+FP}$ for the target class) and recall of the target class and is not affected by unbalanced data.

1.3 Network Science

Networks are a useful way to describe complex relational data, as often is the case with real-world data. Network science has provided a variety of powerful tools for describing, representing, and analyzing a variety of real-world systems, including social networks, the internet, neural networks, and biomolecular networks [43, 44]. Real-world networks are generally perceived to be dynamic, and static networks can be viewed as their state at a particular time point. Static networks are often characterized using topological features (see section 1.3.1), while network dynamics can pertain to the changes in topological features over time or to diffusion on the network (i.e., the spread across the network, of an idea or a disease, for example) [45, 46].

1.3.1 Network Topological Features

The topological structure of a network can be described using different features that pertain to the properties of the network as a whole or to the nodes in the network in local and global contexts [46, 47]. Features at the network level include, but are not limited to, its size (number of nodes), its density (the number of existing edges out of the number of possible edges in a network of the same size), and node features

averaged across all nodes in the network. Common node features attempt to capture the significance of a node in the context of its local or global connectivity. Node features include, but are not limited to, degree (the number of immediate neighbors), betweenness centrality (portion of shortest paths crossing the node), and clustering coefficient (connectivity between immediate neighbors). Intermediate level features may be used to describe cliques of nodes, modules, or other types of sub-networks. For example, participation coefficient (the distribution of a node's connectivity to different sub-networks) is an intermediate level feature.

1.4 Fundamental Research Objectives in FC Analysis

In this dissertation, we employ informed development and application of ML techniques in the domain of FC. First, we address characteristics of existing and emerging data in the domain of FC, identifying limitations and opportunities. Then, we address a selection of these issues using ML techniques, informed by domain knowledge.

1.4.1 Uncovering Population-Dependent and -Independent Condition-Predictive Features in Static FC

Neurological disorders, aging, etc. are often associated with observable differences (e.g., accuracy, efficiency, and response time) in performing particular cognitive tasks as compared to a respective control population. Thus, in order to understand the characteristics of a target population, it is often useful to have not only a control population, but also a control task. That is, in addition to a task in which the populations differ in performance, have a cognitively-related task in which their performance is comparable. The two-way comparison may give better insight into altered, as well as preserved, cognitive processes.

In the present work, we use the term static FC to refer to a single FC network constructed from a single fMRI recording session. In order to uncover population-dependent and -independent condition-predictive features in static FC, we use task-based fMRI for two conditions: (1) performing a task with population-distinct perfor-

mance and (2) performing a task with population-comparable performance. The aim of this objective is two-fold: first, to identify features that can reliably discriminate between conditions, and second, to examine whether these features differ between the populations. We address this objective in chapter 3.

1.4.2 Combining Multi-Site Static FC for Sample Size Augmentation

In recent years, the modest sample size of fMRI studies has been heavily criticized (e.g., [48, 49]), calling into question the reproducibility and generalizability of these studies and their findings. As a result, multiple efforts have been made to encourage unrestricted public sharing of fMRI data (e.g., [50, 51]). Ensuing initiatives include the creation of fMRI repositories in order to share and aggregate multiple independent data samples from different sites [52] (e.g., [53,54]). While these repositories present an opportunity to augment fMRI sample sizes, they also introduce challenges, such as new confounding variables to be factored in when combining samples from different sites. Samples from different sites may vary in equipment, acquisition protocol, instructions to participants (e.g., eyes open or closed), demographics (e.g., age, severity of studied condition, and used medication), and other variables [55]. In chapter 4 we use data from the Autism Brain Imaging Data Exchange (ABIDE) Preprocessed Connectomes Project (PCP) repository as a case study to examine the opportunities and challenges associated with combining static FC samples from multiple sites to augment sample sizes.

1.4.3 Temporal Alignment of Evolving FC from Different Participants

In recent years, large-scale longitudinal fMRI data have become available [14], presenting opportunities such as studying temporal evolution of FC within an individual [56]. However, alongside these exciting opportunities, longitudinal FC data also present domain-specific challenges (e.g., inter-individual differences and the variable time intervals in the data). These challenges should be well understood and carefully

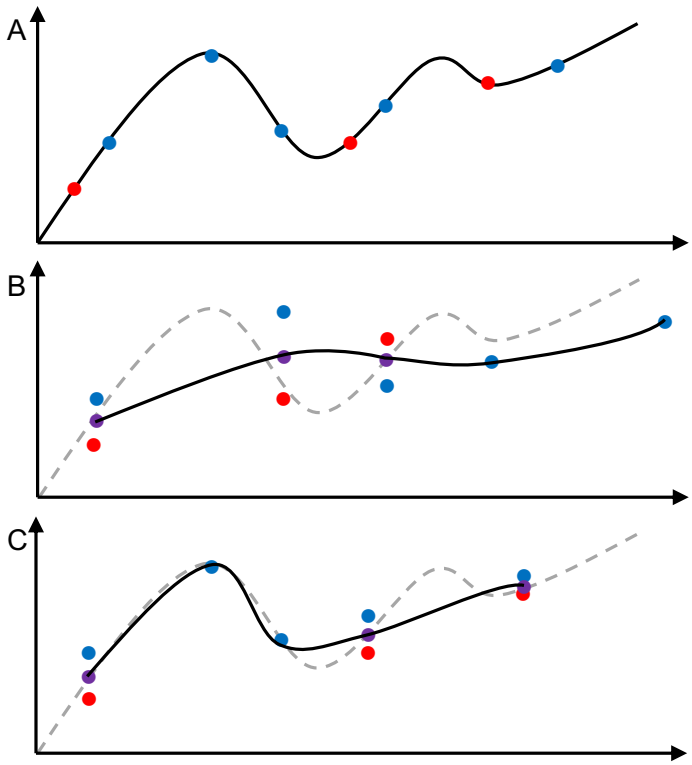

Figure 1.2. The importance of temporal alignment. 2D example of curve estimation from two longitudinal samples (horizontal axis is temporal). Notice the temporal value of the data points might not be available in samples. (A) The original curve and the two samples (Red dots = sample a; Blue dots = sample b). (B) The estimated curve (black solid line) without alignment: assuming $t_{a,i} = t_{b,i}\ \forall 1 \leq i \leq \min\{n, m\} \in \mathbb{N}$ where n and m are the number of data points in samples a and b, respectively, and $t_{a,i}$ and $t_{b,i}$ are the temporal values of the i_{th} observations from samples a and b, respectively. (C) The estimated curve (black solid line) with alignment. Purple dots = average of the "matched" data points from samples a and b. Dashed gray line = the original curve (black solid line in A).

addressed in order to get meaningful results.

We postulate that prior to combining longitudinal fMRI recordings from different participants, the data points from the different participants should be aligned in reference to the relevant temporal axes (e.g., age in years for healthy development and onset and relevant milestones for disease progression). Fig. 1.2 illustrates the importance of alignment in a 2D example, demonstrating how alignment can help to better estimate the ground-truth curve. Thus, in chapter 5 we describe in greater

detail some of the challenges associated with longitudinal fMRI data and propose a novel family of algorithms (see section 5.2) to align evolving FC networks from pairs of participants.

Chapter 2
Background

2.1 Prediction in Network Neuroscience

One of the objectives of network neuroscience (see section 1.1.1.1) is prediction, which includes prediction of networks and prediction on networks. Prediction may refer to prediction of network dynamics (e.g., given a static resting-state FC network of a healthy individual, predict what it will look like a year later) or prediction of network or sub-network properties, associated with a condition of interest (e.g., increased functional connectivity between the subgenual cingulate and thalamus and the default-mode network in individuals suffering from depression, as compared to healthy controls [57]). It should be noted that in neuroscience, prediction is often a mean to an end itself, rather than the end. For example, prediction may serve as an intermediate step in identifying biomarkers or understanding the effect of external parameters (e.g., age or intervention).

To address the first research objective in the present work (see section 1.4.1), we target time-agnostic prediction using network neuroscience and ML techniques. While some approaches using ML techniques may be black-boxed [20], we focus on interpretable methodology. As this research objective pertains to static FC, we use prediction to link FC network properties to a condition of interest, to gain insight into that condition.

2.2 Network Dynamics in Network Neuroscience

Another goal of network neuroscience (see section 1.1.1.1) is to examine and understand the dynamics of different neural networks. Network dynamics may refer to "dynamics on networks", "dynamics of networks", or "dynamics on and of networks". Dynamics on networks refers to diffusion on networks (i.e., changes in node properties, such as activation, over time, while edges remain unchanged). Dynamics of networks refers to changes in connectivity (e.g., edges' existence, absence, or weight) over time, while the nodes and their properties remain unchanged. Dynamics on and of networks refers to changes in both node properties and connectivity over time.

Currently, there are two major types of studies that are used to assess network dynamics in network neuroscience:

1. cross-sectional studies, which assume representative sample populations and examine differences between them (e.g., [58, 59])

2. longitudinal individual-level studies, which provide neuroimaging recordings from the same individuals at different time points (e.g., [60, 61])

Neither type of study is intrinsically better than the other, with both having strengths and weaknesses. For example, cross-sectional studies often require fairly great temporal spaces between the sampled populations to guarantee separation, but data acquisition does not require waiting for the participants to undergo the temporal changes in question. In contrast, longitudinal individual-level studies capture actual temporal changes, as they examine the same brain over time, but they take time to acquire, as they require waiting for the temporal changes in question to take place in each of the individuals. Consequently, large-scale longitudinal fMRI data have become available only in recent years [14].

2.2.1 Network Dynamics of Longitudinal fMRI

Longitudinal fMRI data present new opportunities, as well as some unique domain-specific challenges. One of these challenges is the aggregation of individual-level data for the purpose of generalization to other individuals. There are inter-individual

differences in brain development [62], and one might be wrong to assume, for example, that the FC of two 12-year-old healthy male individuals is comparable. Addressing this challenge is highly critical in settings such as development during adolescence, for example. We therefore propose a novel family of algorithms to align layers between pairs of multilayer graphs in chapter 5. These algorithms can be applied, for example, to multilayer graph representation of evolving FC from different individuals. We postulate that such alignment would lead to more accurate comparison between individuals, and would therefore lead to more meaningful results. The currently-available longitudinal datasets are still fairly small (see appendix C), thereby leaving the application of the algorithms to real-world longitudinal FC data outside the scope of the present dissertation.

Chapter 3
Condition- and Population-Effects in Static FC

In this chapter we address the use of ML techniques to uncover condition- and population-effects in static FC. We use a case study of language-related cognitive processes in younger and older adults [63]. Specifically, we look for age-related differences in task-specific FC in phonological and semantic picture-based match-mismatch tasks in the presence of distractor written words.

3.1 Introduction

Aging is accompanied by changes in cognitive processes, often viewed as cognitive decline [64–67], and many studies have shown differences in functional anatomy between younger and older adults (e.g., [67–69]). Some language-related cognitive processes, such as phonological processes, appear to be susceptible to age-related decline. For example, older adults display deficits in phonological access, leading to an increase in errors in naming objects [65, 66, 70, 71]. Age-related decline in phonological retrieval can also be observed in increased "tip of the tongue" (TOT) occurrences [65]. TOTs are characterized by a strong sense of knowing the target word, while temporarily being unable to produce it. During TOTs an individual has access to semantic and grammatical information about the target word, but is unable to retrieve some of the phonological information. Unlike phonological processes, other language-related cognitive processes, such as many semantic processes, do not appear

to decline with age. For example, semantic vocabulary is well preserved or even improved with age [65–67, 71], semantic retrieval processes are largely unaffected by aging [66, 71, 72], and several studies have even suggested that older adults may have richer semantic representations than younger adults [65, 70].

Despite the apparent difference in the way phonological and semantic processes are affected by aging, both engage additional cognitive processes, such as attention and working memory, which themselves exhibit age-related decline [66, 73, 74]. The inhibition deficit hypothesis [75] offers a possible mechanism for some age-related declines, suggesting that older adults experience greater difficulty in ignoring irrelevant information, compared to younger adults.

Diaz et al. [63] examined behavioral and fMRI correlates of age-related differences in processing irrelevant information during phonological, semantic and perceptual judgment tasks. They used picture-based match-mismatch tasks in the presence of task-irrelevant written words. The words were phonologically and semantically related to the pictures in the phonological and semantic trials, respectively. All participants were less efficient when making phonological as opposed to semantic judgment, and older adults responded more slowly than younger adults. Notably, no significant interactions between age-group and phonological/semantic task were found. The semantic task elicited activation in typical left-hemisphere language regions in all participants, regardless of age-group, with activation in these regions being positively correlated with efficiency. The phonological task elicited activation in bilateral precuneus and cingulate; regions implicated in task control [76], however there was only a trend toward a significant positive relationship with efficiency for younger adults. The apparent absence of a relationship between activation and performance under the phonological condition among older adults, at least on the surface, appears to be inconsistent with the inhibition deficit hypothesis. Moreover, older adults elicited greater activation than younger adults throughout the brain, including brain regions that were unrelated to behavioral performance, suggesting an age-related decline in neural efficiency, consistent with the transmission deficit hypothesis [77].

We note that fMRI studies report additional brain regions to be activated in older adults outside of typical brain networks found to be activated in younger

adults (e.g., [63,66,70,78–80]). Hence, the traditional model-driven approaches to FC analyses [23,81,82] which focus on a few "seed" brain regions that are selected based on the existing literature can lead to biased conclusions. Additionally, although many FC studies have used resting state fMRI (e.g., [83–86]), Bassett et al. [60] have shown that there can be significant differences between task-based and resting-state-based FC networks. Hence, there is a need for unbiased data-driven analyses of task-based whole-brain fMRI data, together with behavioral measures, to understand how age-related differences in FC relate to age-related differences in phonological and semantic processing.

Against this background, we investigated the interaction between phonological and semantic processes in the presence of task-irrelevant stimuli (phonologically and semantically related, respectively), in younger and older adults. Specifically, we focused on an analysis of age-group-related and age-group-independent differences in FC in younger and older adults engaged in phonological and semantic picture-based judgment tasks in the presence of task-irrelevant written words, to complement and extend previous analyses [63].

We hypothesized that while no age-group-dependent phonological/semantic task effects were found in the fMRI activation [63], there might be age-group-related differences in task-specific FC. In order to test our hypothesis, we generated task-specific whole-brain FC networks, based on the general linear model (GLM) analysis performed by Diaz et al. [63]. We characterized the resulting task-specific FC networks using network topological features. We used statistical ML algorithms to train predictive models to reliably discriminate between the phonological and semantic judgment tasks based on FC data. We used the resulting models to identify network topological features that can reliably discriminate between FC induced during the tasks in younger adults and older adults together, as well as in each age-group separately.

Our findings complement those of Diaz et al. [63] in several important aspects: We found that the left frontal pole (FP) shows FC patterns that appear to be characteristic of differentiating phonological and semantic processing, regardless of age-group. Our results show greater similarity in FC between phonological and semantic judgment tasks in older adults as compared to younger adults, which is consistent with the

dedifferentiation hypothesis [87]. We also found age-group-related differences in the FC of specific brain regions between phonological and semantic judgment tasks (bilateral frontal operculum cortex (FO) in the case of older adults). Our results suggest that the FC associated with phonological and semantic processing in older adults differs across brain regions that have been implicated in alertness, a prerequisite for selective attention to aspects of the tasks. Finally, our results demonstrate the potential of data-driven ML algorithms to contribute evidence to support, refute, or generate novel hypotheses regarding the differences in FC of different brain regions associated with age-group, task characteristics, or other aspects.

3.2 Materials and Methods

3.2.1 fMRI Data

Functional data of a fast event-related design from 19 younger adults ($19-35$ years of age, mean = 25.0; 10 male) and 19 older adults ($59-76$ years of age, mean = 67.3; 9 male) were collected during phonological and semantic match-mismatch tasks in the presence of task-irrelevant stimuli. Each participant provided informed consent and was paid for their participation. All experimental procedures were approved by the Institutional Review Board of the Duke University School of Medicine. Each trial consisted of a 1 s display of cue (phonological or semantic), followed by a 2 s display of two images and a centrally located task-irrelevant written word. The word was phonologically or semantically related to at least one of the images in the phonological and semantic trials, respectively. Pictures were high-resolution images of everyday objects against a white background. Each task had 40 match and 40 mismatch trials. The participants were asked to decide whether both images matched the cue (i.e., match trials) or not (i.e., mismatch trials). In the phonological task, the cue referred to the first letter of the names of the objects in the images (e.g., starts with P). In the semantic task, the cue referred to a perceptual or functional attribute of the objects in the images (e.g., edible). Overall, 80 match and 80 mismatch trials were presented to each participant in random order across eight runs. Inter-trial fixation intervals varied between 3 and 10.5 s, with mean of 4.8 s to allow for the convolution of the

hemodynamic response. Trial order across all trial types and inter-trial intervals were randomized and optimized using the Optseq2 program [88].

MRI scanning was conducted on a 3.0 Tesla GE MR 750 whole-body 60 cm bore human scanner equipped with 50 mT/m gradients and a 200 T/m/s slew rate. An eight-channel head coil was used for Radio Frequency reception (General Electric, Milwaukee, WI USA). High-resolution structural images were acquired using a 3D fSPGR pulse sequence (TR = 8.14 ms; TE = 3.22 ms; ti = 450 ms; FOV = 24 cm^2; flip angle = 12°; voxel size = $0.9375 \times 0.9375 \times 1$ mm; matrix = 256×256; 162 contiguous slices; averages = 1; Phase Encoding: RL; bandwidth: 62.5). Functional images sensitive to blood oxygen level-dependent (BOLD) contrast were acquired using an inverse spiral pulse sequence with SENSE acceleration (TR = 2.0s; TE = 30 ms; FOV = 25.6 cm^2; flip angle = 60°; SENSE factor = 2; voxel size = $3.75 \times 3.75 \times 4$ mm; matrix = 64×64; 38 contiguous oblique axial slices, parallel to the AC-PC line, interleaved acquisition; Phase Encoding: RL; bandwidth: 32). Four initial RF excitations were performed to achieve steady state equilibrium and were subsequently discarded.

3.2.1.1 Preprocessing

FSL [89] version 5.0.1 was used for brain extraction, slice-timing correction (to middle of TR period), motion-correction using FSL's MCFLIRT [90] (using six rigid-body transformations), high-pass filtering (cutoff = 50 s), spatial smoothing (using 8 mm FWHM Gaussian kernel), co-registration to structural images (in native space), and normalization to Montreal neurological institute (MNI) standard space (using FSL's MNI Avg152 T1 $2mm^3$ standard brain template).

3.2.1.2 GLM analysis

FEAT [91] version 6.00 was used with a double-γ hemodynamic response function (HRF). We focused on four contrasts in the GLM, corresponding to the four phonological/semantic × match/mismatch tasks relative to implicit baseline. Significant activation in voxels was assumed for $p < 0.01$, and clusters of voxels with significant activation were then corrected using Gaussian random fields theory. The thresholded

z-transformed output (z-map) of the GLM analysis of [*trial type* > *implicit baseline*] contrast (per run, per participant, per trial type) was later used in the current analysis.

3.2.1.3 Normalization

FSL does not store post-normalization functional data files for individual participants (i.e., first level analysis), but the parameters used for transformation are available (in ./reg/example_func2standard.mat in the respective FEAT folder). Therefore, using FSL's FLIRT and the parameters used during preprocessing for the respective run, functional data and the thresholded z-maps were normalized to standard space (FSL's MNI Avg152 T1 $2mm^3$ standard brain template) for our analysis.

3.2.2 Regions of Interest

In this paper we used functionally based regions of interest (ROIs), restricted by 111 anatomical cortical and subcortical ROIs in the Harvard-Oxford atlases (maximum probability through 50 percent). The cerebellum was excluded from this analysis.

For every phonological/semantic match/mismatch task, eight voxels per ROI were selected: one per run, such that the voxel selected per run showed the greatest activation response to the respective task in that run. Activation was determined by the contrast-specific normalized thresholded z-map for the run. In the case where no supra-threshold voxels were found for an ROI for a particular run, no voxel was selected per that run for that ROI. The time series of the eight (or less) voxels in each ROI were then averaged into a single time series for further analysis. In case no voxels were selected for an ROI in any run, the ROI's node did not have any edges in the resulting FC network.

3.2.3 FC Analyses

Task-specific FC was generated using a method closely related to the extension of Psychophysiological Interactions [92] to event-related designs, introduced by Gitelman et al. [93]. Overview of the procedure is shown in Fig. 3.1. Each time series (from the different ROIs) was first mean-corrected and de-convolved, using AFNI's 3dTFitter

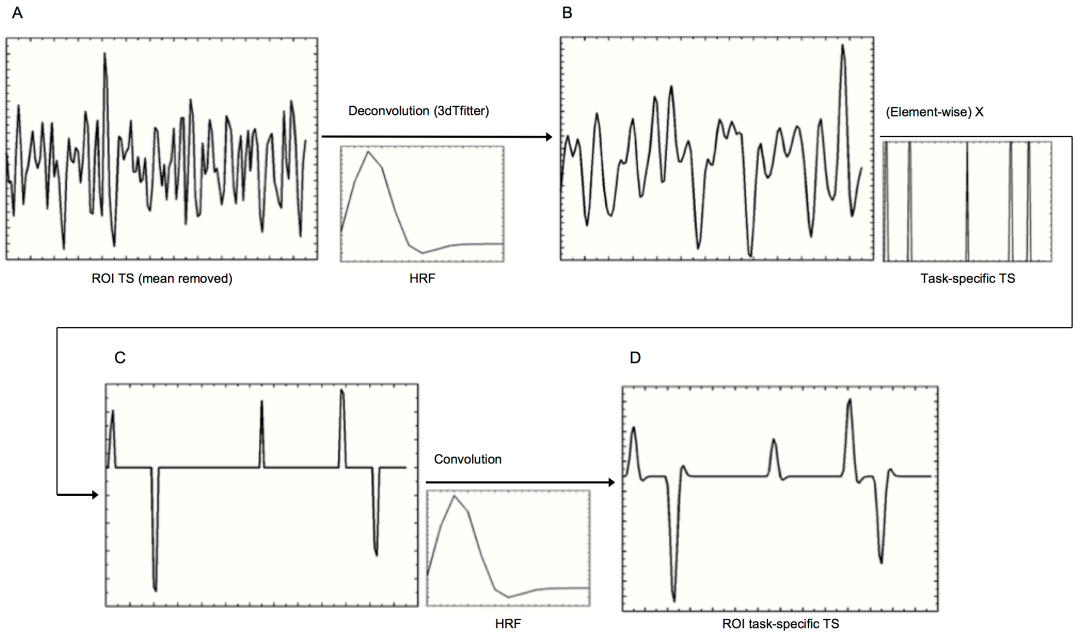

Figure 3.1. Flowchart showing the PPI-like processing steps of a run-specific match/mismatch task-specific time series of an ROI. For an example ROI: a) The time series representing the ROI after mean subtraction; b) Deconvolution with base HRF of FSL (sampled every 2 seconds, to match the recording TR), using 3dTfitter command of AFNI; c) Element-wise multiplication by match/mismatch task-specific time series (binary time series: 1 for three seconds of display starting at specific match/mismatch task onset for the run, 0 otherwise); d) Re-convolution with the same HRF that was used in step b, to transform the time series back to BOLD space. PPI = psychophysiological interactions; ROI = region of interest; TS = time series; HRF = hemodynamic response function (double γ); BOLD = blood oxygenated level-dependent.

command, with FSL's first basis double-γ HRF (provided in hrfbasisfns.txt of FSL's FLOBS). The HRF was sampled at a similar temporal rate to the functional recording's TR ($2s$) prior to deconvolution. Each time series was then element-wise multiplied by the task-specific time series, assuming the three seconds of display as the activation relevant to the task. Each time series was then re-convolved with the same HRF used in the deconvolution step, transforming the time series back to BOLD space. The procedure was performed separately for match and mismatch trials for each of the two tasks, resulting in four task-specific FCs per participant.

Task-specific FC networks were then computed using Pearson's correlation coeffi-

cient on the post-processing time series from the different ROIs. ROIs were set as nodes, and the weights of the edges between every two nodes were set to the correlation coefficient between the time series of the respective ROIs when $p < 0.05$, and zero (no edge) otherwise. Further details, including tests of stability and robustness of the resulting FC networks are provided in Appendix A.

3.2.4 Representation of FC Networks

From each task-specific FC network (four networks per participant), we extracted a weighted graph based on 20 different choices of thresholds on the edge weights to define positive and negative edges. We used lower thresholds for positive edges in the range from 0 to 0.9 in increments of 0.1 and upper thresholds for negative edges in the range from 0 to -0.9 in increments of -0.1. Thus, from each task-specific FC network, we had two types of networks: a 'positive network' in which the edge weights represented the magnitude of positive weights that exceeded the threshold; and a 'negative network' in which the edge weights represented the magnitude of weights that fell below the threshold. Note that the edges in the resulting networks were weighted edges, where the weights were within the range enforced by the chosen thresholds. The representation of a task-specific FC network using the 'positive' and 'negative' networks allowed us to extract features based on standard network topology measures (e.g., weighted degree (WD)) that were designed for networks with non-negative weighted edges.

3.2.5 Extraction of Network Topological Features

We extracted from both 'positive' and 'negative' networks topological features using six basic centrality metrics: WD, which is the sum of weights of all edges of a node; Average neighbor degree, which is the sum of WDs of first-level neighbors of a node, divided by the WD of the node; Eigen centrality, which measures the influence of a node in a network based on the Eigenvector, corresponding to the greatest Eigenvalue; Betweenness centrality measure, which estimates the importance of the node in a network by quantifying the number of times the node is crossed along the shortest path between two other nodes; Clustering coefficient (CC), which is the average

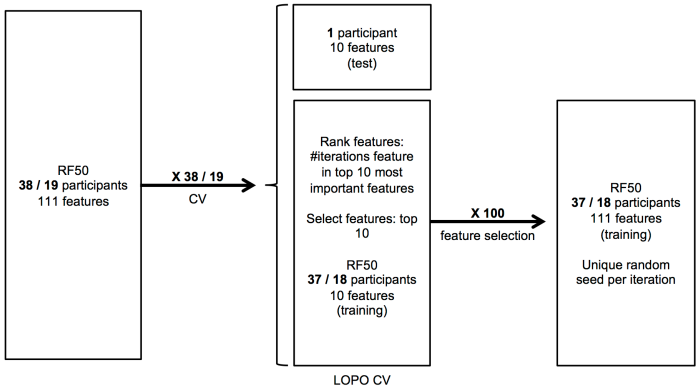

Figure 3.2. Scheme of feature selection procedure. The procedure consists of 38 and 19 iterations for the age-group-agnostic and age-group-specific datasets, respectively. In each iteration, one participant makes up the test set. The rest of the participants (37 and 18 participants in the age-group-agnostic and age-group-specific datasets, respectively) make up the training set and are used in the iteration's feature selection. All of the participants in the training set are used to train 100 different-seed RF50 models, and the 10 features selected in most of the models are selected for the respective training set. RF50 = random forest with 50 trees; LOPO = leave one participant out; CV = cross validation.

local clustering, which quantifies how close the neighbors of a node are to forming a clique; and Weighted CC, which is similar to clustering coefficient, except that it takes into account the weights on the edges (as opposed to only the presence or absence of edges). Given the 20 instantiations of 'positive' and 'negative' networks and six different topological features, we had 120 possible combinations to consider for training and testing classifiers for discriminating between the phonological and semantic tasks. Eigen centrality was computed using 'centrality.m' function, built into MATLAB_R2018b; the other five features were computed using Octave Networks Toolbox v2 [94].

3.2.6 Training Classifiers

We trained classifiers to discriminate between phonological and semantic judgment tasks (match/mismatch-agnostic) on the basis of the topological features of the corresponding FC networks, in younger and older adults separately and combined together. Because the number of samples (152), is not large enough relative to

the number of features (111), we needed to select a method that can effectively cope with the curse of dimensionality. We used random forest (RF) [95] classifiers, which have been shown to perform well in such settings by avoiding over-fitting [96]. Additionally, RF provides feature importance scores, which enhance the interpretability of the resulting classifiers. We used RF with 50 trees (RF50), as RF50 performance was not significantly different from RF with 500 trees, based on McNemar test of significance ($\chi^2 < 3.84$ for 117 of 120 feature-threshold combinations; $\chi^2 > 3.84$ for: WD for 'positive network' at threshold 0.2, betweenness centrality measure for 'positive network' at threshold 0.7, and CC for 'positive network' at threshold 0.1), as suggested by Latinne et al. [97]. The performance of RF50 on each classification task was evaluated using the leave-one-participant-out (LOPO) CV for all relevant participants (Fig. 3.2). We report the performance of the classifiers in terms of the average accuracy (percentage of correctly classified instances in the test set) across labels (phonological and semantic) per run. Additional details of the computational experiments are provided in Appendix A.

3.2.6.1 Identifying Discriminative Features

The performance of classifiers trained using ML can often be improved using an optimal subset of features as opposed to all features to train the classifiers. In addition to improving performance (by controlling over-fitting), feature selection often helps improve the comprehensibility of the resulting classifiers. We used the feature importance reported by RF to select features as follows: We ran RF50 on the training set with 100 different random seeds and selected the 10 features that were ranked among the top 10 in terms of RF feature importance scores in a majority of the LOPO iterations. We then focused our analyses on the 10 features that were selected using this procedure.

3.2.7 Statistical Analyses

We assessed the statistical significance of the observed differences in performance of classifiers, e.g., those trained on the data from the different age-groups, using two-tailed two-sample heteroscedastic T-test. We used the same test to assess the

statistical significance of the observed differences in the values of features across different groups of data samples (e.g., based on age-group or task). Additionally, we computed post-hoc statistical power (SP) of each T-test using the two per-group means and two per-group SDs). As SP = $1 - \beta$, we report β throughout the current study. Unless otherwise stated, we concluded that the observed differences were statistically significant at $p < 0.05$, if $\alpha = 0.05$ and $\beta \leq 0.20$. However, when even more stringent statistical significance could be established, e.g., $p < 0.01$ with $\alpha = 0.01$ and $\beta \leq 0.20$, we explicitly mention it in the text.

We used independence and conditional independence tests to examine the dependencies between features, tasks, and age-groups, e.g., whether feature dependence on task is conditioned on age-group. Specifically, we used the Hilbert-Schmidt Independence Criterion (HSIC) [98] and Self-Discrepancy Conditional Independence Test (SDCIT) [99] to assess the unconditional independence and conditional independence, respectively. We computed a mean p-value for these tests, based on 100 runs using different random seeds. Further details are provided in Appendix A.

3.3 Results and Discussion

3.3.1 Age-Group-Agnostic Network Topological Features That Discriminate between the Phonological and Semantic Tasks

We first examined the differences between the FC networks induced by the phonological and semantic tasks, regardless of age-group. We searched for a subset of features that reliably discriminated between the two tasks, using RF50 and feature selection (see Materials and methods), selecting the top 10 features per LOPO iteration. For each of the six topological features, we optimized the choice of threshold used to define the positive/negative FC networks, so as to maximize the accuracy (Table 3.1). Maximal accuracy (mean = 0.70, SD = 0.16) was achieved using RF50 with CC for 'positive network' at threshold 0.7. Seven features were selected in more than 28 (~ 75%) of the 38 LOPO iterations, four of which (CCs of the left FP, superior frontal gyrus, FO, and caudate nucleus) were selected in all 38 LOPO iterations.

We then used HSIC to examine the dependence of these seven features on phonolog-

Table 3.1. Phonological/semantic task classification performance of RF50 using different FC representations with feature selection

Feature	Optimal threshold	Accuracy
WD	0.7	0.697 ± 0.208
AND	0.5	0.684 ± 0.220
EC	0.1	0.678 ± 0.151
BCM	0.5	0.691 ± 0.145
CC*	0.7	0.704 ± 0.161
WCC	0.5	0.697 ± 0.183

Different representations of FC networks, referred to in this table, consist of combining one of six graph topological features with one of 20 thresholds. For readability purposes, results for optimal thresholds (thresholds for which highest accuracy is obtained with RF50 with features selection for the respective feature) only are reported. Accuracy was evaluated based on 38 LOPO iterations. 10 features (out of 111) were selected per LOPO iteration. RF50 = random forest with 50 trees; FC = functional connectivity; WD = weighted degree; AND = average neighbor degree; EC = Eigen Centrality; BCM = betweenness centrality measure; CC = clustering coefficient; WCC = weighted CC; ROI = region of interest; LOPO = leave one participant out. *Best performing network topological feature with RF50 using only the 10 selected features for the combination.

ical/semantic task for all participants (regardless of age-group). Our analysis showed two of the seven features (CCs of the left FP and Heschl's gyrus) to be highly correlated with task (mean $p < 1.00e{-}06$ and mean $p = 0.01$, respectively). Additionally, the CC of the left FP was significantly ($p = 3.90e{-}10$, $\alpha = 0.01$, $\beta = 1.74e{-}05$) greater for the semantic task (mean = 0.59, SD = 0.28) as compared to the phonological task (mean = 0.26, SD = 0.33).

3.3.2 Network Topological Features that Discriminate between the Phonological and Semantic Tasks in Older Adults and in Younger Adults

Next, we examined differences between the two tasks in each age-group separately. We compared the performance of the classifiers tested on data from participants from each of the age-groups separately. We found no significant difference in the

Table 3.2. Performance of RF50 with dataset-specific feature selection

Test set	Training set	Accuracy
Younger adults	Combined population	0.68 ± 0.16
	Younger Adults	0.65 ± 0.19
	Older Adults	0.68 ± 0.18
Older adults	Combined population	0.72 ± 0.16
	Younger Adults	0.61 ± 0.17
	Older Adults	0.66 ± 0.19
Combined population	Combined population	0.70 ± 0.16

Performance (accuracy) of RF50 with feature selection (10 features selected per LOPO iteration), using different combinations of training and test sets. Per test set, no significant difference was found between the different training sets, possibly due to low SP (α = 0.05, β > 0.2 for all). RF50 = random forest with 50 trees; LOPO = leave one participant out; SP = statistical power.

performance of the classifier tested on the data from younger adults (mean = 0.68, SD = 0.16; Table 3.2), as compared to that tested on older adults (mean = 0.72, SD = 0.16; Table 3.2).

CC of the left FP was the only feature selected in all LOPO iterations of each of the age-group-specific, as well as the age-group-agnostic, datasets. We used HSIC to test whether it was dependent on phonological/semantic task for each of the age-groups separately. As in the case of the age-group-agnostic dataset, in each of the age-group-specific datasets, the CC of the left FP was significantly (p = 1.28e−05, α = 0.01, β = 0.02 for younger adults; p = 8.11e−06, α = 0.01, β = 0.01 for older adults) greater for the semantic task (mean = 0.63, SD = 0.29 for younger adults; mean = 0.55, SD = 0.26 for older adults) as compared to the phonological task (mean = 0.28, SD = 0.35 for younger adults; mean = 0.23, SD = 0.32 for older adults). These results suggest that the left FP serves to discriminate between phonological and semantic tasks independently of the age-group.

To further examine age-group-specific features, we split the data (CC features at threshold 0.7) into two subsets based on the age-group. We performed feature selection, training, and testing of the classifiers separately for each age-group. We

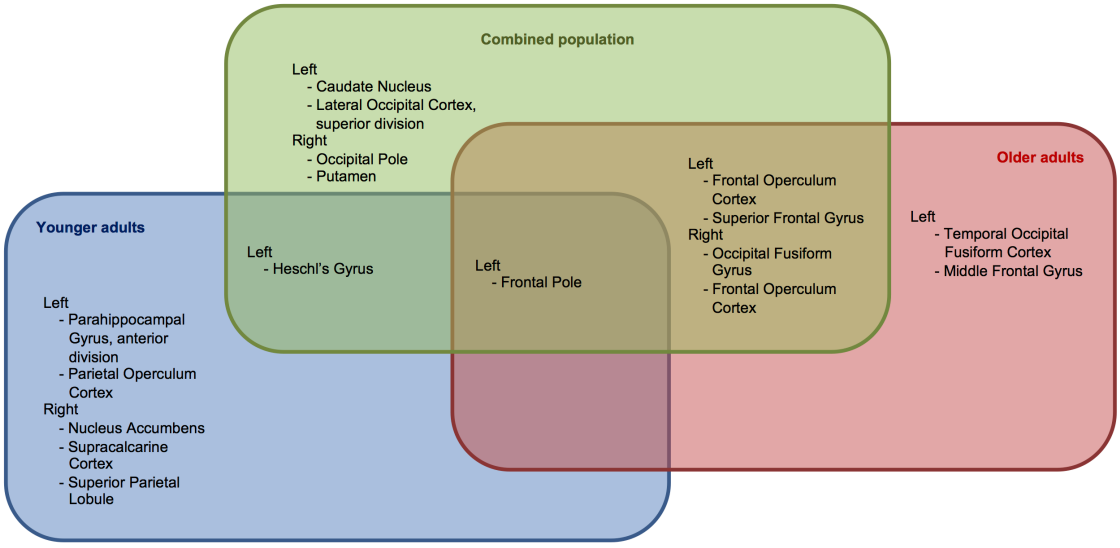

Figure 3.3. Euler diagram of selected features. The ROIs whose CCs were selected in more than 75% of the younger-adults-specific LOPO iterations (blue), more than 75% of the older-adults-specific LOPO iterations (red), and in more than 50% of the age-agnostic LOPO iterations (green), and the overlap between these subsets of features. ROI = region of interest; CC = clustering coefficient; LOPO = leave one participant out.

found no significant difference in performance of the resulting age-group-specific classifiers. For each of the age-groups, seven features were selected in more than 14 ($\sim$ 75%) of the 19 LOPO iterations, with only one feature being shared by the two age-group-specific classifiers (CC of the left FP; Fig. 3.3).

We then examined whether a classifier trained on data from one age-group is useful when testing on data from the other age-group. We trained a single classifier on all of the participants from one age-group and tested on each of the participants in the other age-group. We found no significant difference in performance between the classifiers that were trained on one age-group and tested on the other age-group as compared to the classifiers that were trained and tested on the same age-group or the classifiers that were trained on the data from both age-groups combined. We also found that the top nine of the 10 features selected when the classifier was trained on an entire age-group-specific dataset were identical to the top nine features selected when classifiers were trained and tested using LOPO CV. Specifically, for both age-groups, seven of the top nine features were selected in more than 75% of the respective LOPO

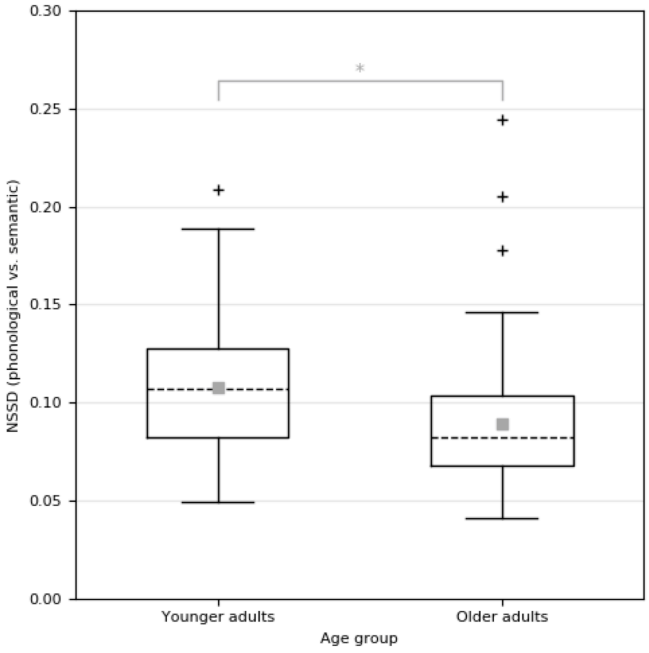

Figure 3.4. Within-participant NSSD between phonological/semantic match/mismatch tasks for the two age-groups. Within-participant NSSDs between either of the two (match/mismatch) FC networks from the phonological task and either of the two (match/mismatch) FC networks from the semantic task per participant. The plot is based on 19 participants per age-group and four comparisons per participant. *The means (dark grey squares) are significantly (p = 1.35e−03, α = 0.05, β = 0.10) different between the two age-groups (mean = 0.09, SD = 0.04 for older adults; mean = 0.11, SD = 0.03 for younger adults). NSSD = normalized sum squared difference; FC = functional connectivity; SD = standard deviation.

iterations. This suggests that the results of feature selection are fairly stable across LOPO iterations, and hence, reliable.

Finally, we computed within-participant normalized sum squared difference (NSSD) between two (match/mismatch) FC networks elicited by the semantic task and two (match/mismatch) FC networks elicited by the phonological task (four comparisons per participant). NSSDs for older adults (mean = 0.09, SD = 0.04) were significantly (p = 1.35e−03, α = 0.05, β = 0.10) smaller than NSSDs for younger adults (mean = 0.11, SD = 0.03; Fig. 3.4), demonstrating that FC networks for the phonological and semantic tasks from the same individual were more similar to each other in older adults than in younger adults. These results support the dedifferentiation

hypothesis [64, 87, 100], which suggests that the patterns of brain activation during similar tasks become less distinct with aging.

3.3.3 Age-Group-Independent Network Topological Features that Discriminate between the Phonological and Semantic Tasks

Our results show that the CC of the left FP was amongst the features selected for the phonological/semantic classifiers constructed from both age-group-specific datasets, as well as the age-group-agnostic dataset. Furthermore, in all three datasets, the CC of the left FP was greater for the semantic task as compared to the phonological task. This suggests that the FC of the left FP may play an important role in discriminating between the phonological and semantic tasks regardless of age-group.

To further explore this hypothesis, we performed feature selection on different variants of the age-group-agnostic dataset. After excluding all previously selected features except the CC of the left FP, we found that the newly selected features included the CC of the left FP along with nine new (i.e., previously not selected) features. There was no significant (p = 0.01, α = 0.05, β = 0.23) difference in performance of the phonological/semantic classifiers between using these 10 features (mean = 0.59, SD = 0.20) and using the original 10 selected features (mean = 0.70, SD = 0.16). In the second variant we excluded only the CC of the left FP from the set of candidate features. We found that the nine originally selected features were selected again, along with one new feature, and the performance (mean = 0.53, SD = 0.21) of the phonological/semantic classifiers was significantly (p = 9.34e−05, α = 0.01, β = 0.06) lower as compared to the performance using the original 10 selected features. These results suggest that the left FP is one of the most informative CC-based features for discriminating between the phonological/semantic tasks, regardless of age-group.

Next, we used HSIC to test whether the CC of the left FP was independent of age-group, as well as independent of the (phonological/semantic) task. While it was found to be independent of age-group (mean p = 0.28), it was highly correlated with the task in both age-groups, separately and combined (mean p < 1.00e−06 for age-group-agnostic dataset, mean p = 5.00e−05 for younger adults, and mean p = 2.00e−05 for older adults), suggesting that the task discrimination provided by this feature

is indeed age-group-independent. Additionally, SDCIT revealed that the CC of the left FP was highly correlated (mean p = 1.70e−03) with the phonological/semantic task, even when conditioned on the age-group, further supporting this hypothesis. This suggests that the CC of the left FP is a reliable discriminator of the phonological versus semantic tasks, and is age-group-independent.

Finally, as the number of immediate neighbors of a node may greatly affect its CC, we used SDCIT to examine whether the CC of the left FP was conditionally independent of the task, given the number of its immediate neighbors, which it was (mean p = 0.53). These results suggest that the number of immediate functional neighbors of the left FP, rather than the density of FC between them, differed between the phonological and semantic tasks. Specifically, the left FP displayed grater FC in the semantic task as compared to the phonological task.

3.3.4 Age-Group-Specific Network Topological Features that Discriminate between the Phonological and Semantic Tasks

Next, we explored if there were age-group-related differences among the features that discriminated between the phonological and semantic tasks. Specifically, we examined the intersection between the top features selected when classifiers were trained on the age-group-agnostic dataset as compared to their counterparts trained on the age-group-specific data. We focused on the features that were selected in more than 75% of the age-group-specific LOPO iterations and in more than 50% of the age-group-agnostic LOPO iterations. In addition to the CC of the left FP, nine features were selected in more than 50% of the age-group-agnostic LOPO iterations, among which one feature (the CC of the left Heschl's gyrus; Table 3.3) was selected in more than 75% of the younger-adults-specific LOPO iterations, and four features (CCs of the left superior frontal gyrus and FO, and the right occipital fusiform gyrus and FO; Table 3.3) were selected in more than 75% of the older-adults-specific LOPO iterations. The CC of the left Heschl's gyrus was not selected in any of the 19 LOPO iterations in older adults, suggesting that it is effective in discriminating between the phonological and semantic tasks only in the younger adults and not in the older adults. Similarly, the four features identified for older adults were each selected in no

more than one of the 19 LOPO iterations in younger adults, suggesting that they are not useful in discriminating between the phonological and semantic tasks in younger adults.

It is worth emphasizing that the selected features collectively, and not necessarily individually, contribute to the discrimination between phonological and semantic FC networks. Nevertheless, these results suggest that there are different age-group-specific subsets of ROIs that discriminate between FC induced by the phonological and semantic tasks.

Next, we used HSIC to examine whether any of the age-group-specific features were associated with the phonological/semantic tasks on their own. In younger adults, the CC of the left Heschl's gyrus was highly correlated with task (mean p = 0.01; Table 3.3), but it was not significantly different between the tasks. In older adults, three of the four features (CCs of the left FO, and the right occipital fusiform gyrus and FO) were highly dependent on task (mean p = 1.60e−03, mean p = 0.02, and mean p = 0.01, respectively; Table 3.3). Additionally, the CCs of the left and right FOs were significantly (p = 1.12e−03, α = 0.05, β = 0.08 for the left FO; p = 3.00e−03, α = 0.05, β = 0.13 for the right FO) greater for the phonological task (mean = 0.80, SD = 0.11 for both) as compared to the semantic task (mean = 0.70, SD = 0.13 for left FO; mean = 0.70, SD = 0.17 for right FO) in older adults.

Finally, SDCIT revealed that CCs of both the left and right FOs were not independent of the (phonological/semantic) task, even when conditioned on the number of their respective immediate functional neighbors (mean p = 0.02 for both). These results suggest that the FC between the immediate neighbors of these ROIs is denser for the phonological task than for the semantic task in older adults.

3.4 Further Discussion: Population-Condition Interactions

3.4.1 Age-Group-Independent Task-Related ROIs

It is worth noting that although our results suggest that the FC of the left FP is greater for the semantic task as compared to the phonological task, the left FP is not typically

considered a language-specific brain region, but is instead associated, more generally, with cognitive control. Badre and D'Esposito [101, 102] proposed a rostro-caudal gradient in the frontal cortex, in which more anterior regions are involved in more abstract, higher-order, levels of cognitive control. While the definition of "abstract" or "higher-order" cognitive control varies across tasks (e.g., referring to temporal abstraction in sequential tasks or relational complexity in association tasks), the FP appears to be consistently associated with higher-order control functions. Therefore, our results suggest that the semantic task demands greater levels of higher-order cognitive control as compared to the phonological task.

Interestingly, activation-based analysis of the current data [63] showed that greater activation in the left FP is elicited by the semantic task as compared to the phonological task (independently of age-group). Devlin et al. [103], in a different study, found greater activation in the left FP during semantic decision making as compared to phonological decision making. Additionally, FC-based analysis revealed that making more difficult semantic decisions is associated with increased connectivity between the posterior cingulate cortex and multiples ROIs, including the left FP [104]. These findings suggest that the semantic task in the current study was more difficult than the phonological task, and that the semantic task, in which participants made perceptual and functional judgments about object features (e.g., edible?, smooth?, flies?) may have required greater abstraction than the phonological task, in which participants decided if two objects started with the same letter (e.g., starts with P?, starts with D?). Behavioral analysis of the current data [63] showed that participants (both younger and older adults) were significantly more efficient in the semantic task than in the phonological task, suggesting difference in abstraction, rather than difficulty, between the tasks. Additionally, while the left FP has been associated with third-order relational complexity (i.e., evaluating the relationship between relationships), the left FC has been engaged during many tasks that do not appear to involve such complexity [102].

Finally, activation-based analysis of the current data [63] revealed significantly greater activation in the left FP during the semantic task as compared to the phonological task, independently of the age-group. The FC-based analysis in the current study suggests that the FC of the left FP provides much of the information needed to

discriminate between the phonological and semantic tasks, independently of the age-group. Together, these findings might explain in part why activation-based analysis of the data did not uncover interaction between the age-group and the task. This also points to the potential of our FC-based analysis to complement the activation-based analysis.

3.4.2 Task-Associated ROIs Specific to Older Adults

While the previous section speaks to the age-group-independent differences in FC between the phonological and semantic tasks, our analysis also revealed some age-group-specific differences in FC between the two tasks. Specifically, our analysis showed that task-dependent differences in FC of the left and right FOs are specific to older adults. Our results further suggest that the FC between the immediate functional neighbors of the bilateral FOs is greater in older adults in the phonological task as compared to the semantic task, a pattern that was not found in younger adults. The bilateral FOs are part of the cingulo-opercular network [105], which has been implicated in task-related attention and alertness (i.e., "non-selective attention"; e.g., [106–108]). Our analysis found significant differences in the density of FC in the immediate functional neighborhoods of the bilateral FOs (i.e., the immediate functional neighbors of each of the FOs are more functionally connected to each other during the phonological task than during the semantic task). These differences may be attributed to the immediate FC sub-networks of the FOs, rather than the FOs alone. Recall that in contrast to the CC of the left FP, excluding all other selected features (including the CCs of the left and right FOs) did not result in a significant drop in the accuracy of phonological/semantic task classification. Additionally, in light of the fact that feature selection aims to eliminate redundant features, it is possible that the task-discriminating information provided by the CCs of the bilateral FOs is redundant with respect to information provided by the CCs of other ROIs (e.g., the anterior cingulate cortex).

Sadaghiani and D'Esposito [107] found significant difference in FC within the cingulo-opercular network between high and low alertness, but not between high and low attention, suggesting that older adults in the current study required greater

alertness, rather than attention, during the phonological task than during the semantic task. In turn, our results are inconsistent with the inhibition deficit hypothesis [75], as they do not suggest inhibition deficit in older adults is the semantic task. Similarly, activation-based analysis of the current data [63] and its comparison to a previous study with similar tasks with no distractors [109] revealed inconsistencies with the inhibition deficit hypothesis. They found that older adults were not more impaired than younger adults by the distractor words. This suggests that differences between the two tasks, specific to older adults, may be explained by differences in activation of brain regions involved in alertness, a prerequisite for the cognitive ability to selectively attend to aspects of the tasks. Interestingly, Diaz et al. [63] reported that their results suggest that both tasks were more attentionally-demanding for older adults than for younger adults. These findings suggest a difference in age-group-related increase in attentional demand between the two tasks, with the increase being greater for the phonological task as compared to the semantic task. We further postulate that, under some conditions, such differences in alertness might contribute to differences in behavioral measures of performance that have been attributed to inhibition deficits (e.g., [75, 110–112]).

Finally, behavioral analysis of the current data [63] reveals that efficiency was significantly worse in the phonological trials as compared to the semantic trials, regardless of age-group. Therefore, it should be noted that it is possible that the difference in attention may be driving the differences between the two tasks, rather than differences between phonological and semantic cognitive processes per-se [80,113].

3.5 Limitations

There are some caveats associated with our interpretation of the results of our analyses. The first caveat has to do with the manner in which we performed voxel selection per ROI: (1) it is possible that the FC of the ROIs is negatively correlated with their size; (2) smaller ROIs might be represented by a single voxel. These factors can impact the interpretation of our results. For further explanation, see Appendix A. The second caveat has to do with the choice of thresholds used to define, and the features used to represent, the FC networks. While we have taken care to examine

the robustness of our findings, it is possible that using a different combination of thresholds and topological features might implicate brain regions other than the ones we have identified in our analyses.

3.6 Concluding Remarks

In this study we discriminated between task-specific FC networks induced by phonological and semantic judgment tasks in the presence of task-irrelevant written words. Specifically, we used topological features of these networks to train classifiers and used ML algorithms to identify a subset of features that can reliably discriminate between the two tasks. The results of our analyses demonstrate that some measures of FC (e.g., CCs of a collection of ROIs) are especially effective in discriminating between the two tasks. Our finding that FC networks for the phonological and semantic tasks from the same individual are more similar to each other in older adults, as compared to in younger adults, lends support to the dedifferentiation hypothesis. We further analyzed the FC network features that reliably discriminated between the phonological and semantic tasks in younger and older adults, separately as well as combined together. These analyses show that while the differences in FC of some of the brain regions (left FP) between the tasks is age-group-independent, those of other brain regions are largely age-group-dependent. Our results suggest that in older adults, differences in task-specific FC between the phonological and semantic tasks may be related to alertness. Specifically, the task-related differences in older adults may be explained in part by differences in density of immediate FC neighborhood of brain regions that have been implicated in alertness, a prerequisite for the cognitive ability to selectively attend to aspects of the tasks. Finally, our analyses also demonstrate the potential of data-driven (using ML algorithms) FC-based analyses to complement activation-based analyses, help identify novel features of FC networks to support or refute alternative hypotheses, or even suggest entirely new hypotheses (e.g., regarding how differences in FC of different brain regions between different cognitive tasks might be impacted by aging, task characteristics, or other factors).

Table 3.3. Age-group-specific selected features and the dependencies between them and the phonological/semantic tasks

	%Iterations			HSIC mean p-value		
ROIs	*YAs*	*OAs*	*All*	*YAs*	*OAs*	*All*
Left						
Frontal pole ***	100%	100%	100%	5.0e−05	2.0e−05	< 1.0e−06
Parahippocampal gyrus, anterior division	95%	-	-			
Parietal operculum cortex	89%	-	-			
Heschl's gyrus **	84%	-	71%	0.0141		0.0069
Frontal operculum cortex *	-	100%	100%		0.0016	0.0640
Superior frontal gyrus †	-	100%	100%		0.0106	0.0488
Temporal occipital fusiform cortex	-	100%	-			
Middle frontal gyrus	-	95%	-			
Right						
Nucleus accumbens	100%	-	-			
Supracalcarine cortex	95%	-	-			
Superior parietal lobule	89%	-	-			
Frontal operculum cortex *	-	100%	61%		0.0125	0.0737
Occipital fusiform gyrus *	-	100%	89%		0.0187	0.2237

Mean (based on 100 different seed runs) p-value of HSIC (rightmost column) between age-group-specific selected features and the phonological/semantic task in the respective age-group ('YAs'/'OAs') and in the combined dataset ('All'). The ROIs correspond to the CC features that were selected in more than 75% of the age-group-specific LOPO iterations and more than 50% of the age-agnostic LOPO iterations (middle column; Figure 3.3). * Significantly ($p < 0.05$) dependent on task in respective age-group only. ** Significantly ($p < 0.05$) dependent on task in respective age-group and combined dataset. *** Significantly ($p < 0.05$) dependent on task in both age-groups, separately and combined. † Significantly ($p < 0.05$) dependent on task in the combined dataset only. YA = younger adult; OA = older adult; HSIC = Hilbert-Schmidt Independence Criterion; CC = clustering coefficient; ROI = region of interest.

Chapter 4
Multi-Site Static FC

Reproducibility and generalizability can be viewed as forms of internal and external empirical validation, respectively. Reproducibility pertains to the ability to replicate results using identical or comparable samples, while generalizability pertains to the validity of results based on one sample, in the context of another sample (e.g., whether differences between TD and ASD populations, found using all-male samples, hold also for all-female or mixed-sex TD and ASD populations).

As fMRI studies typically include a modest number of observations, the reproducibility and generalizability of their results have been brought into question (e.g., [48, 49]). Increasing the sample size may not be a viable option in every study, due to financial constraints, availability of relevant participants, etc. Therefore, different repositories of neuroimaging datasets have been introduced in recent years (e.g., [30, 31, 34–39]; Table 4.1), creating opportunities to work with larger neuroimaging samples. However, the different datasets in these repositories are often collected at different sites, using different protocols, thereby introducing challenges. In this chapter, we address some of these opportunities and challenges with a mindful use of ML techniques, to present a cautionary tale. We focus on static FC from the ABIDE PCP repository [19, 35] as a case study and suggest a few pitfalls associated with multi-site data, we believe are worth taking into consideration in future work.

4.1 Introduction

fMRI provides a non-invasive, high spatial-resolution glimpse into the brain in-vivo.

Table 4.1. Neuroimaging repositories.

Repository	**S**	$\sum N_i$	**Data**	**Populations**
ABIDE I [34]	17	1112	rsfMRI	ASD, TD
ABIDE PCP [35]	17	882	rsFC	ASD, TD
ADHD-200 [36]	8	776	rsfMRI	ADHD, TD
CoRR [37]	33	5093	rsfMRI	TD
ACPI [38]	3	185	rsfMRI,rsFC	*
1000 FCP [39]	33	1380	rsfMRI	TD
Open fMRI [30]†	95	3372	task/rsfMRI	*

The number of samples, number of observations, and type of data in a few examples of neuroimaging repositories. S = Number of sites in repository; $\sum N_i$ = Total number of samples in repository; rs = Resting state; fMRI = Functional Magnetic Resonance Imaging; FC = Functional Connectivity; ASD = Autism Spectrum Disorder; TD = Typically Developing; ADHD = Attention Deficit Hyperactivity Disorder. * Populations vary between sites. † Now part of Open Neuro [31]

FC is the representation of fMRI data in the form of undirected graphs, in which nodes typically represent anatomical, functional, or hybrid brain regions, and edges describe a temporally-based relationship between these regions [9]. Many studies have shown relationships between differences in functional connectivity and other observed differences, such as behavior, age, and disease (e.g., [12–14]), providing insight into potential underlying mechanisms. However, sample sizes in neuroimaging studies may be limited due to different reasons, such as monetary cost and availability of participants. In recent years, the modest sample size in these studies has been heavily criticized (e.g., [48, 49]), calling into question the reproducibility and generalizability of these studies and their findings. As a result, multiple efforts have been made to encourage unrestricted public sharing of fMRI data (e.g., [50, 51]). Ensuing initiatives include fMRI repositories that have been created to share and aggregate multiple independent data samples from different sites [52] (e.g., [53, 54]). Thus, these repositories could potentially augment sample size, but they also introduce new confounding variables to be factored in when combining their data samples. Samples from different sites may vary in equipment, acquisition protocol, instructions to participants (e.g., eyes open or closed), demographics (e.g., age, severity of studied

condition, and medication usage), and other variables [55].

The ABIDE PCP repository [19] includes samples from 17 sites. Each sample contains preprocessed structural and functional neuroimaging data and meta-data from individuals diagnosed with Autism Spectrum Disorder (ASD) and typically developing (TD) individuals. ASD is an umbrella term for a range of neurological disorders [114], making it heterogeneous. Although once considered rare, ASD is now estimated to affect one in 160 children worldwide [115] and one in 68 in the US [116]. ASD is diagnosed in early childhood and tends to persist throughout the individual's lifetime and it is roughly 4.5 times more prevalent in males than in females [114]. In part due to limited understanding of ASD, many misconceptions surround it, including the persisting public debate over causal relationship between vaccines and ASD, despite lack of scientific support (e.g., [117–119]).

While not the sole purpose of fMRI repositories, sample size augmentation was definitely a key incentive [52] to create and use them, as also evident by many studies that use the ABIDE PCP repository (e.g., [40, 120–123]). For 52 of the publications mentioned on the ABIDE PCP's website [35], we were able to determine with certainty the number of ABIDE sites used in their analysis. Out of those 52 publications, roughly 21% (11 publications) used a single site, and only roughly 56% (29 publications) used $15 - 17$ (out of 17) sites (with some using only a sub-sample based on demographics, rather than site). It should be noted that the augmentation provided by these multi-site samples was intended to result in a heterogeneous sample that represents and captures the real-world variance of the disease or relevant population [124]. This heterogeneity of aggregated multi-site samples is often treated with caution by researchers, thereby driving them to use only one site or a few sites, rather than the entire repository [40] (e.g., [125–128] and the above statistics for publications using data from ABIDE PCP).

Many of the studies that have not shied away from the heterogeneity of multi-site samples have reported lower classification accuracy as compared to single-site studies (e.g., [129, 130]). One study [40] compared multi-site classification accuracy between intra-site (both training and test sets include observations from all sites) and inter-site (entire single site as test set and the rest of the sites used for training) CV setups, and found that the variability across folds was greater for inter-site as

compared to intra-site setup, while means were comparable. Another study [131] did not report SD, but their reported mean accuracy was higher for intra-site setup as compared to inter-site setup. Whether leaving an entire site out or training on partial data from all sites, these studies have averaged their results across all sites. It is also generally assumed that including more sites in the training sample increases the relevant variability of the data and thereby the generalizability of the results. However, we postulate that there could be important differences between sites, which in turn could shed light on the generalizability of aggregated multi-site data.

In the present study, we use the ABIDE PCP repository as a case study to examine four main questions:

1. Do models trained on a single site generalize to other sites?
2. Do models trained on multiple sites generalize to other sites?
3. Can performance on a single site be improved by leveraging data from other sites?
4. Can performance on a single site be improved by leveraging selective subsets of data from other sites?

Table 4.2. Demographics and participants' IQ scores in select sites

	N		$\hat{\mu}_{\textbf{Age}}\,(\hat{\sigma})$		**Male:Female**			$\hat{\mu}_{\textbf{FIQ}}\,(\hat{\sigma})$		$\hat{\mu}_{\textbf{VIQ}}\,(\hat{\sigma})$		$\hat{\mu}_{\textbf{PIQ}}\,(\hat{\sigma})$	
Sites	TD	ASD	TD	ASD	TD	ASD	**IQ Test**	TD	ASD	TD	ASD	TD	ASD
Leuven	34	27	18.21 (4.98)	17.98 (4.98)	25 : 9	25 : 2	WAIS III	115 (12)	109 (13)	116 (11)	101 (18)	108 (13)	105 (17)
NYU	98	73	15.67 (6.19)	14.92 (7.04)	72 : 26	64 : 9	WASI	113 (13)	106 (16)	113 (12)	104 (15)	110 (14)	108 (17)
Pitt	23	22	19.13 (6.19)	19.35 (7.35)	20 : 3	18 : 4	WASI	111 (9)	112 (13)	108 (11)	108 (13)	110 (8)	113 (13)
Trinity	23	21	17.48 (3.58)	17.01 (3.04)	23 : 0	21 : 0	Mix	111 (12)	109 (15)	109 (13)	107 (14)	110 (11)	107 (16)
UCLA	39	36	13.18 (1.76)	13.34 (2.56)	33 : 6	34 : 2	Mix	106 (10)	102 (13)	107 (11)	104 (13)	104 (11)	101 (13)
UM	65	47	15.03 (3.64)	13.86 (2.31)	49 : 16	38 : 9	Mix	109 (9)	107 (17)	114 (13)	110 (20)	104 (11)	104 (20)
USM	23	37	22.33 (7.70)	24.82 (8.46)	23 : 0	37 : 0	Mix	116 (15)	100 (17)	114 (16)	96 (20)	113 (13)	105 (17)
Yale	26	22	12.76 (2.78)	13.01 (3.03)	19 : 7	15 : 7	DAS II	105 (18)	94 (23)	107 (16)	96 (25)	101 (17)	90 (19)

The demographics of (TD/ASD) group participants and the IQ test administered in each of the eight biggest sites: sample size (N), estimated mean ($\hat{\mu}$) age and SD ($\hat{\sigma}$), male-female ratio, and mean FIQ and SD, mean VIQ and SD, and mean PIQ and SD of participants for which scores are available. IQ = intelligence quotient; TD = typically developing; ASD = autism spectrum disorder; SD = standard deviation; FIQ = full IQ; VIQ = verbal IQ; PIQ = performance IQ.

Table 4.3. Acquisition variables in select sites

		$\hat{\mu}_{\mathbf{MeanFD}}\,(\hat{\sigma})$			**ACQ**	**Voxel Size**	**TR**	**TA**
Sites	**Eyes**	TD	ASD	**Scanner**	**Matrix**	**[mm^3]**	**[mm]**	**[m:s]**
Leuven	Mix	0.09 (0.03)	0.09 (0.04)	PHILIPS INTERA 3T	64×64	$3.59 \times 3.59 \times 4.0$	1667	07 : 06.7
NYU	Mix	0.05 (0.03)	0.08 (0.04)	SIEMENS MAGNETOM Allegra syngo MR 2004A	80×80	$3.0 \times 3.0 \times 4.0$	2000	06 : 00.0
Pitt	Closed	0.12 (0.03)	0.10 (0.04)	SIEMENS MAGNETOM Allegra syngo MR A30	64×64	$3.1 \times 3.1 \times 4.0$	1500	05 : 06.0
Trinity	Closed	0.07 (0.02)	0.10 (0.04)	PHILIPS 3T Achieva	80×80	$3.0 \times 3.0 \times 3.5$	2000	05 : 06.0
UCLA	Open	0.07 (0.04)	0.09 (0.04)	SIEMENS MAGNETOM TrioTim syngo MR B15	64×64	$3.0 \times 3.0 \times 4.0$	3000	06 : 06.0
UM	Open	0.06 (0.03)	0.09 (0.05)	GE Signa 3T	64×64	$3.438 \times 3.438 \times 3.0$	2000	10 : 00.0
USM	Open	0.10 (0.04)	0.10 (0.05)	SIEMENS MAGNETOM TrioTim syngo MR B17	64×64	$3.4 \times 3.4 \times 3.0$	2000	08 : 06.0
Yale	Open	0.08 (0.04)	0.09 (0.04)	SIEMENS MAGNETOM TrioTim syngo MR B17	64×64	$3.4 \times 3.4 \times 3.0$	2000	06 : 40.0

Acquisition variable in each of the eight biggest sites: eyes (closed or open) during scan, estimated mean ($\hat{\mu}$) head movement (mean FD) and SD ($\hat{\sigma}$) per (TD/ASD) group, scanner, ACQ matrix, voxel size, repetition time, and duration of acquisition. FD = framewise displacement; TD = typically developing; ASD = autism spectrum disorder; ACQ = acquisition; TR = repetition time; TA = acquisition time; SD = standard deviation.

It should be noted that the ABIDE PCP repository includes some variability control. The repository offers preprocessed data (users can choose from four pipelines × with/without band-pass filtering × with/without global signal regression) to provide consistent preprocesing across all sites. Additionally, extracted mean time-series per ROI are available for seven commonly-used structural and functional atlases: Automated Anatomical Labeling (AAL) [132], Eickhoff-Zilles (EZ) [133], Harvard-Oxford (HO) [134] at 25% probability threshold, Talaraich and Tournoux (TT), Dosenbach 160 [12], Craddock 200 [16], and Craddock 400. Nonetheless, samples from different sites still differ in many aspects (e.g., sample size, equipment, acquisition protocol, eyes state (open/closed) during scan, age and sex distributions of participants, and IQ of participants; Tables 4.2 and 4.3). To control for these variables, some researchers have used subsets of the data, focusing on particular demographics (e.g., [41, 135]) and/or sites (e.g., [120, 126, 127, 136]), and some have tried to integrate them as confounding variables into their models (e.g., [40, 123, 129]).

4.2 Materials and Methods

4.2.1 Experimental Design

In order to address our four research questions, we tested a series of models on eight select sites (see section 4.2.2):

1. Intra-single-site models (section 4.3.1)
2. Inter-single-site model (section 4.3.2):
   (a) Without additional training on observations from the target site
   (b) Training additionally on 90% of the target site (the observations that were not in the test set) per fold
3. Inter-multi-site models (section 4.3.3):
   (a) Without additional training on observations from the target site

(b) Training additionally on 90% of the target site (the observations that were not in the test set) per fold

(c) Training on 90% of the target site (the observations that were not in the test set) per fold in addition to a subset of the data from other sites, selected based on that training set (from the target site)

4.2.2 Participants

In the present study we focused on sites from the ABIDE PCP repository that had at least 20 participants per (TD/ASD) group. Eight sites met this requirement: University of Leuven (Leuven; 1 & 2), New York University Langone, Medical Center (NYU), University of Pittsburgh, School of Medicine (Pitt), Trinity Centre for Health Sciences (Trinity), University of California, Los Angeles (UCLA; 1 & 2), University of Michigan (UM; 1 & 2), University of Utah, School of Medicine (USM), and Yale Child Study Center (Yale). Combined, these sites included 616 participants (331 TD, 285 ASD; Table 4.2). The ABIDE PCP repository includes only data that passed quality control based on visual inspection by three human experts. In accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines, all datasets were anonymized, and no protected health information was included in the datasets.

4.2.3 Preprocessing

We used data preprocessed with version X of the Configurable Pipeline for the Analysis of Connectomes (C-PAC) [9] pipeline with the 'filt_global' strategy (both band-pass filtering and global signal regression were applied to the data), readily available from ABIDE PCP. The C-PAC pipeline included slice time correction, motion correction to the average image, skull-stripping, and global mean intensity normalization to $10,000$. Nuisance signal removal included head motion regression using 24 parameters [137], scanner low-frequency drifts regression using linear and quadratic trends, physiological (tissue signal) noise regression using top five principal components from the signal in the white-matter and cerebro-spinal fluid derived from the prior tissue segmentations transformed from anatomical to functional space (CompCor) [138], and global signal

regression. Next, band-pass filtering ($0.01 - 0.10$Hz) was applied to the data. Finally, functional images were coregistered on the anatomical images with boundary-based rigid body method, and the results were then normalized to standard space (MNI152) with non-linear registration from ANTs [139].

4.2.4 Data Representation

We used the mean time-series for ROIs in the Craddock 200 functional atlas [16], available from ABIDE PCP. FC networks were constructed using NILearn's [140] connectome function with Pearson's correlation. All edges in the FC network ($\frac{200\times199}{2} = 19,900$) were used as features.

4.2.5 Machine Learning Classifiers

Since our data representation suffered from the curse of dimensionality [141], having $19,900$ features with only 616 samples overall and $44-171$ per site, we used Naïve Bayes (NB), a simple model that can handle the data with a small number of parameters to adjust. As the features we used were continuous, Gaussian distributions are feasible to assume, commonly used, and allow for simplicity. Thus, we used Gaussian NB (GNB) [142], as implemented in scikit-learn [143] (version 0.20.2).

4.2.5.1 GNB

GNB is a simplified probabilistic model, using Bayes theorem, and assuming all features are mutually independent, conditioned on the class. For binary classification with classes $C = \{c_1, c_2\}$, observation $x = \{x_1, x_2, \ldots, x_n\}$ is classified as c_1 if $0 < \frac{\log p(c_1|x)}{\log p(c_2|x)}$, and:

$$p(c_k \mid x) \propto p(c_k) \prod_{i=1}^{n} p(x_i \mid c_k) = p(c_k) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}_{k,i}^2}} \cdot e^{-\frac{(x_i-\hat{\mu}_{k,i})^2}{2\hat{\sigma}_{k,i}^2}}$$

where $\forall c_k \in C$, $\hat{\mu}_{k,i}$ and $\hat{\sigma}_{k,i}$ are the estimated mean and SD, respectively, of the i^{th} feature for class c_k.

4.2.5.2 Performance Evaluation

We evaluated the performance of models using stratified 10-fold CV on the target site. To guarantee results were comparable between experiments, the same 10 stratified folds per target site were used whether the model was trained on data from the target site or not. As the sites varied in size and folds may have contained as little as four observations (two per class), we corrected the confusion matrix per fold by applying Laplace "add-one" smoothing to it. Performance was measured using the F_1-score of the target ("positive") class (ASD) based on the Laplace-corrected confusion matrices. F_1-score of a class is the harmonic mean of precision and recall of the class:

$$F_1\left(ASD\right) = 2 \cdot \frac{Precision\left(ASD\right) \cdot Recall\left(ASD\right)}{Precision\left(ASD\right) + Recall\left(ASD\right)} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where TP is the number of "true positive" observations (ASD participants correctly classified as ASD), FP is the number of "false positive" observations (TD participants incorrectly classified as ASD), and FN is the number of "false negative" observations (ASD participants incorrectly classified as TD).

4.2.6 Distance between Distributions

To quantify the difference between TD and ASD samples, between and within sites, we examined the distance between their distributions. We used Maximum Mean Discrepancy (MMD) [144] to estimate distances between distributions of different samples:

$$\widehat{MMD}_u^2\left[\mathcal{F}, X, Y\right] = \frac{1}{m\left(m-1\right)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k\left(x_i, x_j\right) + \frac{1}{n\left(n-1\right)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k\left(y_i, y_j\right) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=i}^{n} k\left(x_i, y_j\right)$$

where X and Y are samples with m and n observations, respectively, and $\mathcal{F}$ is a unit ball in a universal reproducing kernel Hilbert space $\mathcal{H}$, defined on the compact metric space $\{X \cup Y\}$, with associated continuous kernel $k\left(\cdot, \cdot\right)$.

4.2.6.1 Kernel Two-Sample Test

A Kernel Two-Sample Test (KTST) is a kernel-embedding-based nonparametric statistical test to determine whether two samples are drawn from different distributions (H_0: the samples are drawn from the same distribution). The KTST used in the present work [144] uses MMD to estimate the difference between distributions.

In the present study, we used the KTST to test whether (TD/ASD) group-specific samples, within or between single or multiple sites, were drawn from different distributions.

4.2.6.2 The Kernel Function

We used a parameter-free positive-definite kernel, based on the Frobenius inner product, which is defined on adjacency matrices (natural representation of FC networks):

Let A and B be two normalized adjacency (square) matrices of FC networks with n nodes each. Then:

$$k(A, B) = \frac{f(A, B)}{\sqrt{f(A, A) \cdot f(B, B)}}$$

where $f(\cdot, \cdot)$ is the Frobenius inner product:

$$f(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} B_{i,j} \quad \forall A, B \in \mathbb{R}^{n \times n}$$

Being parameter-free, it allowed us to avoid possible bias introduced by parameter selection. Being positive-definite, it satisfies properties that are required for reproducing kernel Hilbert space, which is an integral part of MMD (see section 4.2.6). Finally, we confirmed the fitness of this kernel function to our data (see supplementary material section B.1.4).

4.2.7 Selecting Observations Based on Target Site's Training Set

For multi-site models in experiment 3.c (see section 4.2.1), we selected a subset of the data from non-target sites based on the fold's training data from the target site. Per

fold, let $X = \{X_{TD} \cup X_{ASD}\}$ be the target site's training data (the 90% that were not included in the test set), where X_{TD} and X_{ASD} are the TD and ASD observations, respectively. Let $Y = \{Y_{TD} \cup Y_{ASD}\}$ be the combined sample from all non-target sites, where Y_{TD} and Y_{ASD} are the TD and ASD observations, respectively. $y \in Y$ is selected based on:

$$\forall y \in Y_{TD}, \text{ if } \widehat{MMD}_u^2\left[\mathcal{F}, \left\{\tilde{X}_{TD} \cup y\right\}, \tilde{X}_{ASD}\right] > \widehat{MMD}_u^2\left[\mathcal{F}, \tilde{X}_{TD}, \tilde{X}_{ASD}\right]$$

$$\forall y \in Y_{ASD}, \text{ if } \widehat{MMD}_u^2\left[\mathcal{F}, \tilde{X}_{TD}, \left\{\tilde{X}_{ASD} \cup y\right\}\right] > \widehat{MMD}_u^2\left[\mathcal{F}, \tilde{X}_{TD}, \tilde{X}_{ASD}\right]$$

where $\tilde{X}_{TD}$ and $\tilde{X}_{ASD}$ are initialized to X_{TD} and X_{ASD}, respectively, and for every $y \in Y_{TD}$ that is selected, $\tilde{X}_{TD}$ is updated to $\left\{\tilde{X}_{TD} \cup y\right\}$ and for every $y \in Y_{ASD}$ that is selected, $\tilde{X}_{ASD}$ is updates to $\left\{\tilde{X}_{ASD} \cup y\right\}$.

4.3 Results

The results of our experiments (section 4.2.1) varied between sites and are summarized in Table 4.4. To gain additional insight into these results and, consequently, the combined usage of sites in the ABIDE PCP repository, we examined the distance between distributions of TD and ASD observations within and between the eight select sites (section 4.3.4).

4.3.1 Intra-Single-Site GNB Performance

First, we estimated the performance of GNB models for each of the eight select sites on their own (Table 4.4.A). For each site we ran stratified 10-fold CV, training a GNB model on 90% of the data in the site and testing it on the remaining 10%, while preserving in each fold the ratio between the classes in the respective site.

4.3.1.1 Performance Bounds

The sample size varied between sites, leading to the Laplace correction affecting the results differently per site. Thus, we computed the performance bounds per site using the stratified 10-fold CV. The lower bound was the Laplace-corrected performance of

classifying all observations as ASD, which was better than random guessing for all sites (Table B.4). The upper bound was the Laplace-corrected performance of perfect classification.

The intra-single-site performance on all select sites but Pitt fell below the lower bound (Table 4.4), despite Pitt having the second-smallest sample size among the eight select sites (Table 4.2). This suggests that there was a good separation between observations from the two (TD/ASD) groups in this site.

4.3.2 Inter-Single-Site GNB Performance

Next, we trained models on a single site in its entirety and tested on a single target site in 10 stratified folds, testing the model on 10% of the target site at a time (Table 4.4.B). We then tested similar models, trained additionally on the rest (90%) of the target site per fold (Table 4.4.C). The behavior was not uniform across all select sites. For example, the performance on Pitt did not improve for models trained on any of the other sites (with or without additional training data from Pitt). Conversely, the performance on UM improved for models trained on six of the seven other sites, and further improved for three of them when trained additionally on data from UM. Additionally, many single-site models improved performance on certain target sites, but dropped in performance when trained in addition on data from the target site (e.g., training sites NYU, Pitt, UM, and USM with target site Trinity, or training sites Leuven, Pitt, and USM with target site Yale; Table 4.4.B-C). In three cases (training site USM with target site UCLA and training sites Trinity and UCLA with target site USM), performance on the target site improved for single-site models, but only when not trained also on data from the target site (Table 4.4.B-C). These varied behaviors beg the question of how reliable these models and their results were. Therefore, we tested for overfitting in the different models (see section 4.3.5).

4.3.3 Inter-Multi-Site GNB Performance

Finally, we trained models on seven sites (the eight select sites, excluding the target site) and tested on a single target site in 10 stratified folds, testing the model on 10% of the target site at a time (Table 4.4.D). Similarly to our inter-single-site experiments,

we tested similar multi-site models, trained additionally on the rest (90%) of the target site per fold (Table 4.4.E). Again, the behavior varied between sites, with three distinct types in reference to intra-single-site performance:

1. Performance dropped for multi-site models, with and without additional training data from the target site (Leuven and Pitt).

2. Performance improved for multi-site models, but less so when trained additionally on data from the target site (NYU).

3. Performance improved for multi-site models and was maintained or further improved when multi-site models where additionally trained on data from the target site (Trinity, UCLA, UM, USM, and Yale).

Overall, only three of the eight select sites (UM, USM, and Yale) achieved best performance with multi-site models trained additionally on data from the target site.

Table 4.4. Performance on the eight select sites for the different models

		Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
A	**Intra-Single-Site (CV)**	0.55 (0.14)	0.57 (0.09)	0.67 (0.09)	0.52 (0.12)	0.55 (0.16)	0.53 (0.12)	0.68 (0.09)	0.49 (0.15)
B	**Leuven**		0.55 (0.09)	0.53 (0.14)	0.55 (0.11)	0.58 (0.09)	0.59 (0.09)	0.60 (0.08)	0.57 (0.10)
	NYU	0.36 (0.12)		0.48 (0.15)	0.56 (0.14)	0.60 (0.13)	0.62 (0.07)	0.69 (0.11)	0.55 (0.15)
	Pitt	0.59 (0.07)	0.58 (0.05)		0.60 (0.09)	0.59 (0.11)	0.58 (0.11)	0.64 (0.14)	0.59 (0.07)
	Trinity	0.41 (0.13)	0.60 (0.05)	0.52 (0.15)		0.65 (0.05)	0.58 (0.05)	0.68 (0.15)	0.48 (0.18)
	UCLA	0.40 (0.13)	0.61 (0.08)	0.46 (0.12)	0.52 (0.17)		0.61 (0.08)	0.68 (0.09)	0.47 (0.16)
	UM	0.36 (0.16)	0.54 (0.13)	0.57 (0.12)	0.64 (0.10)	0.45 (0.16)		0.59 (0.09)	0.48 (0.15)
	USM	0.47 (0.11)	0.61 (0.03)	0.60 (0.07)	0.60 (0.07)	0.67 (0.04)	0.61 (0.03)		0.57 (0.14)
	Yale	0.33 (0.12)	0.47 (0.10)	0.44 (0.12)	0.44 (0.16)	0.50 (0.16)	0.47 (0.10)	0.66 (0.10)	
C	**Leuven + Target**		0.62 (0.04)	0.62 (0.09)	0.55 (0.08)	0.65 (0.07)	0.60 (0.10)	0.66 (0.07)	0.56 (0.13)
	NYU + Target	0.47 (0.17)		0.57 (0.12)	0.55 (0.14)	0.58 (0.15)	0.64 (0.10)	0.70 (0.12)	0.59 (0.16)
	Pitt + Target	0.59 (0.08)	0.60 (0.09)		0.59 (0.07)	0.63 (0.11)	0.53 (0.16)	0.66 (0.08)	0.57 (0.12)
	Trinity + Target	0.55 (0.07)	0.57 (0.07)	0.62 (0.13)		0.60 (0.16)	0.56 (0.14)	0.66 (0.10)	0.51 (0.09)
	UCLA + Target	0.49 (0.14)	0.60 (0.08)	0.57 (0.12)	0.50 (0.15)		0.61 (0.13)	0.67 (0.10)	0.55 (0.15)
	UM + Target	0.50 (0.14)	0.54 (0.12)	0.67 (0.10)	0.57 (0.06)	0.54 (0.12)		0.73 (0.13)	0.58 (0.10)
	USM + Target	0.47 (0.15)	0.59 (0.12)	0.58 (0.12)	0.53 (0.07)	0.52 (0.16)	0.59 (0.12)		0.50 (0.15)
	Yale + Target	0.56 (0.07)	0.57 (0.09)	0.64 (0.14)	0.53 (0.10)	0.63 (0.10)	0.57 (0.09)	0.71 (0.08)	
D	**Multi-Site**	0.40 (0.14)	0.64 (0.07)	0.51 (0.13)	0.56 (0.15)	0.64 (0.08)	0.63 (0.06)	0.73 (0.07)	0.61 (0.14)
E	**Multi-Site + Target**	0.39 (0.13)	0.60 (0.09)	0.53 (0.12)	0.56 (0.15)	0.65 (0.07)	0.64 (0.06)	0.73 (0.09)	0.61 (0.14)
	Lower Bound	0.58 (0.03)	0.58 (0.10)	0.60 (0.12)	0.59 (0.02)	0.61 (0.02)	0.57 (0.02)	0.69 (0.02)	0.58 (0.03)
	Upper Bound	0.79 (0.02)	0.89 (0.01)	0.76 (0.02)	0.76 (0.02)	0.82 (0.02)	0.85 (0.01)	0.82 (0.02)	0.76 (0.02)

Mean F_1-score(ASD) (SD) for stratified 10-fold CV on the eight select sites (columns) for different model training (rows). A: Intra-single-site. B: Inter-single-site. C: Single site + 90% of the target site training. D: Multi (seven)-site training. E: Multi (seven)-site + 90% of the target site training. CV = cross validation; ASD = Autism Spectrum Disorder; SD = standard deviation.

To examine whether training on more sites yielded better results, we also repeated the inter-multi-site experiments (with and without additional training on the target site) using all 16 other sites per target site (for number of participants in each of the additional sites, see table B.3). Performance did not significantly (two-tailed paired sample t-test $p \geq 0.07$ for all) improve for any of the sites in either model. Actually, performance of models without target-site training slightly dropped using 16 sites, as compared to using seven sites, for NYU, UCLA, USM, and Yale. Performance of models with additional training on target site slightly dropped only for UCLA (Table 4.5). Overall, these results suggest that training on nine additional sites did not improve performance on any of the select sites.

Table 4.5. Performance on the eight select sites for inter-multi-site models using all other (16) sites

Site	no TST	TST
Leuven	0.40 (0.14)	0.41 (0.13)
NYU	0.63 (0.09)	0.61 (0.07)
Pitt	0.57 (0.12)	0.60 (0.09)
Trinity	0.58 (0.15)	0.57 (0.15)
UCLA	0.61 (0.08)	0.64 (0.09)
UM	0.64 (0.06)	0.66 (0.05)
USM	0.66 (0.13)	0.73 (0.11)
Yale	0.59 (0.17)	0.61 (0.14)

Mean F_1-score(ASD) (SD) for stratified 10-fold CV on the eight select sites (rows) for different model training (columns). No TST: train only on the other sites. TST: train additionally on the other 9 folds in the target site. TST = target site training; CV = cross validation; ASD = Autism Spectrum Disorder; SD = standard deviation.

Finally, instead of including entire sites in the multi-site models, we selected a subset of the data from the non-target sites (see section 4.2.7), to check whether this approach would improve results on the target site. Due to the dependence of this approach on the training set from the target site, we ran this experiment only on the biggest site, NYU. The mean number of observations selected per fold was 149.30 (SD = 4.52) and 109.00 (SD = 10.52) for the TD and ASD samples, respectively. Mean performance (0.57, SD = 0.12) did not improve with these models, as compared to both intra-single-site models and multi-site models. It was actually closer to the intra-single-site model performance on NYU, suggesting that selecting observations

based on the target site enforced the site-related bias.

4.3.4 Distance Between TD and ASD Samples' Distributions

We went on to compute the MMD-based distance between the distributions of the TD and ASD samples between (Table 4.6's off-diagonal values) and within (Table 4.6's diagonal) the eight select sites. Amongst the eight select sites, Pitt had the greatest intra-site distance between the samples' distributions of the two (TD/ASD) groups, supporting our hypothesis that there was a good separation between the classes in this site. Additionally, distances between samples' distributions were significantly (two-tailed Welch's t-test $p = 2.56e-22$) greater between sites as compared to within sites, even for sites with comparable sample sizes. This suggests that there were non-negligible differences between sites in the samples' distributions. It is highly plausible that these differences affected the advantage different sites could drew from augmenting their sample size with samples from other sites.

Table 4.6. Distance between distributions of TD and ASD observations between and within select sites

TD \ ASD	Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
Leuven	**0.0082**	0.0545	0.0511	0.0502	0.0433	0.0545	0.0690	0.0700
NYU	0.0359	**0.0073**	0.0265	0.0467	0.0239	0.0699	0.0522	0.0422
Pitt	0.0528	0.0398	**0.0165**	0.0576	0.0465	0.0709	0.0551	0.0603
Trinity	0.0368	0.0548	0.0484	**0.0047**	0.0459	0.0827	0.0664	0.0626
UCLA	0.0316	0.0377	0.0437	0.0416	**0.0070**	0.0793	0.0579	0.0363
UM	0.0588	0.0415	0.0499	0.0599	0.0492	**0.0091**	0.0514	0.0587
USM	0.0515	0.0506	0.0545	0.0626	0.0468	0.0700	**0.0104**	0.0336
Yale	0.0418	0.0286	0.0286	0.0520	0.0211	0.0715	0.0411	**0.0057**

MMD between distributions of a site's TD observations (rows) and a site's ASD observations (columns) for the eight select sites. In bold are the distances between the distributions within site. TD = typically developing; ASD = autism spectrum disorder; MMD = maximum mean discrepancy.

4.3.5 Check for Overfitting

The varied behaviors we observed in the inter-single-site experiments (see section 4.3.2) prompted us to check for overfitting. We checked whether any of the models performed significantly better on the training set as compared to the test set. First, we examined inter-single-site models and found that the performance was significantly different (two-tailed Welch's t-test $p \leq 0.0064$) for all pairs of sites but two: training on Pitt and testing on USM (two-tailed Welch's t-test $p = 0.1458$), and training on Trinity and testing on USM (two-tailed Welch's t-test $p = 0.0175$). The models trained on Pitt and Trinity were independent of the test site and therefore were the same whether tested on USM or any of the other sites, thereby suggesting that there was overfitting in all single-site models. In turn, this suggests that the observed separation between the (TD/ASD) classes in Pitt was specific to the site's small sample and might not reflect ASD-related differences.

Next, we examined inter-multi-site models and found that performance was comparable between the training and test sets (two-tailed Welch's t-test $p > 0.01$) for all select target sites but two: Leuven and Pitt (Table 4.7). The multi-site training sets of target sites Pitt and Trinity had ~ 93% of their observations in common. This suggests that the significant differences in performance observed for target sites Leuven and Pitt were not due to overfitting, but instead suggest that the observations in these sites were not well-represented in the other sites, even when the sites were combined. Interestingly, Leuven and Pitt were the only two sites out of the eight select sites for which performance of the multi-site models (Table 4.4.D-E) was worse as compared to the performance on the target sites on their own (intra-single-site models; Table 4.4.A).

Thus, these results suggest that the inter-single-site models and their results in the present work were most likely unreliable. Conversely, most of the multi-site models in the present work appeared to be reliable, suggesting that the combination of multiple sites reduced the bias towards the training set, making the models more generalizable as compared to models trained on a single site.

Table 4.7. Test for overfitting in inter-multi-site models

	Training set	Test set	p-value †
Leuven *	0.70 (0.10)	0.40 (0.14)	$8.83e-09$
NYU	0.68 (0.12)	0.64 (0.07)	0.1266
Pitt *	0.67 (0.13)	0.51 (0.13)	$3.94e-04$
Trinity	0.68 (0.12)	0.56 (0.15)	0.0141
UCLA	0.68 (0.12)	0.64 (0.08)	0.1667
UM	0.59 (0.20)	0.62 (0.07)	0.2926
USM	0.68 (0.11)	0.73 (0.07)	0.0895
Yale	0.68 (0.13)	0.61 (0.14)	0.1420

Mean F_1-score(ASD) (SD) for inter-multi-site models on their respective training set and test set, and p-value between them. † Two-tailed Welch's t-test. * Significant ($p < 0.01$) difference between the performance on the training and test sets. ASD = autism spectrum disorder; SD = standard deviation.

4.3.6 Data Heterogeneity and Demographics

Some studies (e.g., [40, 122, 135]) reported improved performance for demographics-based sub-populations, as compared to the entire sample per site. We examined whether the demographics used to select the sub-populations in such studies were good candidates for latent variables for meaningful clusters of the (TD/ASD) groups. These studies reported best performance for right-handed adolescent ($9-18$ years of age) male participants. Thus, per group, we assigned all right-handed adolescent males to one cluster and the rest of the participants in the group to the other cluster. We performed the test only on the two biggest sites, NYU and UM. For both sites, based on KTST, these clusters were not drawn from different distributions (NYU: $p = 0.03$ for the TD clusters and $p = 0.51$ for the ASD clusters; UM: $p = 0.04$ for the TD clusters and $p = 0.55$ for the ASD clusters). These results suggest that handedness, age, and sex are not a source of heterogeneity in the samples. Further supporting this notion, Pitt, which was the only site for which intra-single-site models outperformed the performance lower bound, had right- and left-handed, male and female participants, and the age variance of participants in both groups was the second greatest amongst the eight select sites (Table 4.2).

4.4 Discussion

In the present work, we examined generalizability of multi-site FC repositories and their usefulness in sample size augmentation, using the ABIDE PCP repository as a case study. Our results suggest that single-site models were not reliable due to overfitting, thus suggesting that models trained on a single site in the repository did not generalize well to other sites. Our results also suggest that multi-site models generalized to other sites in some cases, though not all (Table 4.7). Increasing the heterogeneity of the training set by combining samples from multiple sites may have improved the generalizability of these models, as suggested in a study that have reported similar results for multi-site data of Schizophrenia [145]. Therefore, we focus in this discussion on the intra-single-site and multi-site models and their results.

Based on these results, the performance on some, but not all, of the select sites in the present study was improved by leveraging data from the other select sites (Table 4.4.D-E). However, performance did not improve on any of the select sites for multi-site models trained with all 16 other sites in the repository, as compared to multi-site models trained with only the seven other select sites. This suggests that having more sites in the combined multi-site sample is not always better and that there might be caveats to the use of other sites to improve performance on a target site. Finally, to address the fourth question we posed in the introduction, multi-site models with selective subset of the data from other sites did not improve performance on the NYU. On the contrary, the performance of these models was lower as compared to the multi-site models without subset selection. This suggests that selecting a subset of the data based on the target site's training set might have enforced the site-specific bias.

It should also be notes that our results revealed that the behavior in our experiments varied between sites. This suggests that the common practice of averaging performance across all sites (e.g., [40, 131]) might lead to misleading conclusions. In the remainder of this discussion, we examine these behaviors in order to shed light on the answers we reached for our research questions.

Based on our results, only four of the eight select sites (UCLA, UM, USM, and Yale) appeared to have benefited from augmenting their sample size using the repository,

as they achieved best performance with the largest sample size available in the eight select sites. The other select sites, however, varied in behavior: NYU achieved best performance with a multi-site model that was trained on all other select sites only (without additional training on data from NYU). The performance on Leuven did not surpass what we deemed lower bound for any of the models, but it improved for multi-site models as compared to intra-single-site models. The performance on Trinity also did not surpass the lower bound for any of the models, but it dropped for multi-site models as compared to intra-single-site models. Lastly, the performance on Pitt surpassed the lower bound only for intra-single-site models.

4.4.1 Sites Demonstrating Desirable Behavior

UCLA, UM, USM, and Yale achieved best performance using the largest sample sizes available to them in the eight select sites, thus displaying the desired results of multi-site training using the select sites. However, multi-site models with all the sites in the repository did not improve performance as compared to multi-site models using only the select sites. This suggests that models trained on multiple sizable sites might generalize to other sites, but adding more sites with small sample sizes did not improve the model's ability to generalize to other sites. As our results also suggest non-negligible differences in the distributions of samples between sites (Table 4.6), it is possible that inter-site differences that are not ASD-relevant could not be discerned using small samples from many different sites.

In the case of UM, the mean performance on the site was significantly (two-tailed paired-sample t-test $p = 0.007$) better for multi-site models with target-site training as compared to intra-single-site models. However, the mean performance on UM for a multi-site model without target-site training was not significantly different from the mean performance on UM for either of those models ($p = 0.02$ with intra-single-site and $p = 0.29$ with multi-site models with target-site training). These results suggest that the multi-site model trained with all other select sites generalized fairly well to UM, and although it did not capture all of the variability in UM's sample, additional training on data from UM fine-tuned the model to fit the site better. Out of 10 folds, nine achieved identical or better performance on the multi-site models with target-site

training as compared to a multi-site model without target-site training (Table B.5).

On the three other sites, mean performance was comparable per site for intra-single site models, a multi-site model without target-site training, and multi-site models with target-site training. The lack of significance could be due to the small sample size, as we used only 10 folds due to the sites' small sample sizes. However, the performance on Yale was identical in all folds for the multi-site model without target-site training and the multi-site models with target-site training. Thus, it is possible that the multi-site model without target-site training captured all the variability in Yale, or that the additional training data from Yale could not fine-tune the model to better fit the site and its variability. The performance on UCLA and USM was highly comparable between the two types of multi-site models (with and without target-site training; $p = 0.76$ and $p = 1.00$ for UCLA and USM, respectively). These results suggest that the multi-site models trained on the non-target select sites indeed generalized to Yale, UCLA, and USM. Furthermore, these results suggest that additional training on the target site did not hurt the generalizability of the model trained on all non-target select sites.

4.4.2 Sites Demonstrating Other Behaviors

The performance on the other four select sites was not optimal for the multi-site models, but the sites varied greatly in behavior. Thus, we believe closer examination of the behaviors of each of these sites could shed light on potential pitfalls in multi-site training.

First, the best performance on NYU was achieved with a multi-site model, trained on all other select sites, but the mean performance dropped when the multi-site model was additionally trained on 90% of NYU per fold. Additionally, NYU had a substantially larger sample as compared to the other seven select sites. All other select sites, combines, had 445 participants (233 TD and 212 ASD), while NYU training set included $153 - 155$ participants ($87 - 88$ TD and $65 - 66$ ASD) per fold. Therefore, for multi-site models with target-site training, the training data from NYU made up $\sim 26\%$ of the training set, making it more likely to shift the GNB model towards the model trained on data from NYU alone (intra-single-site models).

Supporting this hypothesis, a two-tailed paired-sample t-test revealed that the mean performance on NYU was significantly ($p = 8.82e-04$) better for a multi-site model without target-site training as compared to intra-single-site models, but neither of them was significantly different from the mean performance of multi-site models with target-site training ($p = 0.06$ with intra-single-site and $p = 0.02$ with a multi-site model without target-site training). These results suggest that a multi-site model trained on all other select sites generalized to NYU, but additional training on NYU hurt the generalizability of the model and shifted it towards the intra-single-site model, which is suggested by our results to suffer from overfitting (see section 4.3.5).

Next, we consider Leuven and Trinity. The mean performance on these two sites did not surpass what we deemed lower bound for any of the models. Our results suggest poor separation between the two (TD/ASD) groups within Trinity, as the distance between the distributions of the two groups within this site was the smallest amongst the eight select sites (Table 4.6, diagonal). Additionally, the distribution of TD observations from Trinity was closer to the distribution of ASD observations in the combined sample of the rest of the select sites than to the distribution of TD observations in the combined sample (Table B.6). Similar behavior was observed for Leuven and its respective combined sample. This suggests that the distributions of samples from these sites may have been too different from the distributions of samples from the other select sites, resulting in the performance of their respective multi-site models falling below the lower bound.

Finally, Pitt was the only select site for which the mean performance of intra-single-site models (see section 4.3.1) surpassed what we deemed lower bound (Table 4.4). That is despite Pitt having the second smallest sample size (23 TD and 22 ASD participants) amongst the eight select sites (Table 4.2). Additionally, no other model (training on one or more additional sites) improved the performance on Pitt. We therefore hypothesized that the samples in this site were relatively homogeneous, and that the source distributions of the two (TD/ASD) groups were relatively separable. Supporting this hypothesis, the distance between the groups' distributions within-site were substantially greater for Pitt than for any of the other select sites (Table 4.6).

4.4.3 Heterogeneity within Samples

We used a GNB mixture model (GNBMM; see section B.1.2) to test the hypothesis that the eight select sites, excluding Pitt, had heterogeneous samples, drawn from multiple distributions, but the results were inconclusive (see section B.2.2). This could be due to the small sample sizes or an incorrect number of components (perhaps more components could have yielded better results). Alternatively, it is possible that a different clustering method (e.g., one that takes into account the density of both of the groups' samples) would have yielded better results.

4.5 Concluding Remarks

In conclusion, in the present work we addressed four main questions (posed in the introduction) and answered them in respect to the ABIDE PCP repository and our methods:

1. Models trained on a single site did not generalize to other sites.

2. Some, but not all, of the models trained on multiple sites generalized to other sites.

3. Performance on some single sites improved by leveraging data from other sites. Moreover, improvement using multiple sites was often better and more reliable, as compared to improvement using a single site.

4. Performance on NYU did not improve by leveraging selective subsets of data from other sites. While this could be attributed to our method of subset selection, we postulate that selecting the additional data based on the target site's training set enforced the site-specific bias and resulted in models that were closer to intra-single-site models as compared to the multi-site models without subset selection.

Additionally, we suggested a few pitfalls associated with multi-site data, we believe are worth taking into consideration in future work: While training on multiple samples can improve the generalizability of models, there are a couple of caveats.

First, samples should be sizable; including too many small samples might hinder the generalizability of models. Second, multiple samples with one of which being substantially larger than the others might also hurt the generalizability of models, as the larger sample may dominate the model and result in a bias, specific to that sample's site. Finally, our results suggest that demographics such as age, handedness, and sex, which are commonly used to get a "more homogeneous" subset of the data, may not be a good determinant of useful sub-populations.

4.6 Limitations

There are some caveats associated with the results of our analyses and our interpretation of them. The first caveat has to do with the dimensions of the data representation: the data representation in the present work suffers from the curse of dimensionality, having $19,900$ features and less than 900 observations overall. This might affect distance metrics, clustering, classifiers, and statistical power. Nevertheless, using even the smallest atlas available from the ABIDE PCP repository (the HO or TT atlases, with 110 ROIs each), would have resulted in $\frac{110\times109}{2} = 5,995$ features, still exceeding the number of observations by a factor of 6. The second caveat has to do with the version of the data: As mentioned in the main text, there are 16 options available from ABIDE PCP for the data preprocessing. Different preprocessing pipelines might have yielded different results. Finally, the third caveat relates to the model we used in this work: GNB is a simple model, and therefore might be less sensitive to nuances in the data as compared to more complex models.

Chapter 5
Longitudinal FC

The recent availability of large-scale longitudinal fMRI data [14] has provided an opportunity to study the evolution of FC that accompanies temporal phenomena, such as healthy development, aging, and disease progression, within individuals [56]. Section 5.2 in this chapter is based on the results from [146].

5.1 Domain-Specific Challenges

Alongside exciting opportunities, longitudinal FC data also present domain-specific challenges, such as inter-individual differences and the variable time intervals in the data. These challenges should be well understood and carefully addressed in order to get meaningful results.

5.1.1 Inter-Individual Differences

No two individuals are identical, and neither are their brains. Some people are early developers, while others are late bloomers, the same disease may progress at different rates in different individuals, and different people have different experiences, which in turn might affect their FC. For example, inter-individual differences have been shown in brain development [62]. Thus, one might be wrong to assume, for example, that FC networks from two 12-year-old healthy male individuals are comparable. While researchers try to minimize these differences in the participant recruitment stage, they cannot be completely avoided.

5.1.2 Longitudinal Data

Longitudinal data, as opposed to a time series of data, have variable time intervals between consecutive data points. In longitudinal fMRI data the time intervals vary not only within participants, but also between participants. It is important to correctly align data points from different participants in relation to the temporal axis of the subject matter. For example, when examining the progression of a disease with variable onset age, aligning participants based on age may compromise results. If the disease also progresses at different rates, a single reference point at its onset may compromise results as well.

In most domains, seldom do we find multiple network instantiations with the same nodes. For example, different social networks have different actors, and different protein-protein interaction networks have possibly overlapping, but different, sets of proteins. Corresponding nodes in (whole-brain or seed-based) FC networks are assumed to have the same identity. That is, the left amygdala, for example, is considered to be the same brain region and node in FC networks from different participants. FC networks from different participants can therefore be viewed essentially as independent instantiations of the same network. They are instantiations (one per participant) in the sense that we assume complete node correspondence between FC networks from different participants. They are independent in the sense that a perturbation in one instantiation has no baring on another instantiation (in contrast to the network science's sub-field of inderdependent networks). Thus, while longitudinal studies in other domains often examine a single instantiation of a network over time, longitudinal FC studies examine multiple independent instantiations of a network.

To further complicate matters, currently available longitudinal fMRI data often have fewer than 10 data points per participant. Thus, while we postulate that alignment using our novel family of algorithms (see section 5.2) would lead to more accurate comparison between individuals, and thereby more meaningful results, the application of the algorithms to real-life longitudinal FC data is outside the scope of the present work. Nevertheless, in contrast to the previously-mentioned challenges, this challenge could dissipate over time, as more samples become available.

5.2 PATENet: Pairwise Alignment of Temporaly-Evolving Networks

Network science has provided a variety of powerful tools for describing, representing, and analyzing a variety of real-world systems including social networks, the internet, functional brain networks, and biomolecular networks [43,44]. While many of the tools and techniques of network science, e.g., topological analyses and network alignment, focus on networks with fixed topologies, the structure of networks that represent real-world systems change over time. Such networks are naturally modeled as time-evolving networks (TENs) [147, 148]. TENs can display dynamics on networks (where the network structure does not change over time, but the activity of the nodes does); dynamics of networks (where the activity of nodes does not change but the structure does); and dynamics of and on networks (where both the structure and activity change over time) [20]. The relatively young sub-field of TENs [44] has already yielded a substantial body of work, focusing primarily on models of time-evolving networks and the characterization of network dynamics [148]. However, there is limited work on methods for comparative analyses of TENs.

To motivate the underlying problem, consider experimental subjects who undergo fMRI recordings of resting state brain activity at different points in time, e.g., in the context of a longitudinal study of changes in FC as a function of development, aging, or disease progression [56]. The resulting data from each subject are naturally represented as a temporally ordered sequence of FC networks. To complicate matters, it may not be straightforward to establish one-to-one correspondence between the recording times across subjects because of differences in the timing of recordings, missed recording sessions, etc. Furthermore, even in the case of subjects with recordings obtained at what appears to be temporally-matching data points (e.g., age in years), because of differences in the onset and progression of development, aging, or disease, and the trajectories across subjects, the networks at the respective time points may not be comparable. With the exception of [149], which focuses on temporal registration of deforming meshes, to the best of our knowledge, there is no work on aligning (temporally) ordered sequences of networks (OSNs). The

most closely related body of work focuses on aligning ordered sequences of letters over a finite alphabet, e.g., DNA or protein sequences [150], video frames [151], and clinical histories [152]. However, with the exception of methods for aligning sequences of letters [150], the methods used are ad-hoc and are not supported by a sound mathematical rationale and hence lack precise mathematical characterization and are not amenable to generalization to other related problem domains.

Against this background, we focus on the problem of aligning a pair of OSNs. Specifically, we describe PATENet, a mathematically sound family of algorithms for aligning a pair of OSNs. PATENet requires as input, in addition to a pair of OSNs to be aligned, a measure of pairwise similarity of fixed topology networks, a monotonically increasing function, and a match threshold. It produces as output an optimal alignment of the given pair of OSNs. Specifically, PATENet generalizes the Smith-Waterman (SW) algorithm [150], a dynamic programming algorithm for aligning two ordered sequences of letters, given a pairwise measure of substitutability of letters and gap penalties. SW produces an optimal local alignment, i.e., aligned segments of the given pair of sequences with the largest cumulative similarity. Conceptually, adapting the SW algorithm to yield a mathematically sound algorithm for aligning a pair of OSNs is straightforward; we replace letters by networks, and replace pairwise substitutability of letters by a well-behaved measure of pairwise similarity of (fixed topology) networks. However, in order for this approach to yield both mathematically sound and practically useful algorithms for aligning OSNs, several challenges need to be addressed; there are a variety of measures of similarity or distance between networks that are tailored [153] to meet the needs of specific applications [154]. We need to adjust such measures so as to ensure that the algorithms that use them for aligning OSNs are mathematically well-behaved. In the current work we also show that the PATENet family of algorithms can be readily extended to align ordered sequences of elements other than networks, provided suitable and well-behaved measures of similarity between elements are available.

5.2.1 Preliminaries

We use $G = G(\mathcal{V}, \mathcal{E}_G)$ to denote a network, where $\mathcal{V}$ is its set of nodes and $\mathcal{E}_G$ is its set of edges. We define OSN $\mathcal{G}$ to be a sequence of n networks, $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$, where $\forall 1 \leq i \leq n \in \mathbb{N}, G_i = \left(\mathcal{V}, \mathcal{E}_{G_i}\right)$ denotes the i^{th} element of $\mathcal{G}$, which is a snapshot of a TEN at time t_i, and $\forall 1 < i \leq n \in \mathbb{N}, t_{i-1} < t_i$. We use upper case letters, e.g., H, to denote matrices or networks, lower case letters, e.g., n, to denote scalars, and script letters, e.g., $\mathcal{V}$, to denote sets.

Definition 1. Let $G = (\mathcal{V}, \mathcal{E})$ and $G' = \left(\mathcal{V}, \mathcal{E}'\right)$ be two networks with the same set of nodes $\mathcal{V}$ and respective sets of edges $\mathcal{E}$ and $\mathcal{E}'$ (either identical or different). A function $s\left(G, G'\right)$, mapping two graphs to $[0, 1]$, is said to be a well-defined unsigned normalized network similarity measure (UNNSM) if it satisfies the following properties (adapted from [155]):

1. Identity property: $s\left(G, G'\right) \leq s(G, G) = 1, \forall G, G'$.
2. Symmetry property: $s\left(G, G'\right) = s\left(G', G\right), \forall G, G'$.
3. Minimum property: $s\left(G, G'\right) \xrightarrow{|\mathcal{V}| \to \infty} 0$ where WLOG G is the complete network, and G' is the empty network (i.e., $\mathcal{E}^C = \mathcal{E}'$).

Definition 2. Similarly, a function $s'\left(G, G'\right)$, mapping two graphs to $[-1, 1]$, is said to be a well-defined signed normalized network similarity measure (SNNSM) if it satisfied the properties described in definition 1, with the minimum property adjusted to the signed range: $s'\left(G, G'\right) \xrightarrow{|\mathcal{V}| \to \infty} -1$ (rather than 0).

For simplification purposes we assumed G and G' to have the same set of nodes $\mathcal{V}$. However, if $\mathcal{V}_G \neq \mathcal{V}_{G'}$, where $\mathcal{V}_G$ and $\mathcal{V}_{G'}$ denote the sets of nodes of G and G', respectively, then $\mathcal{V} = \mathcal{V}_G \bigcup \mathcal{V}_{G'}$ for the definitions above.

5.2.2 The Smith-Waterman Sequence Alignment Algorithm

The SW algorithm is a local sequence alignment algorithm, designed to find pairs of segments with high cumulative degree of similarity between two sequences of amino

acids (AAs), $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$. There are 22 AAs, and the similarity between every pair of AAs is specified by the entries of a 'substitution matrix' $SM \in \mathbb{R}^{22\times22}$. The SW algorithm uses dynamic programming to generate a 'scoring matrix' $H = H(A, B) \in \mathbb{R}^{(n+1)\times(m+1)}$, which is defined as follows:

$$\begin{aligned}
&\forall 0 \le i \le n \in \mathbb{N}, \forall 0 \le j \le m \in \mathbb{N}, H_{i,0} = H_{0,j} = 0 \\
&\forall 0 < i \le n \in \mathbb{N}, \forall 0 < j \le m \in \mathbb{N}, \\
&\quad H_{i,j} = \max\left\{H_{i-1,j-1} + s\left(a_i, b_j\right), \max_{1\le k\le i}\left\{H_{i-k,j} - w_k\right\}, \max_{1\le l\le j}\left\{H_{i,j-l} - w_l\right\}, 0\right\}
\end{aligned} \tag{5.1}$$

Where $s\left(a_i, b_j\right)$ is the similarity score between the two AAs $a_i \in A$ and $b_j \in B$, according to SM, and w_k is a value assigned to deletions or insertions of length k. Insertions and deletions refer to cases where an element (or a few) within one sequence is not aligned with an element (or a sequence of elements) within the paired sequence. The length of insertions and deletions is the number of consecutive insertions and/or deletions. $w_1 \in \mathbb{R}$ is referred to as 'gap penalty' and is the value assigned to a gap of length 1, and $w_k = f(w_1, k) \in \mathbb{R}$ is the penalty for a gap of length k, where $f(w_1, k)$ can be affine or linear, for example, in relation to w_1.

Let $\mathcal{X}$ denote the maximum value in H, then $\mathcal{X} : A \times B \to \mathbb{R}$ is the local alignment score between the two sequences A and B, and is used to reveal the best local alignment by way of backtracing on H. Starting at a cell holding $\mathcal{X}$, backtracing until a cell holding 0 is reached according to the following logic:

1. If $H_{i,j} = H_{i-1,j-1} + s\left(a_i, b_j\right)$, then a_i is aligned with b_j and the process continues from $H_{i-1,j-1}$.

2. Else if $H_{i,j} = H_{i-1,j} - w_1$, then a_i has no alignment in B and the process continues from $H_{i-1,j}$.

3. Else if $H_{i,j} = H_{i,j-1} - w_1$, then b_j has no alignment in A and the process continues from $H_{i,j-1}$.

The solution is not guaranteed to be unique; there could be multiple cells in H holding $\mathcal{X}$, in which case the backtracing process can be initiated at any of these cells, resulting in different, yet equally good, local alignments.

5.2.3 PATENet

In this paper we focus on aligning a pair of OSNs. To accommodate OSNs resulting from longitudinal recordings from subjects, we impose the following natural desiderata on the alignments returned by PATENet:

1. Preservation in the alignment of the relative order of elements within the sequences. E.g., if the 3^{rd} element of the first sequence is aligned with the 5^{th} element of the second sequence, then the 4^{th} element of the first sequence can be aligned only with elements in positions 6 or greater in the second sequence.

2. Accommodation of unaligned elements in both sequences. I.e., aligning two sequences, S_1 and S_2, of length n and m, respectively, should not force the alignment of $\min\{n, m\}$ elements.

3. Accommodation of longitudinal gaps. E.g., allowing for time points existing in one sequence but missing in the other.

5.2.3.1 Alternative Substitution Matrix Construction

The SW algorithm requires a well-defined SM, holding both positive values for possible matches and negative values for non-matches. Furthermore, unlike in the case of AA sequences, where the sequence elements are drawn from a fixed alphabet, OSNs can contain arbitrary networks defined over a given set of vertices and edges. Hence, we will adapt existing network similarity measures to define pairwise similarity of elements (networks) in OSNs.

Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively. Let s be a well-defined UNNSM. Finally, let $0 < \varphi < 1 \in \mathbb{R}$ be a threshold on s, where $\forall G_i \in \mathcal{G}, \forall G'_j \in \mathcal{G}', match\left(G_i, G'_j\right) = \begin{cases} 1, & \text{if } \varphi \leq s\left(G_i, G'_j\right) \\ 0, & \text{if } \varphi > s\left(G_i, G'_j\right) \end{cases}$, and let $\ell(x), \ell : [0,1] \rightarrow [-1,1]$ be a signed normalized monotonically increasing transform, with $\ell(\varphi) = 0$, $\ell(0) = -1$, and $\ell(1) = 1$. We propose

$$SM_{i,j} = \ell\left(s\left(G_i, G'_j\right)\right), \forall 1 \leq i \leq n \in \mathbb{N}, \forall 1 \leq j \leq m \in \mathbb{N} \tag{5.2}$$

to construct an 'alternative substitution matrix' $SM = SM\left(\mathcal{G}, \mathcal{G}'\right) \in [-1,1]^{n \times m}$. For example, for $\alpha = \frac{\varphi}{1-\varphi}$ and $0.5 < \varphi < 1$, $\ell(x) = 1 - \log_\alpha\left[\alpha^2 + \left(1-\alpha^2\right) \cdot x\right]$ is such a signed normalized monotonically increasing transform (proof omitted), and along with $DeltaCon\left(G, G'\right)$ [155] as the well-defined UNNSM (by definition), can be used to construct an alternative SM. Another example includes $\tilde{s}\left(G, G'\right) = \left(1 - NSSD\left(G, G'\right)\right)$ as the well-defined UNNSM (proof omitted), where $NSSD\left(G, G'\right)$ is the normalized sum squared difference, and $\tilde{\ell}(x) = \begin{cases} \frac{x-\tilde{\varphi}}{1-\tilde{\varphi}}, & \text{if } x \geq \tilde{\varphi} \\ \frac{x-\tilde{\varphi}}{\tilde{\varphi}}, & \text{if } x < \tilde{\varphi} \end{cases}$ for $0 < \tilde{\varphi} < 1$ as the signed normalized monotonically increasing transform (proof omitted).

Lemma 5.2.1. Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively, and let s be a well-defined UNNSM and $\ell : [0,1] \to [-1,1]$ be a signed normalized monotonically increasing transform, as described above. Then $\ell\left(s\left(G, G'\right)\right)$, mapping two graphs to $[-1,1]$, satisfies the properties of a well-defined SNNSM.

Proof. Identity: $s\left(G, G'\right) \leq s(G,G) \implies \ell\left(s\left(G, G'\right)\right) \leq \ell(s(G,G)) = \ell(1) = \ell\left(\max\left\{s\left(G, G'\right)\right\}\right) = \max\left\{\ell\left(s\left(G, G'\right)\right)\right\} = 1$ □
Symmetry: $s\left(G, G'\right) = s\left(G', G\right) \implies \ell(s(G, G')) = \ell(s(G', G))$ □
Minimum: $\ell\left(s\left(G, G'\right)\right) \xrightarrow{|\mathcal{V}| \to \infty} \ell(0) = \ell\left(\min\left\{s\left(G, G'\right)\right\}\right) = \min\left\{\ell\left(s\left(G, G'\right)\right)\right\} = -1$ □

5.2.3.2 From SW to PATENet

The SW algorithm meets the first two desiderata of PATENet (preservation of temporal order and accommodation of possible unaligned elements in both sequences). To satisfy the third desideratum (accommodation of longitudinal gaps), we set the gap penalty to zero. Therefore, for an alternative SM, following the construction described above, the scoring matrix of PATENet $\tilde{H} = \tilde{H}\left(\mathcal{G}, \mathcal{G}'\right) \in \mathbb{R}^{(n+1)\times(m+1)}$, hereafter referred to as 'OSN scoring matrix', is specified as follows:

$$\begin{aligned} &\forall 0 \leq i \leq n \in \mathbb{N}, \forall 0 \leq j \leq m \in \mathbb{N}, \tilde{H}_{i,0} = \tilde{H}_{0,j} = 0 \\ &\forall 0 < i \leq n \in \mathbb{N}, \forall 0 < j \leq m \in \mathbb{N}, \\ &\quad \tilde{H}_{i,j} = \max\left\{\tilde{H}_{i-1,j-1} + SM_{i,j}, \max_{1 \leq k \leq i}\left\{\tilde{H}_{i-k,j}\right\}, \max_{1 \leq l \leq j}\left\{\tilde{H}_{i,j-l}\right\}, 0\right\} \end{aligned} \tag{5.3}$$

Lemma 5.2.2. Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively. Let $\tilde{H} \in \mathbb{R}^{(n+1)\times(m+1)}$ be their OSN scoring matrix, then $\forall 1 \leq i \leq n \in \mathbb{N}, \forall 1 \leq j \leq m \in \mathbb{N}$:

1. $\tilde{H}_{i-1,j} \leq \tilde{H}_{i,j}$
2. $\tilde{H}_{i,j-1} \leq \tilde{H}_{i,j}$

Proof. 1. $\tilde{H}_{i,j} = \max\left\{\tilde{H}_{i-1,j-1} + SM_{i,j}, \max\limits_{1\leq k\leq i}\left\{\tilde{H}_{i-k,j}\right\}, \max\limits_{1\leq l\leq j}\left\{\tilde{H}_{i,j-l}\right\}, 0\right\} \geq \max\limits_{1\leq k\leq i}\left\{\tilde{H}_{i-k,j}\right\} \geq \tilde{H}_{i-1,j}$ □

2. $\tilde{H}_{i,j} = \max\left\{\tilde{H}_{i-1,j-1} + SM_{i,j}, \max\limits_{1\leq k\leq i}\left\{\tilde{H}_{i-k,j}\right\}, \max\limits_{1\leq l\leq j}\left\{\tilde{H}_{i,j-l}\right\}, 0\right\} \geq \max\limits_{1\leq l\leq j}\left\{\tilde{H}_{i,j-l}\right\} \geq \tilde{H}_{i,j-1}$

□

Lemma 5.2.3. Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively. Let $\tilde{H} \in \mathbb{R}^{(n+1)\times(m+1)}$ be their OSN scoring matrix, then $\forall 1 \leq i \leq n \in \mathbb{N}, \forall 1 \leq j \leq m \in \mathbb{N}$:

1. $\max\limits_{1\leq k\leq i}\left\{\tilde{H}_{i-k,j}\right\} = \tilde{H}_{i-1,j}$
2. $\max\limits_{1\leq l\leq j}\left\{\tilde{H}_{i,j-l}\right\} = \tilde{H}_{i,j-1}$

Proof. Intuitive based on Lemma 5.2.2 □

Therefore, the OSN scoring matrix $\tilde{H}$ of PATENet is equivalent to:

$$\begin{aligned} &\forall 0 \leq i \leq n \in \mathbb{N}, \forall 0 \leq j \leq m \in \mathbb{N}, \tilde{H}_{i,0} = \tilde{H}_{0,j} = 0 \\ &\forall 0 < i \leq n \in \mathbb{N}, \forall 0 < j \leq m \in \mathbb{N}, \\ &\quad \tilde{H}_{i,j} = \max\left\{\tilde{H}_{i-1,j-1} + SM_{i,j}, \tilde{H}_{i-1,j}, \tilde{H}_{i,j-1}, 0\right\} \end{aligned} \tag{5.4}$$

Lemma 5.2.4. Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively. Let $\tilde{H} = \tilde{H}\left(\mathcal{G}, \mathcal{G}'\right) \in \mathbb{R}^{(n+1)\times(m+1)}$ and $\tilde{H}' = \tilde{H}'\left(\mathcal{G}', \mathcal{G}\right) \in \mathbb{R}^{(m+1)\times(n+1)}$ be their OSN scoring matrices, and let SM and SM' be the alternative substitution matrices of $\tilde{H}$ and $\tilde{H}'$, respectively. Then $\forall 1 \leq i \leq n \in \mathbb{N}, \forall 1 \leq j \leq m \in \mathbb{N}$:

1. $SM_{i,j} = SM'_{j,i}$

2. $\tilde{H}_{i,j} = \tilde{H}'_{j,i}$

Proof. 1. $SM_{i,j} = \ell\left(s\left(G_i, G'_j\right)\right) = \ell\left(s\left(G'_j, G_i\right)\right) = SM'_{j,i}$ □

2. For $i = j = 1$:

$$\begin{aligned}\tilde{H}_{1,1} &= \max\left\{\tilde{H}_{0,0} + SM_{1,1}, \tilde{H}_{0,1}, \tilde{H}_{1,0}, 0\right\} = \max\left\{SM_{1,1}, 0\right\} = \max\left\{SM'_{1,1}, 0\right\} \\ &= \tilde{H}'_{1,1}\end{aligned}$$

$\forall 2 \leq i \leq n \in \mathbb{N}$ and $j = 1$ we can safely assume $\tilde{H}_{i-1,1} = \tilde{H}'_{1,i-1}$ for induction:

$$\begin{aligned}\tilde{H}_{i,1} &= \max\left\{\tilde{H}_{i-1,0} + SM_{i,1}, \tilde{H}_{i-1,1}, \tilde{H}_{i,0}, 0\right\} = \max\left\{SM_{i,1}, \tilde{H}_{i-1,1}, 0\right\} \\ &= \max\left\{SM'_{1,i}, \tilde{H}'_{1,i-1}, 0\right\} = \tilde{H}'_{1,i}\end{aligned}$$

$\forall 2 \leq j = k \leq m \in \mathbb{N}$ and $i = 1$:

$$\begin{aligned}\tilde{H}_{1,k} &= \max\left\{\tilde{H}_{0,k-1} + SM_{1,k}, \tilde{H}_{0,k}, \tilde{H}_{1,k-1}, 0\right\} = \max\left\{SM_{1,k}, \tilde{H}_{1,k-1}, 0\right\} \\ &= \max\left\{SM_{1,k}, \max\left\{SM_{1,k-1}, \tilde{H}_{1,k-2}, 0\right\}, 0\right\} \\ &= \max\left\{SM_{1,k}, SM_{1,k-1}, \tilde{H}_{1,k-2}, 0\right\} = \dots \\ &= \max\left\{SM_{1,k}, SM_{1,k-1}, \dots, SM_{1,1}, 0\right\} \\ &= \max\left\{SM'_{k,1}, SM'_{k-1,1}, \dots, SM'_{1,1}, 0\right\} = \tilde{H}'_{k,1}\end{aligned}$$

$\forall 2 \leq j = k \leq m \in \mathbb{N}$ and $\forall 1 \leq i \leq n \in \mathbb{N}$, we can safely assume $\tilde{H}_{i,k-1} = \tilde{H}'_{k-1,i}$ as well as $\tilde{H}_{i-1,k} = \tilde{H}'_{k,i-1}$, and therefore also $\tilde{H}_{i-1,k-1} = \tilde{H}'_{k-1,i-1}$ for induction:

$$\begin{aligned}\tilde{H}_{i,k} &= \max\left\{\tilde{H}_{i-1,k-1} + SM_{i,k}, \tilde{H}_{i-1,k}, \tilde{H}_{i,k-1}, 0\right\} \\ &= \max\left\{\tilde{H}'_{k-1,i-1} + SM'_{k,i}, \tilde{H}'_{k,i-1}, \tilde{H}'_{k-1,i}, 0\right\} = \tilde{H}'_{k,i}\end{aligned}$$

□

Lemma 5.2.5. Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively. Let $\tilde{H} = \tilde{H}\left(\mathcal{G}, \mathcal{G}'\right) \in \mathbb{R}^{(n+1)\times(m+1)}$ be their OSN scoring matrix, and let SM be its alternative substitution matrix. Then the alignment score $\tilde{\mathcal{X}} = \max\left\{\tilde{H}\right\}$[1] is equivalent

[1]Notice that $\tilde{\mathcal{X}} : \mathcal{G} \times \mathcal{G}' \to \mathbb{R}$.

to $\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i, G'_j\right)\cdot SM_{i,j}\right]$, where
$\rho\left(G_i, G'_j\right) = \begin{cases} 1, & \text{if } \left(G_i, G'_j\right) \text{ are aligned with each other} \\ 0, & \text{Otherwise} \end{cases}$.

Proof. By definition of the SW algorithm, $\forall 1 \leq i \leq n \in \mathbb{N}, \forall 1 \leq j \leq m \in \mathbb{N}, \tilde{H}_{i,j}$ = the maximum similarity of two segments ending in G_i and G'_j. The similarity score of the alignment is the sum of similarity scores between every pair of aligned elements and weights of all insertions and deletions in the alignment. Since $w_i = 0$ in PATENet, the weights of all insertions and deletions is always 0, leaving only the sum of similarity scores between every pair of aligned elements, which can be written as: $\tilde{\mathcal{X}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i, G'_j\right)\cdot SM_{i,j}\right]$, where $\rho\left(G_i, G'_j\right) = \begin{cases} 1, & \text{if } \left(G_i, G'_j\right) \text{ are aligned with each other} \\ 0, & \text{Otherwise} \end{cases}$. □

5.2.3.3 OSN Alignment Score

Alignment of elements across a pair of OSNs may be informative by itself and reveal temporally-preserved similarities between the two OSNs. However, another concept worth borrowing from sequence alignment is that of the alignment score $\tilde{\mathcal{X}} = \max\left\{\tilde{H}\right\}$.

Lemma 5.2.6. An OSN alignment score $\tilde{\mathcal{X}} = \max\left\{\tilde{H}\right\}$ satisfies properties that are similar to those of a well-defined UNNSM, except for the normalization-related upper bound:

1. Identity property[2]: $\tilde{\mathcal{X}}\left(\mathcal{G}, \mathcal{G}'\right) \leq \tilde{\mathcal{X}}\left(\mathcal{G}, \mathcal{G}\right), \forall \mathcal{G}, \mathcal{G}'$.
2. Symmetry property: $\tilde{\mathcal{X}}\left(\mathcal{G}, \mathcal{G}'\right) = \tilde{\mathcal{X}}\left(\mathcal{G}', \mathcal{G}\right), \forall \mathcal{G}, \mathcal{G}'$.
3. Minimum property: $\tilde{\mathcal{X}}\left(\mathcal{G}, \mathcal{G}'\right) \xrightarrow{|\mathcal{V}|\to\infty} 0$ where WLOG $\mathcal{G}$ is the complete OSN, and $\mathcal{G}'$ is the empty OSN (i.e., $\forall 1 \leq i \leq n \in \mathbb{N}, \forall 1 \leq j \leq m \in \mathbb{N}, \mathcal{E}^{C}_{G_i} = \mathcal{E}_{G'_j}$).

Proof. Lemma 5.2.5 $\Rightarrow \tilde{\mathcal{X}}\left(\mathcal{G}, \mathcal{G}'\right) = \max\left\{\tilde{H}\right\} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i, G'_j\right)\cdot SM_{i,j}\right]$, where $\rho\left(G_i, G'_j\right) = \begin{cases} 1, & \text{if } \left(G_i, G'_j\right) \text{ are aligned with each other} \\ 0, & \text{Otherwise} \end{cases}$.

[2]Notice that $\tilde{\mathcal{X}}\left(\mathcal{G}, \mathcal{G}\right) = 1$ is not required, as the alignment score has no upper bound.

1. Identity: $\tilde{\mathcal{X}}(\mathcal{G},\mathcal{G}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\rho\left(G_i,G_j\right)\cdot SM_{i,j}\right] = \sum_{i=1}^{n}\left[1\cdot 1\right] = n$ and $\tilde{\mathcal{X}}\left(\mathcal{G},\mathcal{G}'\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i,G'_j\right)\cdot SM_{i,j}\right] \leq \sum_{i=1}^{n}\left[1\cdot 1\right] = n = \tilde{\mathcal{X}}(\mathcal{G},\mathcal{G})$ □

2. Symmetry: $\tilde{\mathcal{X}}\left(\mathcal{G},\mathcal{G}'\right) = \max\left\{\tilde{H}\right\} = \max\limits_{1\leq i\leq n,1\leq j\leq m}\left\{\tilde{H}_{i,j}\right\} = \max\limits_{1\leq j\leq m,1\leq i\leq n}\left\{\tilde{H}'_{j,i}\right\}$ $= \max\left\{\tilde{H}'\right\} = \tilde{\mathcal{X}}\left(\mathcal{G}',\mathcal{G}\right)$ □

3. Minimum: $\forall 1\leq i\leq n\in\mathbb{N}, \forall 1\leq j\leq m\in\mathbb{N},\ s\left(G_i,G'_j\right)\xrightarrow{|\mathcal{V}|\to\infty}0 \Rightarrow \rho\left(G_i,G'_j\right)\xrightarrow{|\mathcal{V}|\to\infty}0 \Rightarrow \tilde{\mathcal{X}}\left(\mathcal{G},\mathcal{G}'\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i,G'_j\right)\cdot SM_{i,j}\right]\xrightarrow{|\mathcal{V}|\to\infty}$ $\sum_{i=1}^{n}\sum_{j=1}^{m}\left[0\right] = 0$

□

We observe that PATENet can be used to extend the UNNSM, used for constructing SM, into an unsigned normalized order-aware OSN similarity measure. Let $\mathcal{G}$ and $\mathcal{G}'$ be two OSNs with n and m elements, respectively. Let $\tilde{H} = \tilde{H}\left(\mathcal{G},\mathcal{G}'\right)\in\mathbb{R}^{(n+1)\times(m+1)}$ be the corresponding OSN scoring matrix, and let SM be its alternative substitution matrix. Let s be the well-defined UNNSM used for constructing SM, and $\rho\left(G_i,G'_j\right) = \begin{cases} 1, & \text{if } \left(G_i,G'_j\right) \text{ are aligned with each other} \\ 0, & \text{Otherwise} \end{cases}$. Then $h : \mathcal{G}\times\mathcal{G}'\to[0,1]$ can be defined as:

$$h\left(\mathcal{G},\mathcal{G}'\right) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i,G'_j\right)\cdot s\left(G_i,G'_j\right)\right]}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho\left(G_i,G'_j\right),1\right\}} \tag{5.5}$$

which is hereafter referred to as an 'OSN similarity score'.[3]

Lemma 5.2.7. An OSN similarity score $h\left(\mathcal{G},\mathcal{G}'\right)$ satisfies identity, symmetry and minimum properties, similar to those that hold for UNNSM:

1. Identity property: $h\left(\mathcal{G},\mathcal{G}'\right)\leq h(\mathcal{G},\mathcal{G}) = 1, \forall\mathcal{G},\mathcal{G}'$.

[3]Notice that the OSN similarity score measures similarity in the context of the locally aligned segments of the sequences. That is, if OSNs $\mathcal{G}$ and $\mathcal{G}'$ have k elements aligned with average element-wise similarity of l, whether $k = \min\{n,m\}$ or $k < \min\{n,m\}$, $h(\mathcal{G},\mathcal{G}') = l$. Additionally, if OSNs $\mathcal{G}$ and $\mathcal{G}'$ have one element aligned with element-wise similarity of 1.0, while OSNs $\mathcal{G}$ and $\mathcal{G}''$ have four elements aligned with each element-wise similarity being 0.9, $h\left(\mathcal{G},\mathcal{G}'\right) > h\left(\mathcal{G},\mathcal{G}''\right)$ (but $\tilde{\mathcal{X}}\left(\mathcal{G},\mathcal{G}'\right) < \tilde{\mathcal{X}}\left(\mathcal{G},\mathcal{G}''\right)$)

2. Symmetry property: $h\left(\mathcal{G},\mathcal{G}'\right)=h\left(\mathcal{G}',\mathcal{G}\right),\forall\mathcal{G},\mathcal{G}'$.

3. Minimum property: $h\left(\mathcal{G},\mathcal{G}'\right)\xrightarrow{|\mathcal{V}|\to\infty}0$ where WLOG $\mathcal{G}$ is the complete OSN, and $\mathcal{G}'$ is the empty OSN (i.e., $\forall 1\leq i\leq n\in\mathbb{N},\forall 1\leq j\leq m\in\mathbb{N},\mathcal{E}^{C}_{G_i}=\mathcal{E}_{G'_j}$).

Proof. WLOG, we assume $n\leq m$.

1. Identity: $h\left(\mathcal{G},\mathcal{G}\right)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\rho(G_i,G_j)\cdot s(G_i,G_j)\right]}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\rho(G_i,G_j),1\right\}}=\frac{\sum_{i=1}^{n}\left[1\cdot s(G_i,G_i)\right]}{\max\left\{\sum_{i=1}^{n}1,1\right\}}=\frac{n}{n}=1$ and $h\left(\mathcal{G},\mathcal{G}'\right)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i,G'_j\right)\cdot s\left(G_i,G'_j\right)\right]}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho\left(G_i,G'_j\right),1\right\}}\leq\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho\left(G_i,G'_j\right)}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho\left(G_i,G'_j\right),1\right\}}\leq 1=h\left(\mathcal{G},\mathcal{G}\right)$ □

2. Symmetry: $h\left(\mathcal{G},\mathcal{G}'\right)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i,G'_j\right)\cdot s\left(G_i,G'_j\right)\right]}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho\left(G_i,G'_j\right),1\right\}}=\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\left[\rho\left(G'_j,G_i\right)\cdot s\left(G'_j,G_i\right)\right]}{\max\left\{\sum_{j=1}^{m}\sum_{i=1}^{n}\rho\left(G'_j,G_i\right),1\right\}}$ $=h\left(\mathcal{G}',\mathcal{G}\right)$ □

3. Minimum: $\forall 1\leq i\leq n\in\mathbb{N},\forall 1\leq j\leq m\in\mathbb{N},\ s\left(G_i,G'_j\right)\xrightarrow{|\mathcal{V}|\to\infty}0\Rightarrow\rho\left(G_i,G'_j\right)\xrightarrow{|\mathcal{V}|\to\infty}0\Rightarrow h\left(\mathcal{G},\mathcal{G}'\right)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\rho\left(G_i,G'_j\right)\cdot s\left(G_i,G'_j\right)\right]}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\rho\left(G_i,G'_j\right),1\right\}}\xrightarrow{|\mathcal{V}|\to\infty}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}[0]}{\max\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}0,1\right\}}=0$

□

5.2.4 Experiments

We now proceed to describe a set of experiments that explore the behavior of PATENet under a variety of conditions. Because "ground truth" alignments for real-world OSNs are unavailable, we generated synthetic OSNs for this purpose. Although PATENet has three user-defined parameters, we experimented with different match thresholds, while keeping the other two parameters (a well-defined UNNSM and a signed normalized monotonically increasing transform) constant, as they are more application- and domain-specific.

5.2.4.1 Empirical Design

We experimented with PATENet with a substitution matrix based on DeltaCon [155] and a logarithmic signed normalized monotonically increasing transform function ($\ell\left(x\right)=1-\log_{\alpha}\left[\alpha^2+\left(1-\alpha^2\right)\cdot x\right]$ where $\alpha=\frac{\varphi}{1-\varphi}$ for $0.5<\varphi<1$). DeltaCon

assesses node affinities similarity between two undirected networks with known node correspondence. It is a well-defined UNNSM (by definition).

To examine the robustness of PATENet to noise in the data, we corrupted one of the OSNs - containing otherwise identical subset of (in our experiments with synthetic data, six) elements in the OSNs to be aligned - with different levels of Gaussian noise added to the edge weights. Since PATENet uses a static match threshold, we also examined the interaction between the effect of noise on PATENet's performance and the choice of match threshold φ. We experimented with $\varphi = \{0.51, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90\}$ and Gaussian noise with $\mu = 0$ and $\sigma = \{0.1, 0.3, 0.5, \ldots, 3.9\}$.

Performance of alignment was evaluated using 'goodness of alignment', defined as the percentage of elements with known ground-truth match (based on the construction of the OSNs) that were aligned with their ground-truth matches.

5.2.4.2 Synthetic Data Generation

We constructed three sets of synthetic data: (1) random dynamic OSNs, (2) Barabasi-Albert (BA) [156] dynamic OSNs, and (3) Dorogovtsev-Mendes (DM) [157] dynamic OSNs. The BA and DM models describe evolving (rather than dynamic) networks, hence we adapted only the edge addition/removal portions of these models. For each dataset we examined three temporal conditions: linear, a single change in trend, and two changes in trend. The resulting OSNs consisted of 25 elements each, starting from an undirected random network with 50 nodes and a connectivity rate of ~ 0.12 (141 edges out of possible 1225). We use $\mathcal{O}$ to denote such a 25-element OSN. We further experimented with the percent of edges added/removed from one element to the next in $\mathcal{O}$, using one of four percentages: 1%, 2%, 4%, or 8%.

Random dynamic OSNs were generated as follows: element 1 was generated using the Erdos-Renyi (ER) model [158]. In case of linear $\mathcal{O}$s, edges were added at random to generate elements 2-25 (see Fig. 5.1A). Single trend change $\mathcal{O}$s were generated by adding edges at random to generate elements 2-13, and then removing edges at random to generate elements 14-25 (see Fig. 5.1B). For $\mathcal{O}$s with two trend changes, elements 2-9 were generated by adding edges at random, elements 10-17 were generated by removing edges at random, and elements 18-25 were generated by

adding edges at random (see Fig. 5.1C).

BA and DM dynamic OSNs were generated in a manner similar to random dynamic OSNs, with the following changes: the BA model [159] with 50 nodes and $n = 3$ (resulting in 141 edges out of possible 1225, similarly to the ER-based element 1) was used to generate element 1, and edges to be added were selected based on the corresponding model. Edge removal is done at random according to both models.

In any OSN $\mathcal{O}$, 12 (roughly half) of the elements were selected at random and kept in order to make up a new OSN, denoted by $\mathcal{M}$. Half (six) of the elements selected for $\mathcal{M}$ were then removed from $\mathcal{O}$ to generate a new OSN with 19 elements, denoted by $\mathcal{O}'$. Consequently, any pair of OSNs $\left(\mathcal{O}', \mathcal{M}\right)$ constructed according to the preceding procedure, shares six random elements, and $\mathcal{M} \nsubseteq \mathcal{O}'$ and $\mathcal{O}' \nsubseteq \mathcal{M}$ (Fig. 5.1D). Gaussian noise (see section 5.2.4.1) was added only to $\mathcal{M}$ prior to alignment.

5.2.4.3 Results

In all three synthetic datasets, experiments revealed a similar relationship between the performance of PATENet, user-specified threshold φ, and the added Gaussian noise (Fig. 5.2). For lower values of φ, PATENet showed a high degree of noise tolerance, significantly outperforming random alignment over a broad range of Gaussian noise levels. As φ increased, so did PATENet's susceptibility to noise, but for tolerable levels of noise, its performance was similar or better, as compared to PATENet with lower φ for the same noise level. We conclude that the choice of φ affects multiple aspects of the performance of PATENet in the presence of noise.

5.2.5 Discussion

5.2.5.1 Additional Considerations and Future Directions

In real world OSN data, e.g., those derived from longitudinal studies of functional brain connectivity networks, at present, there are no effective approaches to estimating the noise level in the data. Our results demonstrate a trade-off between PATENet's resistance to noise and performance with low levels of noise as a function of the choice of match threshold. Hence, in practical settings, it might be worth exploring a probabilistic combination of different match thresholds.

Some natural directions include PATENet as an OSN kernel, to use in classification and regression problems where the input to the classifier is an OSN. Possible applications include assigning subjects to different categories (e.g., normal development, accelerated development, retarded development) based on the observed development from longitudinal studies. Another natural direction for future work is to extend PATENet to align multiple OSNs (as opposed to a pair of OSNs). The resulting multi-sequence variant of PATENet can also be used to cluster OSNs.

5.2.5.2 Generalizations

The empirically-demonstrated version of PATENet is limited to the case where the elements of the OSNs are undirected networks with pre-specified correspondence between nodes in each element of one OSN and nodes in each element of any other OSN to be aligned with it. It would be interesting to explore variants of PATENet that can work with OSNs consisting of directed graphs, graphs with both directed and undirected edges, or colored graphs (with multiple types of nodes and/or edges), etc. It would also be interesting to consider variants of PATENet that can work in settings where the correspondence between nodes in each element of one OSN and nodes in each element of any other OSN to be aligned with it is not specified, but instead needs to be established based on some node similarity criteria [160].

Furthermore, while in this paper we have focused on the pairwise alignment of OSNs, the PATENet algorithms can be further generalized to work with ordered sequences of arbitrary elements (instead of networks) so long as we can specify a well-behaved unsigned normalized similarity measure between such elements.

5.2.6 Concluding Remarks

Networks that change over time, e.g., functional brain networks that change their structure due to processes such as development or aging, are naturally modeled by TENs. Longitudinal measurements of such TENs are naturally represented as OSNs, where each network in the sequence represents a static snapshot of the TEN at a specific time of observation. In this paper we proposed PATENet, a novel family of algorithms for optimal local alignment of pairs of OSNs. The algorithms require three

user-defined inputs in addition to a pair of OSNs to be aligned: a well-defined UNNSM, a signed normalized monotonically increasing transform, and a match threshold. We showed how PATENet can be used to compute an alignment score, as well as a similarity score, for a pair of aligned OSNs.

Our experiments using PATENet to align synthetic OSNs produced using different generative models of OSNs with their noise corrupted counterparts show that: at lower match thresholds, PATENet displays a high degree of noise tolerance, significantly outperforming random alignment over a broad range of noise levels; at higher match thresholds (more stringent match criteria), PATENet shows increased susceptibility to noise.

PATENet offers a mathematically sound approach to aligning OSNs, which is amenable to being generalized along a number of dimensions, e.g., OSNs consisting of directed networks, labeled networks, or even ordered sequences of other types of elements.

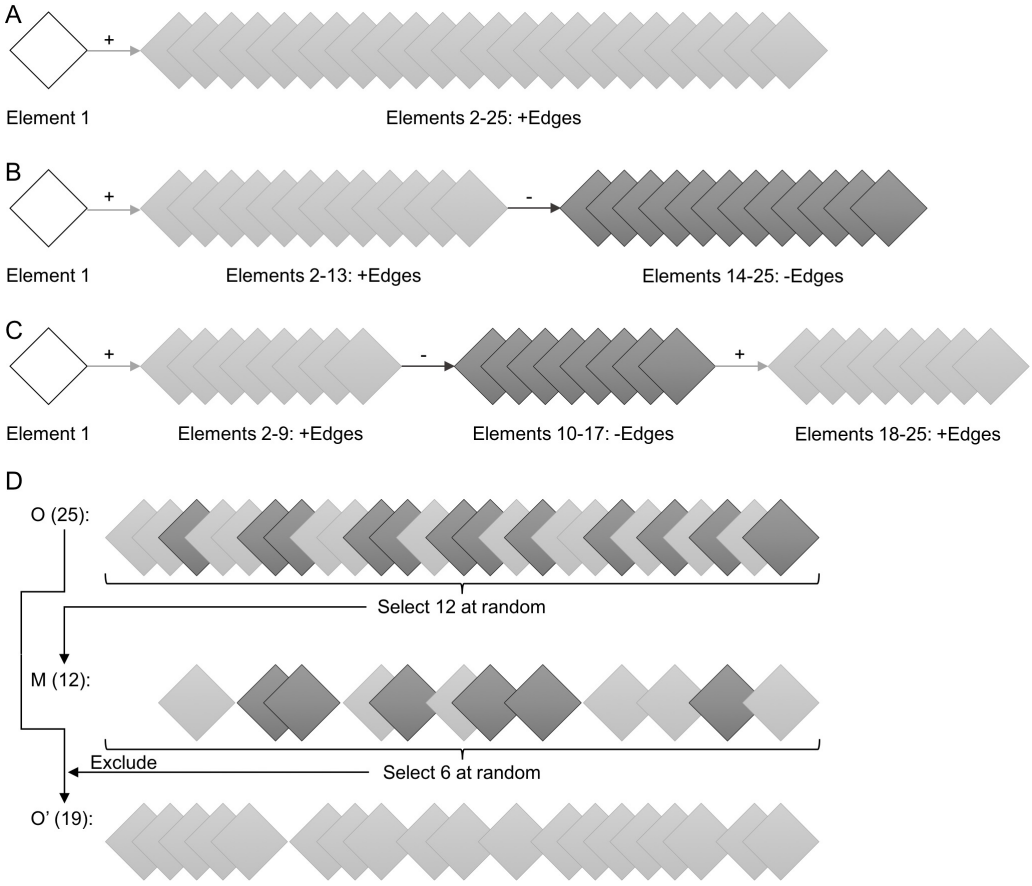

Figure 5.1. Synthetic data generation. (A-C) Generation of a linear OSN (A), a single trend-change OSN (B) and an OSN with two trend changes (C). All OSNs $\mathcal{O}$s (A-C) consist of 25 elements, the first element (white) being a random graph (ER for random dynamics OSNs and BA for BA and DM OSNs). Rectangles represent elements, with light gray indicating increase trend (edges added between elements) and dark gray indicating decrease trend (edges removed between elements); + edges are added between elements; − edges are removed between elements. (D) Generation of OSNs $\mathcal{M}$ and $\mathcal{O}'$ from OSN $\mathcal{O}$. In dark gray are the elements selected according to the corresponding description in the text. Starting from 25 elements in $\mathcal{O}$, 12 elements are selected at random to create $\mathcal{M}$, six of which are removed from the copy of $\mathcal{O}$ to $\mathcal{O}'$ (resulting in 19 elements in $\mathcal{O}'$).

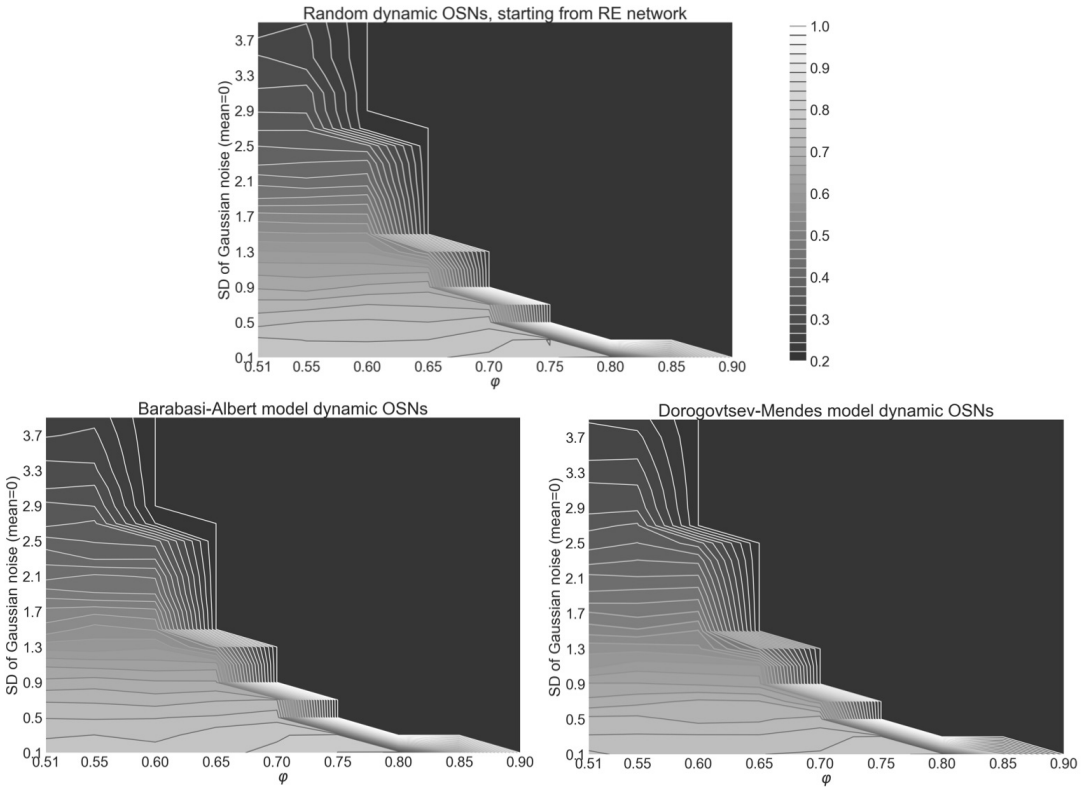

Figure 5.2. Effect of noise and match threshold on PATENet's performance. Goodness of alignment of PATENet with the synthetic data as a function of added Gaussian noise (vertical axis) and match threshold φ (horizontal axis). Goodness of alignment was normalized and averaged across all four percentages (of edges added/removed between elements: 1%, 2%, 4%, and 8%) and three temporal conditions (linear, single trend-change, and two trend-changes OSNs; 12 conditions overall), as their patterns were comparable. Top: random dynamic OSNs, starting from ER network. Bottom left: BA dynamic OSNs. Bottom right: DM dynamic OSNs. The same color bar is used in all three plots, ranging from the average performance of random alignment (comparable in all three datasets) to perfect alignment (1.0).

Chapter 6
Conclusions

In this chapter, we summarize the contributions of this dissertation and list some of the opportunities for future work.

6.1 Summary and Contributions

The present dissertation identifies existing and emerging challenges in the FC domain, based on its current state, and addressed them using ML-based techniques.

- **Demonstrated the potential of ML-based data-driven FC analyses to complement activation-based analyses:** In chapter 3 we applied ML techniques to static FC networks from data that have been previously examined using activation-based analyses. While no population-condition interactions were found in the activation-based analyses, our FC analyses reveled such interactions. E.g., our results suggested that in older adults (but not in younger adult), participants' alertness differed between the conditions.

- **Uncovered population-dependent and -independent features of static FC that discriminated between phonological and semantic conditions:** Our results (in chapter 3) suggested that the left FP had grater immediate FC in the semantic task as compared to the phonological task, independently of population. Conversely, our results suggested that the immediate functional neighbors of the left and right FOs were more functionally connected to each other for the phonological task than for the semantic task.

- **Suggested cautionary notes for multi-site learning:** One of the main purposes of multi-site fMRI repositories is to increase the size and heterogeneity of samples in order to improve the reproducibility and generalizability of results. In chapter 4 we used the ABIDE PCP repository as a case study for multi-site FC learning. Our results landed support to the hypothesis that models trained on multiple sizable samples may generalize better as compared to models trained on a single sample. However, we revealed potentials pitfalls to be aware of when using multi-site FC repositories such as ABIDE PCP: First, training on many small samples may hurt the generalizability of the model (possibly making it more difficult to discern inter-site differences from population-related differences). Second, multi-site training that includes one considerably larger sample as compared to the rest (e.g., NYU in our experiments) may also hurt the generalizability of the model (as that site may dominate the combined sample and lead to a model closer to a single-site model).

- **Novel family of algorithms for local alignment of pairs of evolving FC networks:** In chapter 5 we introduced PATENet, a novel family of algorithms for optimal local alignment of pairs of OSNs. PATENet offers a mathematically sound approach to aligning OSNs. Furthermore, it is amenable to being generalized along a number of dimensions (e.g., OSNs consisting of directed networks, labeled networks, and even ordered sequences of none-network complex elements).

- **Novel similarity measure between pairs of evolving FC networks:** We introduced an "OSN similarity score" (chapter 5), based on PATENet. The OSN similarity score is an unsigned normalized order-aware OSN similarity measure, which can be used to extend any well-defined UNNSM.

- **Domain-aware application of ML techniques to neuroimaging:** Throughout the present work we applied ML techniques to the data in a domain-aware fashion. As the size of FC samples is often considerably smaller than their dimension, FC data can be challenging for simple ML algorithms, let alone for complex ML algorithms, which have more parameters to estimate. For that reason and due to prioritizing interpretability over performance, we favored simple algorithms over complex ones. Additionally, we occasionally employed

state-of-the-art advanced nonparametric tests, such as HSIC, SDCIT, and KTST, which are more robust as compared to their parametric counterparts. Finally, in chapter 4 we applied Laplace correction to our results per site, to account for small and variable sample sizes.

6.2 Future Research Directions

In the present dissertation we examined a few of the existing and emerging challenges and opportunities in ML-based analyses of FC, which have led us to identify additional challenges and opportunities. Here we list some possible directions for future research:

- **Mindful dimensionality reduction in FC:** The multi-site analyses in the present dissertation (chapter 4) reveled that despite combining samples from multiple sites, the data that are currently available in FC repositories still suffer from the curse of dimensionality. Dimensionality reduction can be achieved using feature selection (e.g., see chapter 3) or feature projection (e.g., principal component analysis). Mindful application refers to ensuring that the results are interpretable and that the new representation is not overfitted to the training data.

- **Probabilistic-combination PATENet:** Our analysis of PATENet demonstrated a trade-off between PATENet's resistance to noise and performance with low levels of noise as a function of the choice of match threshold. Hence, in practical settings, where the level of noise in the data is unknown, it might be worth exploring a probabilistic combination of different match thresholds in PATENet instead of using a single match threshold.

- **PATENet-based OSN kernel:** PATENet can be used as a basis to define an OSN kernel. Such kernel may be useful in classification and regression problems where the input to the model is an OSN.

- **Multi-sequence PATENet:** In the present dissertation we have introduced a single-pair variant of PATENet. This variant aligns a single pair of OSNs at a time with each other. However, it can be extended to a multi-sequence

variant of PATENet, aligning multiple OSNs at a time with each other. Such multi-sequence PATENet may be useful for clustering of OSNs.

- **Applying PATENet-based OSN similarity score to bigger samples of real-world evolving FC data on availability:** The longitudinal FC sample that was made available to us during the present dissertation was revealed to not be sufficiently large to yield conclusive results. However, as more longitudinal FC data are collected and shared, the PATENet-based OSN similarity score might become more applicable to real-world evolving FC data. E.g., using the OSN similarity score (or the OSN alignment score) to cluster evolving FC networks from different participants may reveal differently-evolving sub-types of conditions (e.g., healthy development or disease progression), which in turn may provide insight into the condition and its progression.

Appendix A
Condition- and Population-Effects in Static FC: Supplementary Material

A.1 FC Analyses

The functional dataset met two important criteria for applicability of the method we used to generate FC networks: First, the number of trials per task per participant was greater than 30, as emphasized in the closely related generalized form of PPI [161]. Second, as demonstrated by [162], while deconvolution can recover even strongly overlapping sequential events, the events need to be separated by at least 4 s.

FC was computed using Pearson's correlation coefficient on the post-processing time series from the different ROIs. ROIs were set as nodes, and the weights of the edges between every two nodes were set to the correlation coefficient between the time series of the ROIs when $p < 0.05$, and zero otherwise. Stability of p-value threshold was assessed by examining nearby thresholds, as well as much lower thresholds (0.01 and 0.001), which resulted in adjacency matrices very close to the ones generated in our analysis, based on NSSD. NSSDs (mean±SD) between the FC networks used in the current study (with $p < 0.05$) and those generated with $p < 0.045$ (Fig. A.1), $p < 0.055$ (Fig. A.2), $p < 0.01$ (Fig. A.3), and $p < 0.001$ (Fig. A.4) were $2.41e{-}06 \pm 2.17e{-}06$, $2.04e{-}06 \pm 1.99e{-}06$, $4.45e{-}05 \pm 3.28e{-}05$, and $1.31e{-}04 \pm 9.61e{-}05$, respectively.

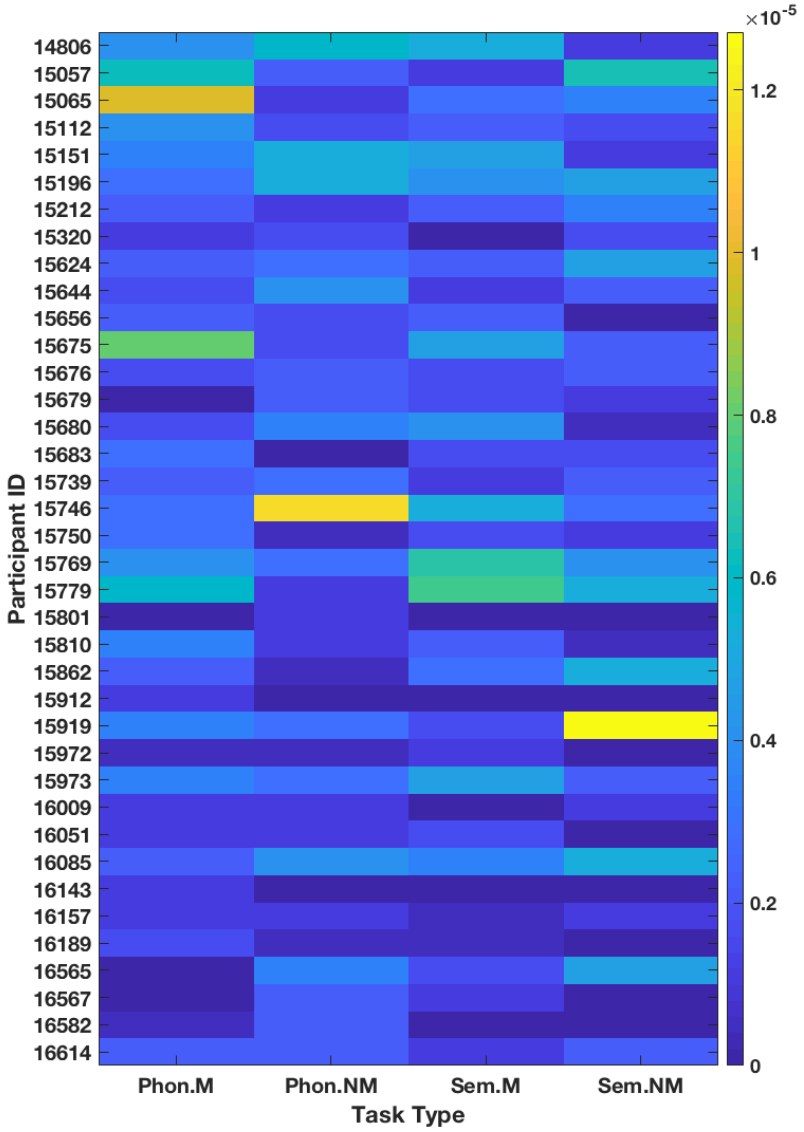

Figure A.1. NSSD between p > 0.05 and p > 0.045 Pearson's-based full correlation matrices, using up to eight voxels per ROI. For all participants, for all conditions, NSSD ≤ 1.27e−05. NSSD = normalized sum squared difference; ROI = region of interest; Phon = phonological; Sem = semantic; M = match; NM = mismatch

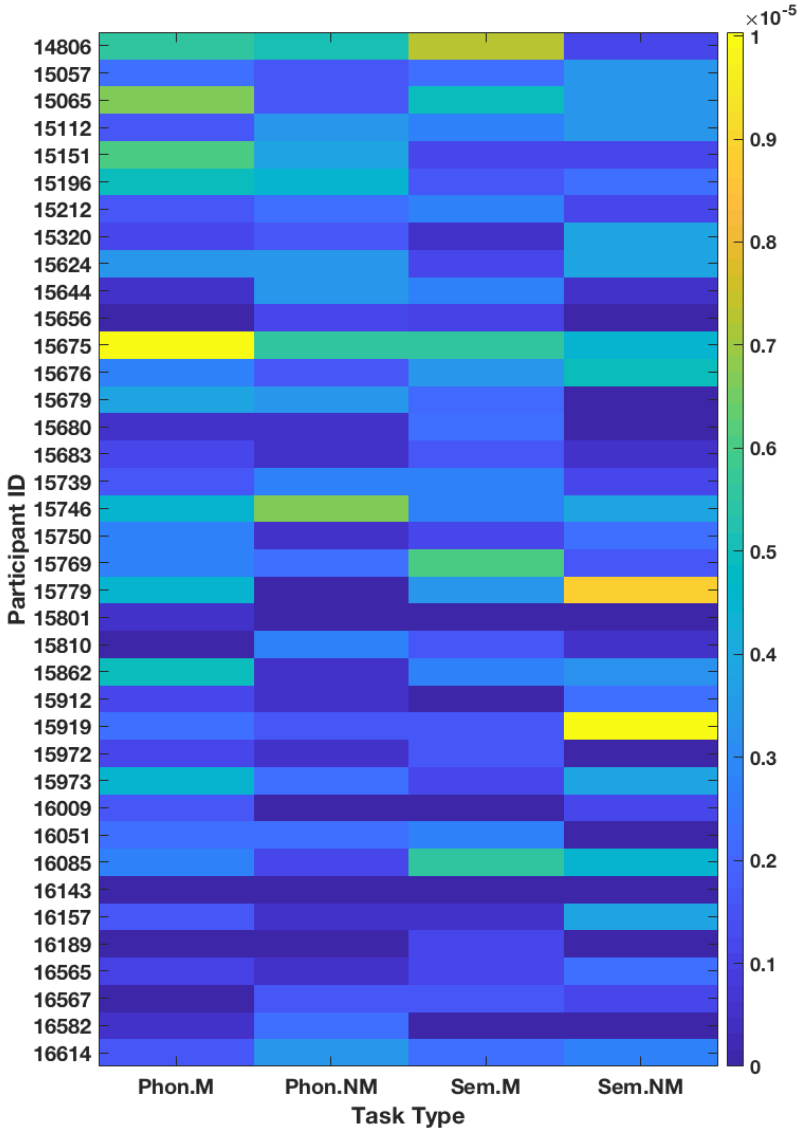

Figure A.2. NSSD between p > 0.05 and p > 0.055 Pearson's-based full correlation matrices, using up to eight voxels per ROI. For all participants, for all conditions, NSSD ≤ 1.00e−05. NSSD = normalized sum squared difference; ROI = region of interest; Phon = phonological; Sem = semantic; M = match; NM = mismatch

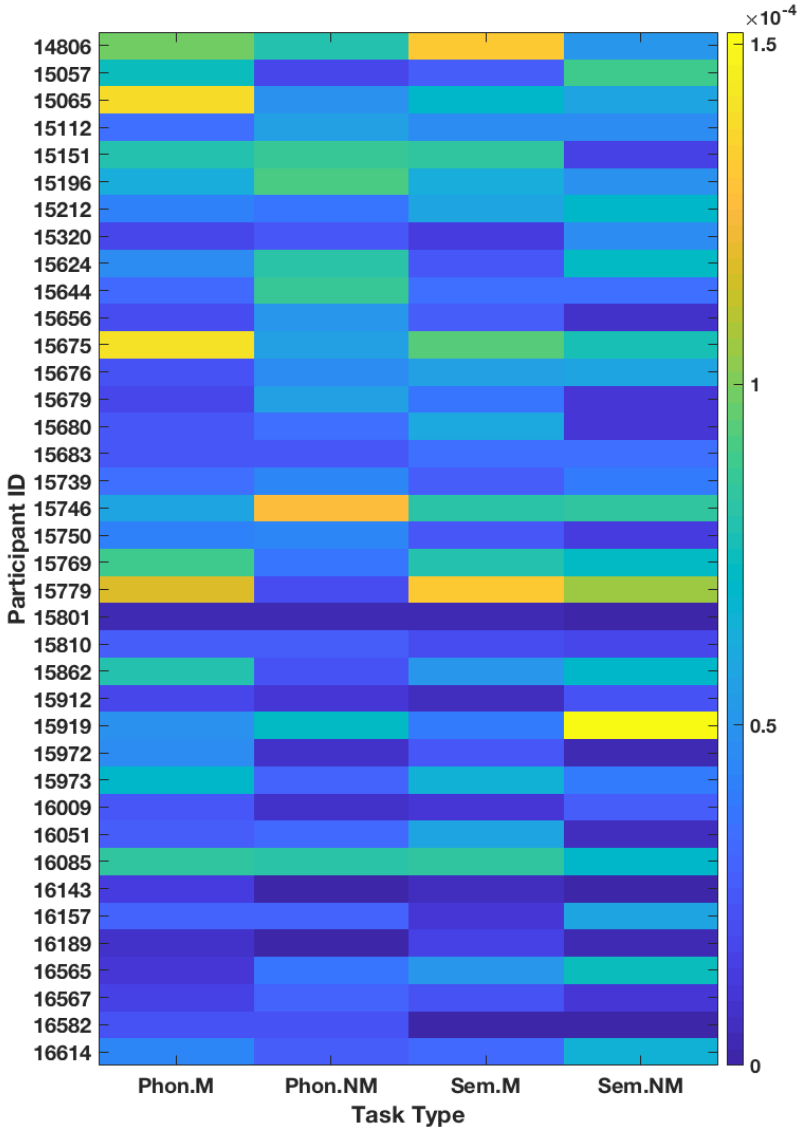

Figure A.3. NSSD between p > 0.05 and p > 0.01 Pearson's-based full correlation matrices, using up to eight voxels per ROI. For all participants, for all conditions, NSSD ≤ 0.0002. NSSD = normalized sum squared difference; ROI = region of interest; Phon = phonological; Sem = semantic; M = match; NM = mismatch

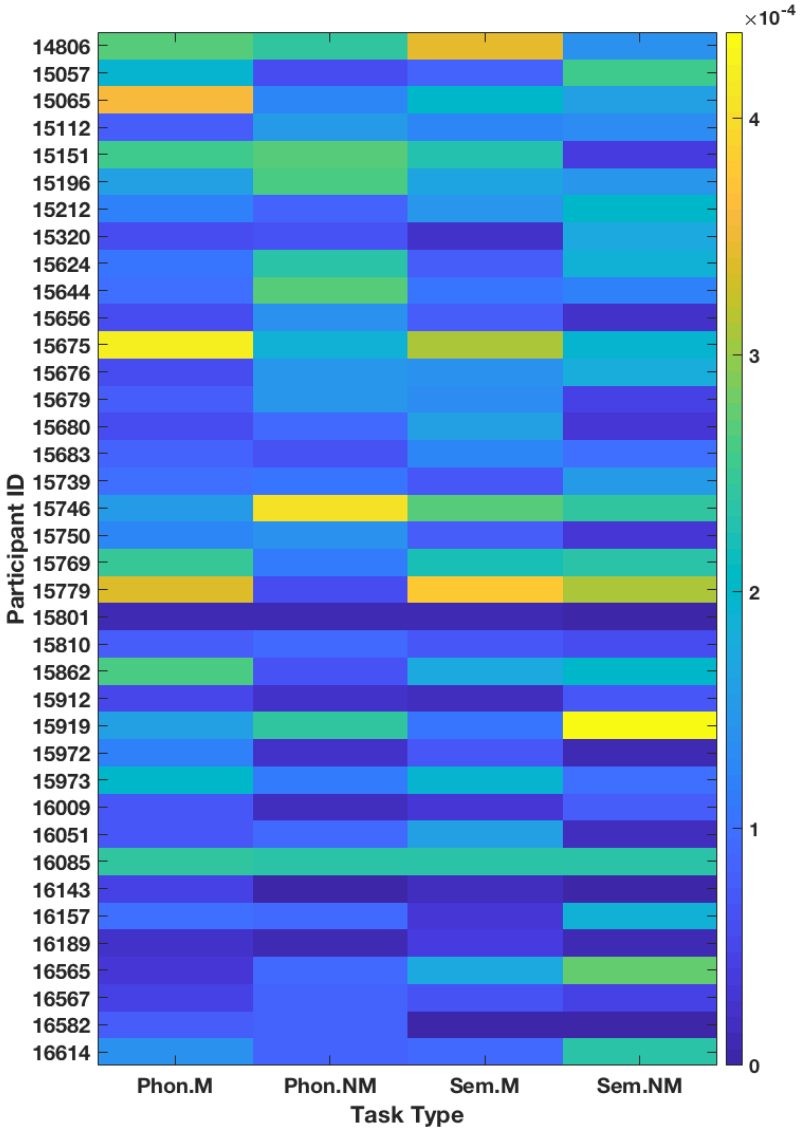

Figure A.4. NSSD between $p > 0.05$ and $p > 0.001$ Pearson's-based full correlation matrices, using up to eight voxels per ROI. For all participants, for all conditions, NSSD ≤ 0.0004. NSSD = normalized sum squared difference; ROI = region of interest; Phon = phonological; Sem = semantic; M = match; NM = mismatch

For reference, NSSDs between FC networks (with $p < 0.05$) of match and mismatch phonological/semantic trials within participants were 0.05 ± 0.02, and were significantly greater than the NSSDs above, based on two-tailed two-sample heteroscedastic T-test ($p = 8.66e{-}31$ for $p < 0.045$ and $p < 0.055$ FC networks, $p = 9.12e{-}31$ for $p < 0.010$ FC networks, and $p = 1.01e{-}30$ for $p < 0.001$ FC networks; $\alpha = 0.01$, $\beta \to 0$ for all).

Additionally, we examined the effect of the number of voxels selected per ROI, generating FC networks with up to two peak voxels per run per ROI. Similarly to the generation of the FC networks used in this study, if no supra-threshold voxels existed for an ROI in a particular run, no voxel was selected for the ROI for the run. If only one supra-threshold voxel existed for an ROI in a particular run, only that voxel was selected for the ROI for the run. The mean NSSD between these FC networks and their respective counterparts in the FC networks used in this study (Fig. A.5) was $1.40e{-}35$ (SD = $8.10e{-}04$) and significantly ($p = 4.78e{-}30$, $\alpha = 0.01$, $\beta \to 0$) lower than the mean NSSD between FC networks of match and mismatch phonological/semantic trials within participants (0.05; SD = 0.02).

Finally, wavelet correlation [163] was examined as a more time-sensitive alternative to Pearson's correlation. Maximal overlap discrete wavelet correlation, using scale 2 (corresponding to $(0.0625, 0.125]$ Hz frequencies) Daubechies external phase wavelet with four vanishing moments (db4) [164], also resulted in adjacency matrices very close to the ones in our analysis, based on NSSD (Fig. A.6; mean$\pm$SD = 0.664 ± 0.201), which were significantly ($p = 3.25e{-}30$, $\alpha = 0.01$, $\beta \to 0$) lower than the mean NSSDs between FC networks of match and mismatch phonological/semantic trials within participants (0.05; SD = 0.02).

A.2 Training Classifiers

We used in-house written Python 3.7 code (available at https://github.com/SGur/LangAge) in PyCharm 2017.3.4 (Community Edition) with scikit-learn version 0.19.1 [143] implementation on MAC (3.2 GHz Intel Core i5; 32 GB 1600 MHz DDR; iOS High Sierra version 10.13.3).

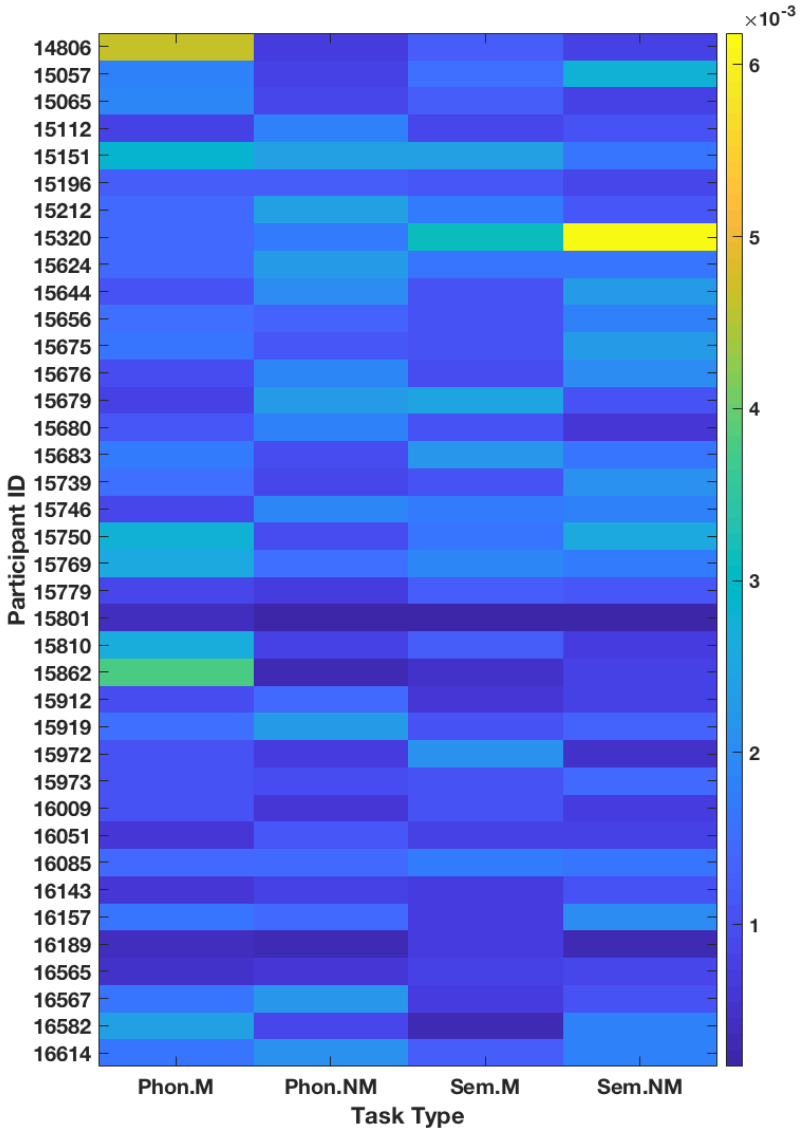

Figure A.5. NSSD between p > 0.05 Pearson's-based full correlation matrices, using up to one voxel and up to two voxels per run per ROI. For all participants, for all conditions, NSSD ≤ 0.01. NSSD = normalized sum squared difference; ROI = region of interest; Phon = phonological; Sem = semantic; M = match; NM = mismatch

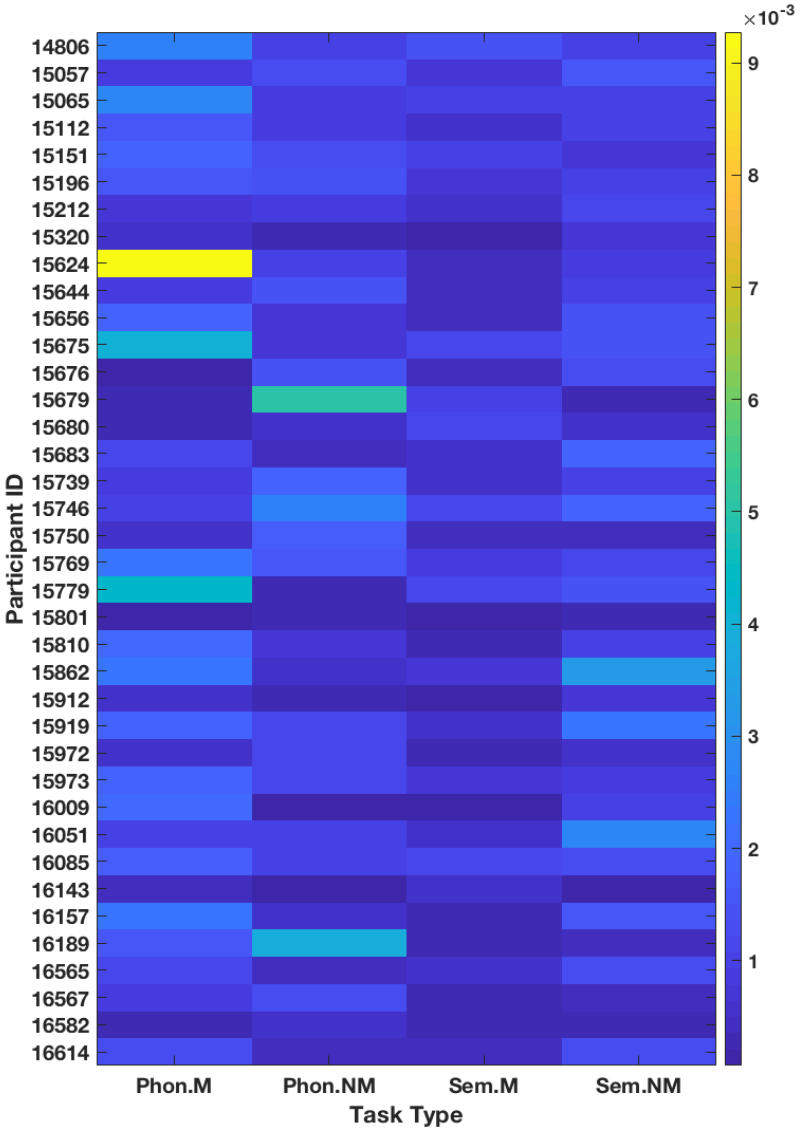

Figure A.6. NSSD between p > 0.05 Pearson's-based and scale 2 db4 MODWT-based full correlation matrices, using up to eight voxels per ROI. For all participants, for all conditions, NSSD ≤ 0.01. NSSD = normalized sum squared difference; ROI = region of interest; Phon = phonological; Sem = semantic; M = match; NM = mismatch; db4 = Daubechies external phase wavelet with four vanishing moments; MODWT = maximal overlap discrete wavelet transform.

A.2.1 Data Availability

A post-preprocessing (ROIs' time series) version of the data analyzed in this study is available at https://github.com/S-Gur/LangAge. The raw data are available from mtd143@psu.edu on reasonable request.

A.3 Statistical analyses

In the present work, we assessed statistical significance of differences in performance between different combinations of instances, features, and algorithms. For this purpose, we used two-tailed two-sample heteroscedastic T-test. This stricter form was used also in cases where looser significance tests (e.g., paired T-test or two-sample homoscedastic T-test) might be applicable. The difference in condition-dependent NSSD between younger and older adults had its statistical significance assessed in the same way.

In addition, we employed nonparametric tests: HSIC [98] for unconditional independence queries and SDCIT [99] for conditional independence queries. These tests depend on the choice of kernel that determines how similar two observations are to each other. For real values (e.g., CCs of left FP and bilateral FO), we employed a Gaussian radial basis function kernel, $k(x, x') = e^{-\gamma|x-x'|^2}$, where γ is determined by median heuristic [98]. For categorical values (age-group and condition), Dirac delta kernel was used, where kernel value for two observations is 1 if they are identical and 0 otherwise.

A.4 Limitations

The main text discusses a couple of limitations to be considered in interpreting the results, which might require further explanation. (1) Big ROIs might have spatially distant, potentially functionally different, voxels selected for them in the different runs from the same participant. In such cases the time series of these voxels might very differently correlate with the time series of other ROIs, resulting in the averaged time series of the ROI to be weakly correlated with the time series of the other ROIs.

However, it should be noted that this risk exists also in the case of averaging across all voxels in each ROI, as practiced in some related work. (2) Smaller ROIs might have their activation peak per run overlap or be very close to each other, so the correlation between their respective time-series and time-series of other ROIs is similar and preserved in the averages time-series. Alternatively, smaller ROIs might not have any supra-threshold voxels in any of the runs, in which case the ROI was functionally isolated.

Appendix B
Multi-Site Static FC: Supplementary Material

B.1 Additional Materials and Methods

B.1.1 Participants

The IDs of the participants in the eight select sites that were used in the present study are listed in table B.1.

B.1.2 GNB Mixture Model

To accommodate the possibility of heterogeneous data, we used GNBMM, allowing for up to two sub-populations per class. Sub-populations were identified using spectral clustering, as implemented in scikit-learn [143] (version 0.20.2), using a kernel function (see section 4.2.6.2).

Let $\ell_{k,1}$ and $\ell_{k,2}$ be the two sub-populations identified for class $c_k \in C = \{c_1, c_2\}$, and let $X = \{x_1, \ldots, x_n\}$ be an observation with n features, then:

$$p(x \mid c_k) = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{p\left(x \mid \ell_{j,i}\right) \cdot p\left(\ell_{j,i} \mid c_k\right)}{p(c_k)} = \sum_{i=1}^{2} \frac{p\left(x \mid \ell_{k,i}\right) \cdot p\left(\ell_{k,i} \mid c_k\right)}{p(c_k)}$$

and therefore:

$$p\left(c_k \mid x\right) \propto \sum_{i=1}^{2}\left[p\left(\ell_{k,i} \mid c_k\right) \cdot \prod_{j=1}^{n} p\left(x_j \mid c_k\right)\right]$$
$$= \sum_{i=1}^{2}\left\{p\left(\ell_{k,i} \mid c_k\right) \cdot \prod_{j=1}^{n}\left[\frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\ell_{k,i},j}}} \cdot e^{-\frac{\left(x_{i,j}-\hat{\mu}_{\ell_{k,i},j}\right)^2}{2\hat{\sigma}^2_{\ell_{k,i},j}}}\right]\right\}$$

For binary classification with classes $C = \{c_1, c_2\}$, GNBMM, similarly to GNB, classifies observation $x = \{x_1, x_2, \ldots, x_n\}$ as c_1 if $0 < \frac{\log p(c_1|x)}{\log p(c_2|x)}$. Due to the sum in the log, the computation of the log likelihood ratio might lead to overflow errors. We addressed it using the "log-sum-exp trick":

$$\begin{cases}\log\left(e^a + e^b\right) = b + \log\left(e^{a-b} + 1\right), & \text{if} |a-b| < \log\max\left(float\_type\right) \\ \log\left(e^a + e^b\right) \simeq \max\left\{a, b\right\}, & \text{otherwise}\end{cases}$$

B.1.3 Kernel Two-Sample Test

We used KTST also to test whether group-specific clusters identified in GNBMM were drawn from different distributions (see section B.1.2).

B.1.4 Kernel Fitness to Data

In order to be able to use our choice of kernel function (see section 4.2.6.2 in main text) as a distance metric for other methods (e.g., KTST and spectral clustering), we validated that it made a meaningful distance metric on our data. We tested the fitness of the kernel to the data by validating that $\widehat{MMD}_u^2\left[\mathcal{F}, X_1, X_2\right]$ for random splits $X = \{X_1 \bigcup X_2\}$, where X is the data from a select site, was drawn from a uniform distribution. For each of the eight select sites, we generated $\widehat{MMD}_u^2$ for 1000 random splits and used the Kolmogorov-Smirnov test [165] (as implemented in scipy (scipy.stats.kstest), version 1.2.1) to validate a uniform distribution of these values. We ran the test for TD and ASD samples separately and combined (the entire site's sample) per site. For all samples, the results indicated uniform distributions (p $\leq$ 0.09; Table B.2).

B.2 Additional Results

In addition to the experiments mentioned in the main text, we examined intra-single-site GNBMM in comparison to intra-single-site GNB for the biggest site in the repository (NYU; section B.2.2).

B.2.1 Distance Between TD and ASD Samples Distributions

We computed the MMD-based distance between the distributions of the TD and ASD samples between each of the eight select sites and the combined samples from all other select sites (Table B.6). For example, for USM and Yale, $\widehat{MMD}_u^2\left[\mathcal{F}, TD_{Site}, TD_{Ref}\right] < \widehat{MMD}_u^2\left[\mathcal{F}, TD_{Site}, ASD_{Ref}\right]$ and $\widehat{MMD}_u^2\left[\mathcal{F}, ASD_{Site}, ASD_{Ref}\right] < \widehat{MMD}_u^2\left[\mathcal{F}, ASD_{Site}, TD_{Ref}\right]$, where Ref is the combined samples from all other select sites. Conversely, for Trinity and Leuven, $\widehat{MMD}_u^2\left[\mathcal{F}, TD_{Site}, ASD_{Ref}\right] < \widehat{MMD}_u^2\left[\mathcal{F}, TD_{Site}, TD_{Ref}\right]$.

B.2.2 Intra-Single-Site GNBMM Performance

To test the hypothesis that the eight select sites, excluding Pitt, had heterogeneous samples drawn from multiple distributions, we used GNBMM models with two sub-populations per class. As the sample sizes per site were fairly small and clustering would shrink them further, we focused on the largest site, NYU. Similarly to previous experiments, we performed stratified 10-fold CV. We validated in every fold that the identified sub-populations per (TD/ASD) group were drawn from different distributions (KTST $p < 0.01$).

The mean F_1-score(ASD) using GNBMM (0.36; SD = 0.12) was significantly (two-tailed Welch's t-test $p = 5.09e-04$) lower than the mean using GNB (0.57; SD = 0.09; Table 4.4.A). We examined the sub-populations per fold and found that in every fold the spectral clustering resulted in the majority of the observations per class assigned to clusters that had worse group-separation as compared to using all observations per group. That is, for sub-populations $\ell_{TD,1}, \ell_{TD,2}$ of TD and sub-populations $\ell_{ASD,1}, \ell_{ASD,2}$ of ASD, $\widehat{MMD}_u^2\left[\mathcal{F}, \ell_{TD,1}, \ell_{ASD,1}\right] < \widehat{MMD}_u^2\left[\mathcal{F}, TD, ASD\right]$ (KTST $p < 0.01$ for both), and $\left|\ell_{TD,1}\right| \geq \left|\ell_{TD,2}\right|$ and $\left|\ell_{ASD,1}\right| \geq \left|\ell_{ASD,2}\right|$.

Table B.1. Participants

Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
S1	0050954*	0050004*	0050232*	*S1*	*S1*	0050433	0050551
0050682	0050955*	0050005*	0050233*	0051205*	0050272*	0050434	0050552
0050683	0050956*	0050006*	0050234*	0051206*	0050273*	0050435	0050553
0050685	0050957*	0050007*	0050235*	0051208*	0050274*	0050436	0050554
0050686*	0050958*	0050008*	0050236*	0051210*	0050275*	0050437	0050555
0050687	0050959*	0050009*	0050237*	0051211*	0050276*	0050438	0050556
0050688	0050960*	0050010*	0050239*	0051212*	0050277*	0050439	0050557
0050689*	0050961*	0050011*	0050240*	0051214*	0050278*	0050440	0050558
0050690*	0050962*	0050012*	0050241*	0051216*	0050280*	0050441	0050559
0050691	0050964*	0050014*	0050243*	0051217*	0050282*	0050442	0050560
0050692	0050965*	0050015*	0050245*	0051218*	0050283*	0050443	0050561
0050693*	0050966*	0050016*	0050246*	0051220*	0050284*	0050444	0050562
0050694*	0050967*	0050020*	0050247*	0051221*	0050285*	0050445	0050563
0050695*	0050968*	0050022*	0050248*	0051222*	0050287*	0050446	0050564
0050696*	0050969*	0050023*	0050249*	0051223*	0050289*	0050447	0050566
0050697*	0050970*	0050024*	0050250*	0051224*	0050290*	0050448	0050567
0050698	0050972*	0050025*	0050251*	0051225*	0050291*	0050449	0050568
0050699	0050973*	0050027*	0050252*	0051226*	0050292*	0050455	0050569
0050700*	0050974*	0050029*	0050253*	0051229*	0050293*	0050463	0050570
0050701	0050976*	0050030	0050254*	0051230*	0050294*	0050466	0050571
0050702*	0050977*	0050031	0050255*	0051234*	0050295*	0050467	0050572
0050703	0050978*	0050032	0050257	0051235*	0050297*	0050468	0050573
0050704*	0050979*	0050033	0050259	0051236*	0050298*	0050469	0050574
0050705*	0050981*	0050034	0050260	0051237*	0050300*	0050477*	0050575
0050706	0050982*	0050035	0050261	0051239*	0050301*	0050480*	0050576
0050707	0050983*	0050036	0050262	0051240*	0050302*	0050482*	0050577
0050708*	0050984*	0050037	0050263	0051241*	0050307*	0050483*	0050601*
0050709	0050985*	0050038	0050264	0051248*	0050310*	0050485*	0050603*
0050710	0050986*	0050040	0050266	0051249*	0050314*	0050486*	0050604*
0050711*	0050987*	0050041	0050267	0051250	0050315*	0050487*	0050605*
S2	0050988*	0050042	0050268	0051251	0050318*	0050488*	0050606*
0050722	0050989*	0050044	0050269	0051252	0050319*	0050490*	0050607*
0050723	0050990*	0050045	0050270	0051253	0050320*	0050491*	0050609*

Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
0050724	0050991*	0050046	0050271	0051254	0050321*	0050492*	0050611*
0050725	0050992*	0050047	0051132	0051255	0050324*	0050493*	0050612*
0050726	0050993*	0050048	0051133	0051256	0050325*	0050494*	0050613*
0050727	0050994*	0050050	0051134	0051257	0050326*	0050496*	0050614*
0050728	0050995*	0050051	0051135	0051260	0050327	0050497*	0050616*
0050730	0050996*	0050052	0051137	0051261	0050329	0050498*	0050617*
0050731	0050997*	0050055*	0051138	0051262	0050330	0050499*	0050619*
0050732	0050998*	0050056*	0051139	0051264	0050331	0050500*	0050620*
0050733	0050999*	0050057*	0051140	0051265	0050332	0050502*	0050621*
0050735	0051000*	0050058	0051141	0051266	0050334	0050503*	0050623*
0050736	0051001*	0050059	0051142	0051267	0050335	0050504*	0050624*
0050737	0051002*	0050060		0051268	0050336	0050505*	0050625*
0050738	0051003*			0051269	0050337	0050507*	0050626*
0050739	0051006*			0051271	0050338	0050514*	0050627*
0050740	0051007*			0051272	0050339	0050515*	0050628*
0050741	0051008*			0051273	0050340	0050516*	
0050742	0051009*			0051275	0050341	0050518*	
0050743*	0051010*			0051276	0050342	0050520*	
0050745*	0051011*			0051277	0050343	0050521*	
0050746*	0051012*			0051279	0050344	0050524*	
0050747*	0051013*			0051280	0050345	0050525*	
0050748*	0051014*			0051281	0050346	0050526*	
0050749*	0051015*			0051282	0050347	0050527*	
0050751*	0051016*			*S2*	0050348	0050528*	
0050752*	0051017*			0051291*	0050349	0050529*	
0050753*	0051018*			0051293*	0050350	0050530*	
0050754*	0051019*			0051294*	0050351	0050532*	
0050755*	0051020*			0051295*	0050352		
0050756*	0051021*			0051298*	0050353		
0050757*	0051023*			0051301*	0050355		
	0051024*			0051302*	0050356		
	0051025*			0051303	0050357		
	0051026*			0051305	0050358		
	0051027*			0051306	0050360		

Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
	0051028*			0051307	0050361		
	0051029*			0051308	0050362		
	0051032*			0051309	0050363		
	0051033*			0051311	0050364		
	0051034*			0051312	0050365		
	0051035*			0051313	0050366		
	0051036			0051314	0050367		
	0051038			0051315	0050368		
	0051039			0051316	0050369		
	0051040			0051317*	0050370		
	0051041				0050372		
	0051042				0050373		
	0051044				0050374		
	0051045				0050375		
	0051046				0050377		
	0051047				0050379		
	0051048				*S2*		
	0051049				0050382		
	0051050				0050385		
	0051051				0050386		
	0051052				0050387		
	0051053				0050388		
	0051054				0050390		
	0051055				0050391		
	0051056				0050397*		
	0051057				0050399*		
	0051058				0050404*		
	0051059				0050405*		
	0051060				0050406*		
	0051061				0050407*		
	0051062				0050408*		
	0051063				0050410*		
	0051064				0050411*		
	0051065				0050412*		

Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
	0051066				0050413*		
	0051067				0050414		
	0051068				0050415		
	0051069				0050416		
	0051070				0050417		
	0051071				0050418		
	0051072				0050419		
	0051073				0050421		
	0051074				0050424		
	0051075				0050425		
	0051076				0050426		
	0051077				0050427		
	0051078				0050428		
	0051079						
	0051080						
	0051081						
	0051082						
	0051083						
	0051084						
	0051085						
	0051086						
	0051087						
	0051088						
	0051089						
	0051090						
	0051091						
	0051093						
	0051094						
	0051095						
	0051096						
	0051097						
	0051098						
	0051099						
	0051100						

Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
	0051101						
	0051102						
	0051103						
	0051104						
	0051105						
	0051106						
	0051107						
	0051109						
	0051110						
	0051111						
	0051112						
	0051113						
	0051114						
	0051116						
	0051117						
	0051118						
	0051121						
	0051122						
	0051123						
	0051124						
	0051126						
	0051128						
	0051129						
	0051130						
	0051131						
	0051146						
	0051147						
	0051148						
	0051149						
	0051150						
	0051151						
	0051152						
	0051153						
	0051154						

Leuven	NYU	Pitt	Trinity	UCLA	UM	USM	Yale
	0051155						
	0051156						

List of the participant IDs per select site that were used in the present work. *ASD participants (the rest are TD participants). For sites with multiple samples (Leuven, UCLA, and UM), participants from samples 1 and 2 are listed under *S1* and *S2*, respectively. TD = Typically Developing; ASD = Autism Spectrum Disorder.

Table B.2. Kernel fitness test for each of the eight select sites

Site	TD Sample	ASD Sample	TD+ASD Sample
Leuven	0.13	0.36	0.56
NYU	0.19	0.56	0.51
Pitt	0.19	0.56	0.51
Trinity	0.61	0.72	0.51
UCLA	0.77	0.61	0.96
UM	0.61	0.82	0.15
USM	0.82	0.99	0.93
Yale	0.09	0.56	0.61

Kolmogorov-Smirnov test's p-value for uniform distribution of $\widehat{MMD}_u^2[\mathcal{F}, X_1, X_2]$ for random splits X_1 and X_2 (s.t. $X = \{X_1 \bigcup X_2\}$) based on 1000 random splits per sample. TD = Typically Developing; ASD = Autism Spectrum Disorder; MMD = Maximum Mean Discrepancy.

Table B.3. Number of participants in each of the nine non-select sites

	TD	ASD	Total
Caltech	18	19	37
CMU	2	3	5
KKI	27	12	39
MaxMun	24	18	42
OHSU	11	12	23
Olin	11	14	25
SBL	12	14	26
SDSU	21	12	33
Stanford	19	17	36
Total	145	121	266

ASD = Autism Spectrum Disorder; SD = Standard Deviation.

Table B.4. Performance for random guessing on the select sites

	Leuven	NYU	Pitt	Trinity	UCLA
F_1(ASD)	0.47 (0.02)	0.44 (0.01)	0.50 (0.02)	0.49 (0.02)	0.49 (0.02)
	UM	**USM**	**Yale**		
F_1(ASD)	0.44 (0.01)	0.58 (0.03)	0.48 (0.02)		

Mean Laplace-Corrected F_1-score(ASD) (SD) for 10 stratified folds with classes assigned at random (random guessing) on the eight select sites. ASD = Autism Spectrum Disorder; SD = Standard Deviation.

Table B.5. Performance on select sites per fold for different setups

		Inter-Multi-Site	
Site	**Intra-Single-Site**	**no TST**	**TST**
Leuven	0.73	0.25	0.25
	0.67	0.50	0.50
	0.55	0.50	0.50
	0.55	0.29	0.29
	0.50	0.29	0.29
	0.22	0.67	0.60
	0.73	0.25	0.25
	0.50	0.57	0.57
	0.60	0.33	0.33
	0.50	0.33	0.33
NYU	0.56	0.70	0.64
	0.70	0.73	0.73
	0.64	0.72	0.64
	0.42	0.50	0.42
	0.71	0.71	0.71
	0.55	0.64	0.55
	0.53	0.56	0.50
	0.53	0.63	0.59
	0.44	0.58	0.57
	0.60	0.64	0.67
Pitt	0.73	0.67	0.67
	0.67	0.29	0.29
	0.50	0.57	0.57
	0.75	0.33	0.33
	0.60	0.50	0.50
	0.57	0.57	0.57
	0.75	0.57	0.57
	0.75	0.67	0.67
	0.75	0.33	0.57
	0.60	0.57	0.57

		Inter-Multi-Site	
Site	**Intra-Single-Site**	**no TST**	**TST**
Trinity	0.50	0.60	0.60
	0.50	0.67	0.67
	0.67	0.60	0.60
	0.33	0.75	0.75
	0.50	0.57	0.57
	0.44	0.60	0.60
	0.57	0.67	0.67
	0.75	0.25	0.25
	0.33	0.33	0.33
	0.57	0.57	0.57
UCLA	0.62	0.46	0.57
	0.60	0.60	0.73
	0.22	0.62	0.55
	0.77	0.77	0.77
	0.33	0.67	0.57
	0.44	0.73	0.73
	0.67	0.67	0.67
	0.60	0.60	0.60
	0.67	0.67	0.67
	0.60	0.67	0.67
UM	0.77	0.71	0.71
	0.57	0.63	0.67
	0.57	0.63	0.63
	0.36	0.60	0.67
	0.50	0.57	0.57
	0.50	0.75	0.67
	0.67	0.67	0.75
	0.60	0.59	0.63
	0.43	0.59	0.59
	0.36	0.56	0.57
USM	0.71	0.71	0.71

		Inter-Multi-Site	
Site	**Intra-Single-Site**	**no TST**	**TST**
	0.71	0.73	0.83
	0.77	0.67	0.62
	0.71	0.83	0.83
	0.77	0.83	0.83
	0.71	0.77	0.77
	0.60	0.73	0.73
	0.44	0.60	0.55
	0.67	0.73	0.73
	0.67	0.67	0.67
Yale	0.25	0.80	0.80
	0.50	0.50	0.50
	0.67	0.50	0.50
	0.75	0.67	0.67
	0.29	0.75	0.75
	0.57	0.75	0.75
	0.57	0.67	0.67
	0.50	0.67	0.67
	0.33	0.33	0.33
	0.50	0.50	0.50

F_1-score(ASD) per fold for the eight select sites (10 folds per site). Intra-Single-Site: train on the other nine folds. Inter-Multi-Site no TST: train only on the seven other sites. Inter-Multi-Site TST: train on the seven other sites and on the other nine folds in the target site). TST = target site training; ASD = Autism Spectrum Disorder.

Table B.6. Distances between distributions of TD and ASD observations between each of the select sites and multi-site samples

	$\widehat{\mathbf{MMD}}^2_{\mathbf{u}}[\mathcal{F}, \mathbf{TD_{Site}}, \mathbf{X}]$		$\widehat{\mathbf{MMD}}^2_{\mathbf{u}}[\mathcal{F}, \mathbf{ASD_{Site}}, \mathbf{X}]$	
Site	$\mathbf{X = TD_{Ref}}$	$\mathbf{X = ASD_{Ref}}$	$\mathbf{X = TD_{Ref}}$	$\mathbf{X = ASD_{Ref}}$
Leuven	0.0273	0.0255	0.0451	0.0343
NYU	0.0165	0.0217	0.0246	0.0165
Pitt	0.0236	0.0218	0.0345	0.0198
Trinity	0.0405	0.0345	0.0406	0.0273
UCLA	0.0224	0.0184	0.0295	0.0149
UM	0.0391	0.0597	0.0351	0.0422
USM	0.0312	0.0371	0.0359	0.0256
Yale	0.0171	0.0314	0.0220	0.0213

Estimated MMD between distributions of a site's (rows) TD and ASD observations and the distributions of TD and ASD observations (columns) in the site's respective reference samples (all of the other select sites combined). TD = Typically Developing; ASD = Autism Spectrum Disorder; MMD = Maximum Mean Discrepancy; Ref = Reference.

Appendix C
PATENet - Example Application to Longitudinal fMRI: Pairwise Alignment in Adolescents and Young Adults

In the present appendix we describe an example application of PATENet (see section 5.2) to real-world longitudinal FC data.

C.1 Introduction

Studies have suggested that the human brain undergoes greater development, accompanied by more changes in FC, during adolescence as compare to during young adulthood (e.g., [62,166–168]). Therefore, we hypothesized that the similarity between longitudinal FC data from pairs of participants would be greater in young adults as compared to the similarity in adolescents. We used PATENet and its OSN alignment score to test this hypothesis.

C.2 Materials and Methods

C.2.1 Participants

In the present work we included a subset of the participants used in [169], selecting only participants that had 3 or more recordings during adolescence ($\leq$ 18.0 years-old) or during young adulthood ($\geq$ 20.0 years-old). We ended up with 19 adolescents ($3-5$ recordings each; mean = 4.16, SD = 0.69) and 23 young adults ($3-8$ recordings each; mean = 4.74, SD = 1.51). The Mean number of recordings per participant was comparable between the two groups (two-tailed Welch's t-test p = 0.11). Written informed consent was obtained from every subject, and all participants were compensated for their time. All experiments complied with the Code of Ethics of the World Medical Association (1996 Declaration of Helsinki) and were approved by the Institutional Review Board at the University of Pittsburgh.

C.2.2 fMRI Acquisition, Preprocessing, and FC

Acquisition and preprocessing details are described elsewhere (see [169]). The data used in the present work were intrinsic FC based on resting-state fMRI extracted from task-based fMRI [170].

C.2.2.1 FC Networks Construction

FC networks in the present work had 267 nodes corresponding to 267 functional 10 mm diameter spheres centered around the 264 coordinates from the Power brain atlas [17] and three additional MNI coordinates: $[22.8, -2.2, -19]$ (right amygdala), $[23.4, -3.4, -18.5]$ (left amygdala), and $[2, 24.5, -13]$). The time series of all voxels within each ROI were averaged, and then Pearson's correlations between the averages were used to produce a 267×267 correlation matrix. Any comparisons made between correlations were transformed to z values using Fisher z(r) transformation, and then reconverted to Pearson's r values.

C.2.3 Participation Coefficient

Development during adolescence is accompanied by changed in FC between modules rather than formation of modules, suggesting participation coefficient (PC) would be a good topological feature to capture these changes [167]. PC is a measure of the diversity of a node's inter-modular connectivity, and is defined by:

$$PC\left(v_i\right) = \sum_{k=1}^{m}\left[\left(\frac{\sum_{j\in M_k} w_{i,j}}{d\left(v_i\right)}\right)^2\right]$$

where v_i is the i^{th} node, m is the number of modules, M_k is the k^{th} module, $w_{i,j}$ is the weight of the edge (strength of the connection) between nodes v_i and v_j, and $d\left(v_i\right)$ is the weighted degree (sum of the weights of all immediate edges/connections) of node v_i.

The findings of [167] focused on five modules, which correspond to seven modules (and 181 ROIs) in the Power brain atlas [17]:

1. Sensory/somatomotor hand/mouth module (SM; 35 ROIs)
2. Cingulo-opercular/Salience module (CO/S; 32 ROIs)
3. Default mode module (DM; 58 ROIs)
4. Visual module (V; 31 ROIs)
5. Fronto-parietal module (FP; 25 ROIs)

In the present work, we computed PC for each of these 181 ROIs in the context of these five modules.

C.2.4 Head Motion Measurements

Three head motion metrics per recording were made available with the data: volume-to-volume framewise displacement (FD), the root mean square derivative of fMRI time series (DVARS), and the proportion of censored volumes when connectivity is computed (the percentage of volumes with FD > 0.3 mm or DVARS > 20; pctCens).

Using FD and DVARS, the mean head motion was significantly (two-tailed Welch's t-test $p < 0.05$ for FD and $p < 0.001$ for DVARS) greater in adolescents as compared to young adults (Table C.1). Using pctCens, the means were comparable (two-tailed Welch's t-test $p = 0.0816$) between the groups.

As head motion is known to affect FC, and considering that head motion was significantly greater in adolescents as compared to young adults (for two of the three metrics), head motion needed to be taken into account, to make sure any observed differences in similarity between the age-groups were not due to difference in head motion between the age-groups. Therefore, we used FD in the remainder of the present work, as it is a commonly-used measure of head motion and was significantly different between the age-groups. We computed participant mean FD (PMFD; the mean FD of the participant's recordings that were included in the experiments) and alignment mean FD (AMFD; the mean between the two PMDFs of the pair of participants being aligned with each other).

Table C.1. Mean head motion (SD) for the two age-groups, using three different metrics

Metric	Adolescents	Young adults
FD	0.2484^{*} (0.1281)	0.1814 (0.0419)
DVARS	$11.6709^{*} *$ (0.2918)	11.9735 (0.2085)
pctCens	0.1162 (0.0780)	0.0764 (0.0628)

Mean head motion (SD) for the two age-groups, using the three different metrics. *Mean significantly (two-tailed Welch's t-test $p < 0.05$) greater than mean in young adults; $^{*}*$Mean significantly (two-tailed Welch's t-test $p < 0.0001$) greater than mean in young adults. FD = volume-to-volume framewise displacement; DVARS = root mean square derivative of fMRI time series; pctCens = the proportion of censored volumes when connectivity is computed (the percentage of volumes with FD > 0.3 mm or DVARS > 20; SD = Standard Deviation.

C.2.5 Behavioral Measurements

13 behavioral measurements from six tasks were made available to us with the data. Four of the six tasks are part of the Cambridge Neuropsychological Test Automated Battery (CANTAB) [171]. The measurements included mean latency and mean error in CANTAB's motor screening test (MOT); %correct and mean correct latency in the CANTAB's delayed matched to sample (DMS) test; span length, total number

of errors, and mean time to first response in the CANTAB's spatial span (SSP) test; number of problems solved in minimum moves and initial response time in CANTAB's Stockings of Cambridge (SOC) test; number of correct trials and latency in correct trials in an anti-saccade task; and mean latency and precision in a memory-guided saccade (MGS) task.

C.2.6 Data Representation

Each FC network was soft-thresholded [172] to generate a non-negative adjacency matrix: $w' = \left(\frac{w+1}{2}\right)^{12}$, where w is the edge weight in the FC network, and $\forall w, w > -1$. Each participant was represented using an OSN of their respective FC networks in chronological order.

C.2.7 Pair-Wise Participant Alignment

In the present work, we used PATENet with match threshold $\varphi = 0.65$ (see section C.2.7.1), DeltaCon [155] as the well-defined UNNSM, and a logarithmic signed normalized monotonically increasing transform: $\ell(x) = 1 - \log_\alpha\left[\alpha^2 + \left(1 - \alpha^2\right) \cdot x\right]$, where $\alpha = \frac{\varphi}{1-\varphi}$.

C.2.7.1 Match Threshold

In order to determine the match threshold to use in PATENet, we computed for each of the age-groups DeltaCon [155] similarity scores between layers between and within participants (Table C.2). We chose to use $\varphi = 0.65$ as the match threshold, as it was the mean similarity between participants in each of the age-groups, rounded to two decimal places. It is worth noting that the mean similarity between adolescents was significantly (two-tailed Welch's t-test $p < 0.05$) lower than the mean similarity between young adults, as well as the mean similarity within adolescents and the mean similarity within young adults).

Table C.2. Mean DeltaCon similarity score (SD) between and within participants from the two age-groups

	Adolescents	**Young adults**
Between participants	0.6480 (0.0184)	0.6516^{*} (0.0208)
Within participants	0.6755^{*} (0.0228)	0.6819^{*} (0.0287)

Mean DeltaCon similarity score (SD) between and within participants from the two age-groups. *Mean significantly (two-tailed Welch's t-test $p < 0.05$) greater than mean similarity between adolescent participants. SD = Standard Deviation.

C.2.8 Statistical Analyses

We used two-tailed Welch's t-test to test for significant difference in means, and we used independence and conditional independence tests to examine the dependencies between OSN alignment scores, head motion, and age-groups. E.g., whether OSN alignment scores dependence on age-group was conditioned on AMFD. Specifically, we used HSIC [98] and SDCIT [99] to assess unconditional independence and conditional independence, respectively. We computed a mean p-value for these tests, based on 100 runs with different random seeds. For all tests, we assumed significance for $p < 0.01$.

C.3 Results and Discussion

First, we examined relationships between the different variables. Mean OSN alignment score was significantly (two-tailed Welch's t-test $p = 3.16e-18$) greater for young adults as compared to adolescents, while mean AMFD was significantly (two-tailed Welch's t-test $p = 4.15e-07$) smaller for young adults as compared to adolescents (Table C.3).

HSIC gave comparable results: OSN alignment score and AMFD were dependent (mean $p < 0.0001$ for OSN alignment score; mean $p = 0.0003$ for AMFD) on age-group. Additionally, HSIC revealed that OSN alignment score and AMFD were dependent on each other (mean $p = 0.0008$).

Next, we used SDCIT to examine conditional independence. OSN alignment score was conditionally independent of age-group given AMFD (mean $p = 0.2450$), as well

Table C.3. OSN alignment results for the two age-groups

	OSN alignment score	**#L**	**AMFD**	**#P**
Adolescents	0.10 (0.08)	2.41 (1.01)	0.25 (0.09)	165 (/171)
Young adults	0.16* (0.12)	2.63 (1.14)	0.18* (0.03)	251 (/253)

Mean OSN alignment score (SD), mean number of aligned layers (SD), and mean FD (SD) between pairs of participants within each age-group, and number of pairs (/maximum possible) that were aligned within each age-group. *Mean significantly (two-tailed Welch's t-test $p = 3.16e-18$ and $p = 4.15e-07$ for OSN alignment score and AMFD, respectively) different from respective mean in adolescents. OSN = ordered sequences of networks; #L = number of aligned layers; AMFD = alignment mean FD; FD = volume-to-volume framewise displacement; #P = number of pairs of OSNs with at least one layer aligned; SD = standard deviation

as conditionally independent of AMFD given age-group (mean $p = 0.0607$; could be due to low SP, as $SD = 0.1293$). However, AMFD remained dependent on age-group even when conditioned on OSN alignment score (mean $p < 0.0001$).

If the results above are reliable, rather than due to low SP, then:

1. $\mathcal{M} \not\perp\!\!\!\perp \mathcal{G}$
2. $\mathcal{S} \not\perp\!\!\!\perp \mathcal{G}$
3. $\mathcal{M} \not\perp\!\!\!\perp \mathcal{S}$
4. $[\mathcal{M} \not\perp\!\!\!\perp \mathcal{G}] \mid \mathcal{S}$
5. $[\mathcal{S} \perp\!\!\!\perp \mathcal{G}] \mid \mathcal{M}$
6. $[\mathcal{M} \perp\!\!\!\perp \mathcal{S}] \mid \mathcal{G}$

where $\mathcal{M}$ is AMFD, $\mathcal{S}$ is OSN alignment score, and $\mathcal{G}$ is age-group. However, no causal DAG with only three elements can satisfy all of the above simultaneously (Fig. C.1). Therefore, the difference in AMFD alone does not cause the difference in OSN alignment score between the two age-groups; there has to be at least one unmeasured confounding variable to satisfy these relationships simultaneously.

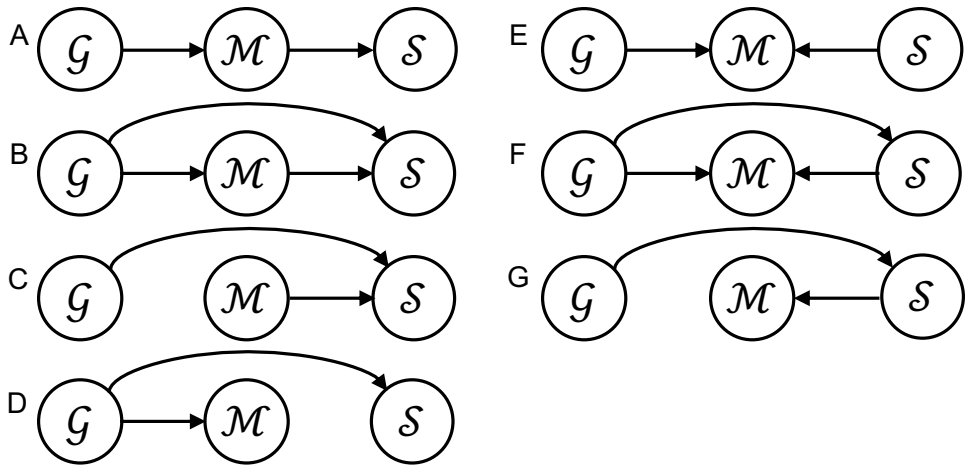

Figure C.1. All possible causal DAGs with three elements. Each of these DAGs violates at least one of the conditional/unconditional dependence/independence suggested by the data. (A) violates 6. (B) violates 5 and 6. (C) violates 4 and 5. (D) violates 5. (E) violates 6. (F) violates 5 and 6. (G) violates 4, 5, and 6. $\mathcal{M}$ = AMFD; $\mathcal{S}$ = OSN alignment score; $\mathcal{G}$ = age-group; AMFD = alignment mean FD; FD = volume-to-volume framewise displacement; OSN = ordered sequences of networks; DAG = directed acyclic graph.

C.3.1 Behavioral Measurements to Identify Confounding Variables Candidates

We went on to examine the behavioral measurements available to us (see section C.2.5) to try to identify possible confounding variables that could be added to the causal DAG. Of the 13 measurements, mean correct latency in DMS and mean time to first response in SSP were the most strongly associated with age (also within participants and within each sex separately) in the records used in the present analyses, as well as in records that were excluded (Table C.4).

Figure C.2 depicts one potential causal DAG that satisfies all the relationships deduced from the present analyses. We used this architecture to guide our search for confounding variables (see criteria in section C.3.2).

Table C.4. Conditional and unconditional independence tests between select behavioral measurements and age

Task	Sample	$(\mathbf{X}, \mathbf{Age})$	$(\mathbf{X}, \mathbf{Age}) \mid \mathbf{ID}$	$(\mathbf{X}, \mathbf{Age}) \mid \mathbf{Sex}$
DMS	**All**	$p < 0.0001^*$	$p = 0.0001^*$	$p < 0.0001^*$
DMS	**Select**	$p < 0.0001^*$	$p = 0.0037^*$	$p = 0.0126^\dagger$
SSP	**All**	$p < 0.0001^*$	$p < 0.0001^*$	$p < 0.0001^*$
SSP	**Select**	$p < 0.0001^*$	$p < 0.0001^*$	$p < 0.0001^*$

p-value of HSIC between age and X (column '(X, Age)') and SDCIT between age and X given participant ID (column '$(X, Age) \mid ID$') and given participant's sex (column '$(X, Age) \mid Sex$'), where X is the select behavioral measurement (mean correct latency in DMS task or mean time to first response in SSP task). In 'Sample' column, 'Select' refers to the records included in the present analyses, while 'All' refers to all the records that were made available to us. *Significantly ($p < 0.01$) dependent; †Marginally significant dependence. DMS = delayed matched to sample; SSP = spatial span; ID = participant ID; AMFD = alignment mean FD; FD = volume-to-volume framewise displacement; HSIC = Hilbert-Schmidt Independence Criterion; SDCIT = Self-Discrepancy Conditional Independence Test.

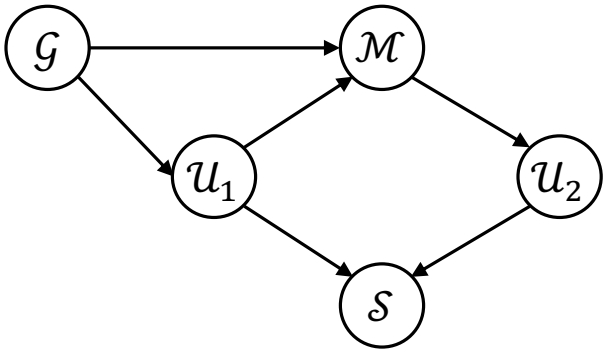

Figure C.2. Potential causal DAG that satisfies all of the conditional/unconditional dependence/independence suggested by our results. $\mathcal{M}$ = AMFD; $\mathcal{S}$ = OSN alignment score; $\mathcal{G}$ = age group; AMFD = alignment mean FD; FD = volume-to-volume framewise displacement; OSN = ordered sequences of networks; $\mathcal{U}_1$ = unobserved confounding variable; $\mathcal{U}_2$ = unobserved confounding variable; DAG = directed acyclic graph.

C.3.2 Participation Coefficients Correlated with Behavioral Measurement

We first examined the mean correct latency in DMS, to check if any node v_i within the 181 nodes with PC values (see section C.2.3) satisfied all of the following requirements:

$$\begin{cases} PC\left(v_i\right) \not\perp\!\!\!\perp \ell \\ PC\left(v_i\right) \not\perp\!\!\!\perp \mathcal{G} \\ \left[\ell \perp\!\!\!\perp \mathcal{G}\right] \mid PC\left(v_i\right) \end{cases}$$

where $PC\left(v_i\right)$ is PC of node v_i, ℓ is mean correct latency in DMS, and $\mathcal{G}$ is age-group. We used HSIC to evaluate the unconditional dependencies and used SDCIT to evaluate the conditional independence.

Five ROIs satisfied all three relationships: v_{20} (AAL: left inferior Parietal gyrus; SM module), v_{31} (AAL: right supplementary motor srea; SM module), v_{37} (AAL: undefined; SM module), v_{49} (AAL: right superior frontal gyrus, dorsolateral; CO/S module), and v_{105} (AAL: right superior frontal gyrus, medial; DM module).

Next, we computed for each of these five ROIs the derivatives of normalized ℓ (mean correct latency in DMS) with respect to normalized $PC\left(v_i\right)$: $\rho\left(v_i\right) = \frac{\partial \ell}{\partial PC(v_i)}$, and tried to characterize how $\rho\left(v_i\right)$ changed with age. $\rho\left(v_i\right)$ was computed within-participant for every two layers of the respective participant that were aligned with at least one other participant's layer. Consequently, there were 48 $\rho\left(v_i\right)$ values from adolescents and 74 $\rho\left(v_i\right)$ values from young adults (with outliers in both age-groups). Our attempts did not seem to yield any coherent results (omitted), possibly in part due to the small sample size. Similar analysis for mean time to first response in SSP yielded similar results, halting the present work.

C.4 Concluding Remarks

In the present work we found that similarity between evolving FC (OSN alignment score) from pairs of participants was significantly lower between adolescents as compared to between young adults. However, head motion (AMFD) was significantly greater in adolescents as compared to young adults, and was strongly negatively

correlated with OSN alignment score. Therefore, we used advanced techniques to estimate whether the observed relationship between AMFD and OSN alignment score was merely a correlation or a causal relationship. Our results suggest that the difference in AMFD between the two age-groups did not cause (at least not by itself) the difference in OSN alignment score between the two age-groups. Furthermore, our results suggest that there had to have been at least one unmeasured confounding variable to have a complete causal DAG that connects AMFD, OSN alignment score, and age-group. However, we were not able to identify a potential confounding variable. We tried to use behavioral measurements to identify confounding variables, but our attempts were unsuccessful, possibly in part due to the small sample size. Perhaps in the future, larger samples will become available and will have enough SP to yield more conclusive results.

Bibliography

[1] Scanziani, M. and Häusser, M. (2009). "Electrophysiology in the Age of Light," *Nature*, **461**(7266), pp. 930–939.

[2] Baars, B. J. and Gage, N. M. (2010). *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*, Academic Press.

[3] Haas, L. F. (2003). "Hans Berger (1873–1941), Richard Caton (1842–1926), and Electroencephalography," *Journal of Neurology, Neurosurgery & Psychiatry*, **74**(1), pp. 9–9.

[4] Cohen, D. (1968). "Magnetoencephalography: Evidence of Magnetic Fields Produced by Alpha-Rhythm Currents," *Science*, **161**(3843), pp. 784–786.

[5] Hounsfield, G. N. (1973). "Computerized Transverse Axial Scanning (Tomography): Part I. Description of System," *British Journal of Radiology*, **46**, pp. 1016–1022.

[6] Gratton, G., Corballis, P. M., Cho, E., Fabiani, M., and Hood, D. C. (1995). "Shades of Gray Matter: Noninvasive Optical Images of Human Brain Responses During Visual Stimulation," *Psychophysiology*, **32**(5), pp. 505–509.

[7] Belliveau, J. W., Kennedy, D. N., McKinstry, R. C., Buchbinder, B. R., Weisskoff, R., Cohen, M. S., Vevea, J. M., Brady, T. J., and Rosen, B. R. (1991). "Functional Mapping of the Human Visual Cortex by Magnetic Resonance Imaging," *Science*, **254**(5032), pp. 716–719.

[8] Chang, E. F. (2015). "Towards Large-Scale, Human-Based, Mesoscopic Neurotechnologies," *Neuron*, **86**(1), pp. 68–78.

[9] Craddock, R. C., Jbabdi, S., Yan, C. G., Vogelstein, J. T., Castellanos, F. X., Di-Martino, A., Kelly, C., Heberlein, K., Colcombe, S., and Milham, M. P. (2013). "Imaging Human Connectomes at the Macroscale," *Nature Methods*, **10**(6), pp. 524–539.

[10] Sporns, O. (2013). "Structure and Function of Complex Brain Networks," *Dialogues in Clinical Neuroscience*, **15**(3), p. 247.

[11] Calhoun, V. D., Miller, R., Pearlson, G., and Adali, T. (2014). "The Chronnectome: Time-Varying Connectivity Networks as the Next Frontier in fMRI Data Discovery," *Neuron*, **84**(2), pp. 262–274.

[12] Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., and Barnes, K. A. (2010). "Prediction of Individual Brain Maturity using fMRI," *Science*, **329**(5997), pp. 1358–1361.

[13] Betzel, R. F., Byrge, L., He, Y., Goñi, J., Zuo, X. N., and Sporns, O. (2014). "Changes in Structural and Functional Connectivity Among Resting-State Networks Across the Human Lifespan," *Neuroimage*, **102**, pp. 345–357.

[14] Stevens, M. C. (2016). "The Contributions of Resting State and Task-Based Functional Connectivity Studies to Our Understanding of Adolescent Brain Network Maturation," *Neuroscience & Biobehavioral Reviews*, **70**, pp. 13–32.

[15] Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., and Uğurbil, K. (2010). "Multiband Multislice GE-EPI at 7 Tesla, with 16-Fold Acceleration using Partial Parallel Imaging with Application to High Spatial and Temporal Whole-Brain fMRI," *Magnetic Resonance in Medicine*, **63**(5), pp. 1144–1153.

[16] Craddock, R. C., James, G. A., Holtzheimer III, P. E., Hu, X. P., and Mayberg, H. S. (2012). "A Whole Brain fMRI Atlas Generated via Spatially Constrained Spectral Clustering," *Human Brain Mapping*, **33**(8), pp. 1914–1928.

[17] Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., and Schlaggar, S. E., B. L. Petersen (2011). "Functional Network Organization of the Human Brain," *Neuron*, **72**(4), pp. 665–678.

[18] Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van-Essen, D. C., and Wager, T. D. (2011). "Large-Scale Automated Synthesis of Human Functional Neuroimaging Data," *Nature Methods*, **8**(8), p. 665.

[19] Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B. S., Lewis, J. D., Li, Q., Milham, M., and

Yan, C. (2013). "The Neuro Bureau Preprocessing Initiative: Open Sharing of Preprocessed Neuroimaging Data and Derivatives," *Neuroinformatics*, **4**.

[20] Bassett, D. S. and Sporns, O. (2017). "Network Neuroscience," *Nature Neuroscience*, **20**(3), pp. 353–364.

[21] Alpaydin, E. (2009). *Introduction to Machine Learning*, MIT Press.

[22] Pereira, F., Mitchell, T., and Botvinick, M. (2009). "Machine Learning Classifiers and fMRI: A Tutorial Overview," *Neuroimage*, **45**(1), pp. S199–S209.

[23] Li, K., Guo, L., Nie, J., Li, G., and Liu, T. (2009). "Review of Methods for Functional Brain Connectivity Detection using fMRI," *Computerized Medical Imaging and Graphics*, **33**(2), pp. 131–139.

[24] 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), pp. 519–534.

[25] Sporns, O. (2013). "The Human Connectome: Origins and Challenges," *Neuroimage*, **80**, pp. 53–61.

[26] Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., and Gotlib, I. H. (2015). "Support Vector Machine Classification of Major Depressive Disorder using Diffusion-Weighted Neuroimaging and Graph Theory," *Frontiers in Psychiatry*, **6**, p. 21.

[27] Zuo, X. N., He, Y., Betzel, R. F., Colcombe, S., Sporns, O., and Milham, M. P. (2017). "Human Connectomics Across the Life Span," *Trends in Cognitive Sciences*, **21**(1), pp. 32–45.

[28] Bassett, D. S., Khambhati, A. N., and Grafton, S. T. (2017). "Emerging Frontiers of Neuroengineering: A Network Science of Brain Connectivity," *Annual Review of Biomedical Engineering*, **19**, pp. 327–352.

[29] Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K. R. (2011). "Introduction to Machine Learning for Brain Imaging," *Neuroimage*, **56**(2), pp. 387–399.

[30] Poldrack, R. (2013). "Open fMRI," `https://legacy.openfmri.org/`, [Online; accessed 5-May-2019].

[31] Poldrack, R. "Open Neuro," `https://openneuro.org/`, [Online; accessed 5-May-2019].

[32] Dvornek, N. C., Ventola, P., Pelphrey, K. A., and Duncan, J. S. (2017). "Identifying Autism from Resting-State fMRI using Long Short-Term Memory Networks," in *International Workshop on Machine Learning in Medical Imaging*, pp. 362–370.

[33] Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. (2017). "Distance Metric Learning using Graph Convolutional Networks: Application to Functional Brain Networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 469–477.

[34] Di-Martino, A. and Mostofsky, S. (2012). "Autism Brain Imaging Data Exchange (ABIDE) I," `http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html`, [Online; accessed 5-May-2019].

[35] Craddock, C. and Bellec, P. (2013). "Autism Brain Imaging Data Exchange (ABIDE) Preprocessed," `http://preprocessed-connectomes-project.org/abide/`, [Online; accessed 5-May-2019].

[36] Milham, M. (2011). "ADHD200," `http://fcon_1000.projects.nitrc.org/indi/adhd200/`, [Online; accessed 5-May-2019].

[37] Milham, M. and Zuo, X. N. (2014). "Consortium for Reliability and Reproducibility (CoRR)," `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html`, [Online; accessed 5-May-2019].

[38] Milham, M. and Craddock, C. (2014). "Addiction Connectome Preprocessed Initiative (ACPI)," `http://fcon_1000.projects.nitrc.org/indi/ACPI/html/index.html`, [Online; accessed 5-May-2019].

[39] Milham, M. (2010). "1000 Functional Connectomes Project (FCP)," `http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html`, [Online; accessed 5-May-2019].

[40] Abraham, A., Milham, M. P., Di-Martino, A., Craddock, R. C., Samaras, D., Thirion, B., and Varoquaux, G. (2017). "Deriving Reproducible Biomarkers from Multi-Site Resting-State Data: An Autism-Based Example," *Neuroimage*, **147**, pp. 736–745.

[41] Subbaraju, V., Sundaram, S., and Narasimhan, S. (2018). "Identification of Lateralized Compensatory Neural Activities within the Social Brain due to Autism Spectrum Disorder in Adolescent Males," *European Journal of Neuroscience*, **47**(6), pp. 631–642.

[42] Castellanos, F. X., Di-Martino, A., Craddock, R. C., Mehta, A. D., and Milham, M. P. (2013). "Clinical Applications of the Functional Connectome," *Neuroimage*, **80**, pp. 527–540.

[43] Holme, P. and Saramäki, J. (2012). "Temporal Networks," *Physics Reports*, **519**(3), pp. 97–125.

[44] Holme, P. (2015). "Modern Temporal Network Theory: A Colloquium," *The European Physical Journal B*, **88**(9), pp. 234–263.

[45] Börner, K., Sanyal, S., and Vespignani, A. (2007). "Network Science," *Annual Review of Information Science and Technology*, **41**(1), pp. 537–607.

[46] Newman, M. (2010). *Networks: An Introduction*, Oxford University Press.

[47] Rubinov, M. and Sporns, O. (2010). "Complex Network Measures of Brain Connectivity: Uses and Interpretations," *Neuroimage*, **52**(3), pp. 1059–1069.

[48] Varoquaux, G. (2018). "Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars," *Neuroimage*, **180**, pp. 68–77.

[49] Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. (2013). "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, **14**(5), pp. 365–376.

[50] Van Horn, J. D. and Gazzaniga, M. S. (2002). "Databasing fMRI Studies - Towards A 'Discovery Science' of Brain Function," *Nature Reviews Neuroscience*, **3**(4), p. 314.

[51] Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., and Dogonowski, A. M. (2010). "Toward Discovery Science of Human Brain Function," *Proceedings of the National Academy of Sciences*, **107**(10), pp. 4734–4739.

[52] Milham, M., Craddock, C., Fleischmann, M., Son, J., Clucas, J., Xu, H., Koo, B., Krishnakumar, A., Biswal, B., Castellanos, F., and Colcombe, S. (2017). "Assessment of the Impact of Shared Data on the Scientific Literature," *BioRxiv*, p. 183814.

[53] Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). "Making Data Sharing Work: The FCP/INDI Experience," *Neuroimage*, **82**, pp. 683–691.

[54] Di-Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., and Deen, B. (2014). "The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism," *Molecular Psychiatry*, **19**(6), pp. 659–667.

[55] Milham, M. P. (2012). "Open Neuroscience Solutions for the Connectome-Wide Association Era," *Neuron*, **73**(2), pp. 214–218.

[56] Madhyastha, T., Peverill, M., Koh, N., McCabe, C., Flournoy, J., Mills, K., King, K., Pfeifer, J., and McLaughlin, K. A. (2017). "Current Methods and Limitations for Longitudinal fMRI Analysis Across Development," *Developmental Cognitive Neuroscience*, **33**, pp. 118–128.

[57] Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., Reiss, A. L., and Schatzberg, A. F. (2007). "Resting-State Functional Connectivity in Major Depression: Abnormally Increased Contributions from Subgenual Cingulate Cortex and Thalamus," *Biological Psychiatry*, **62**(5), pp. 429–437.

[58] Meunier, D., Achard, S., Morcom, A., and Bullmore, E. (2009). "Age-Related Changes in Modular Organization of Human Brain Functional Networks," *Neuroimage*, **44**(3), pp. 715–723.

[59] Zhang, H. Y., Wang, S. J., Liu, B., Ma, Z. L., Yang, M., Zhang, Z. J., and Teng, G. J. (2010). "Resting Brain Connectivity: Changes During the Progress of Alzheimer Disease," *Radiology*, **256**(2), pp. 598–606.

[60] Bassett, D. S., Yang, M., Wymbs, N. F., and Grafton, S. T. (2015). "Learning-Induced Autonomy of Sensorimotor Systems," *Developmental Psychology*, **18**(5), pp. 744–751.

[61] Betzel, R. F., Mišić, B., He, Y., Rumschlag, J., Zuo, X. N., and Sporns, O. (2015). "Functional Brain Modules Reconfigure at Multiple Scales Across the Human Lifespan," *arXiv preprint arXiv:1510.08045*.

[62] Ernst, M. (2014). "The Triadic Model Perspective for the Study of Adolescent Motivated Behavior," *Brain and Cognition*, **89**, pp. 104–111.

[63] Diaz, M. T., Johnson, M. A., Burke, D. M., Truong, T. K., and Madden, D. J. (2018). "Age-Related Differences in the Neural Bases of Phonological and Semantic Processes in the Context of Task-Irrelevant Information," *Cognitive, Affective, & Behavioral Neuroscience*, pp. 1–16.

[64] Li, S. C. and Lindenberger, U. (1999). "Cross-Level Unification: A Computational Exploration of the Link between Deterioration of Neurotransmitter Systems and Dedifferentiation of Cognitive Abilities in Old Age," in *Cognitive Neuroscience of Memory* (L.-G. N. and H. J. M., eds.), Hogrefe and Huber, Seattle, WA, pp. 103–146.

[65] Burke, D. M. and Shafto, M. A. (2008). "Language and Aging," in *The Handbook of Aging and Cognition, 3rd ed.* (F. I. Craik and T. A. Salthouse, eds.), Psychology Press, New York, NY, pp. 373–443.

[66] Dennis, N. A. and Cabeza, R. (2008). "Neuroimaging of Healthy Cognitive Aging," in *The Handbook of Aging and Cognition, 3rd ed.* (F. I. Craik and T. A. Salthouse, eds.), Psychology Press, New York, NY, pp. 1–54.

[67] Meunier, D., Stamatakis, E. A., and Tyler, L. K. (2014). "Age-Related Functional Reorganization, Structural Changes, and Preserved Cognition," *Neurobiology of Aging*, **35**(1), pp. 42–54.

[68] Price, C. J. (2010). "The Anatomy of Language: A Review of 100 fMRI Studies Published in 2009," *Annals of the New York Academy of Sciences*, **1191**(1), pp. 62–88.

[69] Price, C. J. (2012). "A Review and Synthesis of the First 20 Years of PET and fMRI Studies of Heard Speech, Spoken Language and Reading," *Neuroimage*, **62**(2), pp. 816–847.

[70] Wierenga, C. E., Benjamin, M., Gopinath, K., Perlstein, W. M., Leonard, C. M., Rothi, L. J. G., Conway, T., Cato, M. A., Briggs, R., and Crosson, B. (2008). "Age-Related Changes in Word Retrieval: Role of Bilateral Frontal and Subcortical Networks," *Neurobiology of Aging*, **29**(3), pp. 436–451.

[71] Baciu, M., Boudiaf, N., Cousin, E., Perrone-Bertolotti, M., Pichat, C., Fournet, N., Chainay, H., Lamalle, L., and Krainik, A. (2016). "Functional MRI Evidence for the Decline of Word Retrieval and Generation During Normal Aging," *Age*, **38**(1), p. 3.

[72] Verhaegen, C. and Poncelet, M. (2013). "Changes in Naming and Semantic Abilities with Aging from 50 to 90 Years," *Journal of the International Neuropsychological Society*, **19**(2), pp. 119–126.

[73] Badre, D., Lebrecht, S., Pagliaccio, D., Long, N. M., and Scimeca, J. M. (2014). "Ventral Striatum and the Evaluation of Memory Retrieval Strategies," *Journal of Cognitive Neuroscience*, **26**(9), pp. 1928–1948.

[74] Scimeca, J. M. and Badre, D. (2012). "Striatal Contributions to Declarative Memory Retrieval," *Neuron*, **75**(3), pp. 380–392.

[75] Hasher, L. and Zacks, R. T. (1988). "Working Memory, Comprehension, and Aging: A Review and a New View," in *The Psychology of Learning and Motivation, vol. 22* (G. H. Bower, ed.), Academic Press, San Diego, CA, pp. 193–225.

[76] Dosenbach, N. U., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A., Fox, M. D., Snyder, A. Z., Vincent, J. L., Raichle, M. E., and Schlaggar, B. L. (2007). "Distinct Brain Networks for Adaptive and Stable Task Control in Humans," *Proceedings of the National Academy of Sciences*, **104**(26), pp. 11073–11078.

[77] Burke, D. M., Mackay, D. G., Worthley, J. S., and Wade, E. (1991). "On the Tip of the Tongue: What Causes Word Finding Failures in Young and Older Adults?" *Journal of Memory and Language*, **30**, pp. 542–579.

[78] Diersch, N., Jones, A. L., and Cross, E. S. (2016). "The Timing and Precision of Action Prediction in the Aging Brain," *Human Brain Mapping*, **37**(1), pp. 54–66.

[79] Gonneaud, J., Lecouvey, G., Groussard, M., Gaubert, M., Landeau, B., Mézenge, F., de La Sayette, V., Eustache, F., Desgranges, B., and Rauchs, G. (2017). "Functional Dedifferentiation and Reduced Task-Related Deactivations Underlie the Age-Related Decline of Prospective Memory," *Brain Imaging and Behavior*, **11**(6), pp. 1873–1884.

[80] Froehlich, E., Liebig, J., Morawetz, C., Ziegler, J. C., Braun, M., Heekeren, H. R., and Jacobs, A. M. (2018). "Same Same but Different: Processing Words in the Aging Brain," *Neuroscience*, **371**, pp. 75–95.

[81] Craddock, R. C., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2009). "Disease State Prediction from Resting State Functional Connectivity," *Magnetic Resonance in Medicine*, **62**(6), pp. 1619–1628.

[82] Antonenko, D., Brauer, J., Meinzer, M., Fengler, A., Kerti, L., Friederici, A. D., and Flöel, A. (2013). "Functional and Structural Syntax Networks in Aging," *Neuroimage*, **83**, pp. 513–523.

[83] Friston, K. J. (2011). "Functional and Effective Connectivity: A Review," *Brain Connectivity*, **1**(1), pp. 13–36.

[84] Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). "Discriminative Analysis of Resting-State Functional Connectivity Patterns of Schizophrenia using Low Dimensional Embedding of fMRI," *Neuroimage*, **49**(4), pp. 3110–3121.

[85] Lord, A., Horn, D., Breakspear, M., and Walter, M. (2012). "Changes in Community Structure of Resting State Functional Connectivity in Unipolar Depression," *PloS One*, **7**(8), p. e41282.

[86] Vergun, S., Deshpande, A., Meier, T. B., Song, J., Tudorascu, D. L., Nair, V. A., Singh, V., Biswal, B. B., Meyerand, M. E., Birn, R. M., and Prabhakaran, V. (2013). "Characterizing Functional Connectivity Differences in Aging Adults using Machine Learning on Resting State fMRI Data," *Frontiers in Computational Neuroscience*, **7**, p. 38.

[87] Baltes, P. B., Cornelius, S. W., Spiro, A., Nesselroade, J. R., and Willis, S. L. (1980). "Integration Versus Differentiation of Fluid/Crytallized Intelligence in Old Age," *Developmental Psychology*, **16**(6), p. 625.

[88] Dale, A. M. (1999). "Optimal Experimental Design for Event-Related fMRI," *Human Brain Mapping*, **8**(2-3), pp. 1619–1628.

[89] Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). "FSL," *Neuroimage*, **62**(2), pp. 782–790.

[90] Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images," *Neuroimage*, **17**(2), pp. 825–841.

[91] Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., and Niazy, R. (2004). "Advances in Functional and Structural MR Image Analysis and Implementation as FSL," *Neuroimage*, **23**, pp. S208–S219.

[92] Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., and Dolan, R. J. (1997). "Psychophysiological and Modulatory Interactions in Neuroimaging," *Neuroimage*, **6**(3), pp. 218–229.

[93] Gitelman, D. R., Penny, W. D., Ashburner, J., and Friston, K. J. (2003). "Modeling Regional and Psychophysiologic Interactions in fMRI: The Importance of Hemodynamic Deconvolution," *Neuroimage*, **19**(1), pp. 200–207.

[94] Bounova, G. (2015). "Octave Networks Toolbox," [August 2].

[95] Breiman, L. (2001). "Random Forests," *Machine Learning*, **45**(1), pp. 5–32.

[96] Biau, G. and Scornet, E. (2016). "A Random Forest Guided Tour," *Test*, **25**(2), pp. 197–227.

[97] Latinne, P., Debeir, O., and Decaestecker, C. (2001). "Limiting the Number of Trees in Random Forests," in *International Workshop on Multiple Classifier Systems* (K. J. and R. F., eds.), Springer, Berlin, Heidelberg, pp. 178–187.

[98] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). "Measuring Statistical Dependence with Hilbert-Schmidt Norms," in *Algorithmic Learning Theory* (S. Jain, H. U. Simon, and E. Tomita, eds.), Springer, Berlin, Heidelberg, pp. 63–78.

[99] Lee, S. and Honavar, V. G. (2017). "Self-Discrepancy Conditional Independence Test," in *Uncertainty in Artificial Intelligence, Vol. 33*, Sydney, Australia, p. 16.

[100] Li, S. C. and Sikström, S. (2002). "Integrative Neurocomputational Perspectives on Cognitive Aging, Neuromodulation, and Representation," *Neuroscience and Biobehavioral Reviews*, **26**(7), pp. 795–808.

[101] Badre, D. and D'Esposito, M. (2007). "Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex," *Journal of Cognitive Neuroscience*, **19**(12), pp. 2082–2099.

[102] Badre, D. and D'Esposito, M. (2009). "Is the Rostro-Caudal Axis of the Frontal Lobe Hierarchical?" *Nature Reviews Neuroscience*, **10**(9), pp. 659–669.

[103] Devlin, J. T., Matthews, P. M., and Rushworth, M. F. (2003). "Semantic Processing in the Left Inferior Prefrontal Cortex: A Combined Functional Magnetic Resonance Imaging and Transcranial Magnetic Stimulation Study," *Journal of Cognitive Neuroscience*, **15**(1), pp. 71–84.

[104] Krieger-Redwood, K., Jefferies, E., Karapanagiotidis, T., Seymour, R., Nunes, A., Ang, J. W. A., Majernikova, V., Mollo, G., and Smallwood, J. (2016). "Down but not Out in Posterior Cingulate Cortex: Deactivation yet Functional Coupling with Prefrontal Cortex During Demanding Semantic Cognition," *Neuroimage*, **141**, pp. 366–377.

[105] Dosenbach, N. U., Fair, D. A., Cohen, A. L., Schlaggar, B. L., and Petersen, S. E. (2008). "A Dual-Networks Architecture of Top-Down Control," *Trends in Cognitive Sciences*, **12**(3), pp. 99–105.

[106] Dubis, J. W., Siegel, J. S., Neta, M., Visscher, K. M., and Petersen, S. E. (2014). "Tasks Driven by Perceptual Information do not Recruit Sustained BOLD Activity in Cingulo-Opercular Regions," *Cerebral Cortex*, **26**(1), pp. 192–201.

[107] Sadaghiani, S. and D'Esposito, M. (2014). "Functional Characterization of the Cingulo-Opercular Network in the Maintenance of Tonic Alertness," *Cerebral Cortex*, **25**(9), pp. 2763–2773.

[108] Coste, C. P. and Kleinschmidt, A. (2016). "Cingulo-Opercular Network Activity Maintains Alertness," *Neuroimage*, **128**, pp. 264–272.

[109] Diaz, M. T., Johnson, M. A., Burke, D. M., and Madden, D. J. (2014). "Age-Related Differences in the Neural Bases of Phonological and Semantic Processes," *Journal of Cognitive Neuroscience*, **26**(12), pp. 2798–2811.

[110] Hamm, V. P. and Hasher, L. (1992). "Age and the Availability of Inferences," *Psychology and Aging*, **7**, pp. 56–64.

[111] Hasher, L., Quig, M. B., and May, C. P. (1997). "Inhibitory Control Over no-Longer-Relevant Information: Adult Age Differences," *Memory & Cognition*, **25**(3), pp. 286–295.

[112] Lahar, C. J., Isaak, M. I., and McArthur, A. D. (2001). "Age Differences in the Magnitude of the Attentional Blink," *Neuropsychology, and Cognition*, **8**(2), pp. 149–159.

[113] Cabeza, R. and Dennis, N. A. (2012). "Frontal Lobes and Aging," in *Principles of Frontal Lobe Function, 2nd ed.* (D. T. Stuss and R. T. Knight, eds.), Oxford University Press, New York, pp. 628–652.

[114] Hull, J. V., Dokovna, L. B., Jacokes, Z. J., Torgerson, C. M., Irimia, A., and Van Horn, J. D. (2017). "Resting-State Functional Connectivity in Autism Spectrum Disorders: A Review," *Frontiers in Psychiatry*, **7**, p. 205.

[115] (WHO), W. H. O. (2018). "Autism Spectrum Disorders," `https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders`, [Online; accessed 10-July-2019].

[116] Christensen, D. L., Braun, K. V. N., Baio, J., Bilder, D., Charles, J., Constantino, J. N., Daniels, J., Durkin, M. S., Fitzgerald, R. T., Kurzius-Spencer, M., and Lee, L. C. (2018). "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2012," *MMWR Surveillance Summaries*, **65**(13), p. 1.

[117] Dyer, C. (2010). "Lancet Retracts Wakefield's MMR Paper," C696.

[118] Maisonneuve, H. and Floret, D. (2012). "Wakefield's Affair: 12 Years of Uncertainty Whereas butLink between Autism and MMR Vaccine has been Proved," *Presse Medicale*, **41**(9 Pt 1), pp. 827–834.

[119] Taylor, L. E., Swerdfeger, A. L., and Eslick, G. D. (2014). "Vaccines are not Associated with Autism: An Evidence-Based Meta-Analysis of Case-Control and Cohort Studies," *Vaccine*, **32**(29), pp. 3623–3629.

[120] Iidaka, T. (2015). "Resting State Functional Magnetic Resonance Imaging and Neural Network Classified Autism and Control," *Cortex*, **63**, pp. 55–67.

[121] Bhaumik, R., Pradhan, A., Das, S., and Bhaumik, D. K. (2018). "Predicting Autism Spectrum Disorder using Domain-Adaptive Cross-Site Evaluation," *Neuroinformatics*, **16**(2), pp. 197–205.

[122] Mahanand, B. S., Vigneshwaran, S., Suresh, S., and Sundararajan, N. (2016). "An Enhanced Affect-Size Thresholding Method for the Diagnosis of Autism Spectrum Disorder using Resting State Functional MRI," in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, IEEE, pp. 1–6.

[123] Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., and Rueckert, D. (2017). "Spectral Graph Convolutions for Population-Based Disease Prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 177–185.

[124] Van Horn, J. D. and Toga, A. W. (2009). "Multi-Site Neuroimaging Trials," *Current Opinion in Neurology*, **22**(4), pp. 370–378.

[125] Plitt, M., Barnes, K. A., and Martin, A. (2015). "Functional Connectivity Classification of Autism Identifies Highly Predictive Brain Features but Falls Short of Biomarker Standards," *Neuroimage: Clinical*, **7**, pp. 359–366.

[126] Guo, X., Dominick, K. C., Minai, A. A., Li, H., Erickson, C. A., and Lu, L. J. (2017). "Diagnosing Autism Spectrum Disorder from Brain Resting-State Functional Connectivity Patterns using a Deep Neural Network with a Novel Feature Selection Method," *Frontiers in Neuroscience*, **11**, p. 460.

[127] Sartipi, S., Kalbkhani, H., and Shayesteh, M. G. (2017). "Ripplet II Transform and Higher Order Cumulants from R-fMRI Data for Diagnosis of Autism," in *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pp. 557–560.

[128] Hong, S. J., de Wael, R. V., Bethlehem, R. A. I., Lariviere, S., Paquola, C., Valk, S. L., Milham, M. P., Di Martino, A., Margulies, D. S., Smallwood, J., and Bernhardt, B. C. (2019). "Atypical Functional Connectome Hierarchy in Autism," *Nature Communications*, **10**(1), p. 1022.

[129] Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., Lainhart, J. E., and Anderson, J. S. (2013). "Multisite Functional Connectivity MRI Classification of Autism: ABIDE Results," *Frontiers in Human Neuroscience*, **7**, p. 599.

[130] Ghiassian, S., Greiner, R., Jin, P., and Brown, M. R. G. (2016). "Using Functional or Structural Magnetic Resonance Images and Personal Characteristic Data to Identify ADHD and Autism," *PloS One*, **11**(12), p. e0166934.

[131] Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). "Identification of Autism Spectrum Disorder using Deep Learning and the ABIDE Dataset," *Neuroimage: Clinical*, **17**, pp. 16–23.

[132] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). "Automated Anatomical Labeling of Activations in SPM using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain," *Neuroimage*, **15**(1), pp. 273–289.

[133] Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., and Zilles, K. (2005). "A New SPM Toolbox for Combining Probabilistic Cytoarchitectonic Maps and Functional Imaging Data," *Neuroimage*, **25**(4), pp. 1325–1335.

[134] 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., and Albert, M. S. (2006). "An Automated Labeling System

for Subdividing the Human Cerebral Cortex on MRI Scans into Gyral Based Regions of Interest," *Neuroimage*, **31**(3), pp. 968–980.

[135] Vigneshwaran, S., Mahanand, B. S., Suresh, S., and Sundararajan, N. (2015). "Using Regional Homogeneity from Functional MRI for Diagnosis of ASD Among Males," in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

[136] Hong, S. J., Valk, S. L., Di-Martino, A., Milham, M. P., and Bernhardt, B. C. (2017). "Multidimensional Neuroanatomical Subtyping of Autism Spectrum Disorder," *Cerebral Cortex*, **28**(10), pp. 3578–3588.

[137] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. (1994). "Statistical Parametric Maps in Functional Imaging: A General Linear Approach," *Human Brain Mapping*, **2**(4), pp. 189–210.

[138] Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). "A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI," *Neuroimage*, **37**(1), pp. 90–101.

[139] Avants, B. B., Tustison, N., and Song, G. (2009). "Advanced Normalization Tools (ANTS)," *Insight J*, **2**, pp. 1–35.

[140] Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). "Machine Learning for Neuroimaging with Scikit-Learn," *Frontiers in Neuroinformatics*, **8**, p. 14.

[141] Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*, vol. 2045, Princeton University Press.

[142] John, G. H. and Langley, P. (1995). "Estimating Continuous Distributions in Bayesian Classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., pp. 338–345.

[143] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Vanderplas, J. (2011). "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, **12**, pp. 2825–2830.

[144] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, **13**(Mar), pp. 723–773.

[145] Orban, P., Dansereau, C., Desbois, L., Mongeau-Pérusse, V., Giguère, C. É., Nguyen, H., Mendrek, A., Stip, E., and Bellec, P. (2018). "Multisite Generalizability of Schizophrenia Diagnosis Classification Based on Functional Brain Connectivity," *Schizophrenia Research*, **192**, pp. 167–171.

[146] Gur, S. and Honavar, V. G. (2018). "PATENet: Pairwise Alignment of Time Evolving Networks," in *International Conference on Machine Learning and Data Mining in Pattern Recognition* (P. Perner, ed.), vol. 10934, Springer, Cham, pp. 85–98.

[147] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). "Multilayer Networks," *Journal of Complex Networks*, **2**(3), pp. 203–271.

[148] Li, A., Cornelius, S. P., Liu, Y. Y., Wang, L., and Barabasi, A. (2017). "The Fundamental Advantages of Temporal Networks," *Science*, **358**(6366), pp. 1042–1046.

[149] Luo, G., Cordier, F., and Seo, H. (2016). "Spatio-Temporal Segmentation for the Similarity Measurement of Deforming Meshes," *The Visual Computer*, **32**(2), pp. 243–256.

[150] Smith, T. F. and Waterman, M. S. (1981). "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, **147**(1), pp. 195–197.

[151] Caspi, Y. and Irani, M. (2002). "Spatio-Temporal Alignment of Sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(11), pp. 1409–1424.

[152] Lee, W. N. and Das, A. K. (2010). "Local Alignment Tool for Clinical History: Temporal Semantic Search of Clinical Databases," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, pp. 437–441.

[153] Elzinga, C. H. (2014). "Distance, Similarity and Sequence Comparison," in *Advances in Sequence Analysis: Theory, Method, Applications*, Springer, pp. 51–73.

[154] Emmert-Streib, F., Dehmer, M., and Shi, Y. (2016). "Fifty Years of Graph Matching, Network Alignment and Network Comparison," *Information Sciences*, **346**, pp. 180–197.

[155] Koutra, D., Vogelstein, J. T., and Faloutsos, C. (2013). "Deltacon: A Principled Massive-Graph Similarity Function," in *Proceedings of the 2013 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, pp. 162–170.

[156] Albert, R. and Barabasi, A. (2000). "Topology of Evolving Networks: Local Events and Universality," *Physical Review Letters*, **85**(24), pp. 5234–5237.

[157] Dorogovtsev, S. N. and Mendes, J. F. F. (2000). "Scaling Behaviour of Developing and Decaying Networks," *Europhysics Letters*, **52**(1), pp. 33–39.

[158] Erdos, P. and Renyi, A. (1959). "On Random Graphs I," *Publ. Math. Debrecen*, **6**, pp. 290–297.

[159] Barabasi, A. and Albert, R. (1999). "Emergence of Scaling in Random Networks," *Science*, **826**(5439), pp. 509–512.

[160] Towfic, F., Greenlee, M. H., and Honavar, V. (2009). "Aligning Biomolecular Networks using Modular Graph Kernels," in *Proceedings of the 9th International Workshop on Algorithms in Bioinformatics*, Springer, pp. 345–361.

[161] McLaren, D. G., Ries, M. L., Xu, G., and Johnson, S. C. (2012). "A Generalized Form of Context-Dependent Psychophysiological Interactions (gPPI): A Comparison to Standard Approaches," *Neuroimage*, **61**(4), pp. 1277–1286.

[162] Glover, G. H. (1999). "Deconvolution of Impulse Response in Event-Related BOLD fMRI," *Neuroimage*, **9**(4), pp. 416–429.

[163] Bullmore, E., Fadili, J., Maxim, V., Şendur, W. B., L., Suckling, J., Brammer, M., and Breakspear, M. (2004). "Wavelets and Functional Magnetic Resonance Imaging of the Human Brain," *Neuroimage*, **23**, pp. S234–S249.

[164] Zhang, Z., Telesford, Q. K., Giusti, C., Lim, K. O., and Bassett, D. S. (2016). "Choosing Wavelet Methods, Filters, and Lengths for Functional Brain Network Construction," *PloS One*, **11**(6), p. e0157243.

[165] Massey Jr, F. J. (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, **46**(253), pp. 68–78.

[166] Braams, B. R., van Duijvenvoorde, A. C. K., Peper, J. S., and Crone, E. A. (2015). "Longitudinal Changes in Adolescent Risk-Taking: A Comprehensive Study of Neural Responses to Rewards, Pubertal Development, and Risk-Taking Behavior," *Journal of Neuroscience*, **35**(18), pp. 7226–7238.

[167] Marek, S., Hwang, K., Foran, W., Hallquist, M. N., and Luna, B. (2015). "The Contribution of Network Organization and Integration to the Development of Cognitive Control," *PLoS Biology*, **13**(12), p. e1002328.

[168] Shulman, E. P., Smith, A. R., Silva, K., Icenogle, G., Duell, N., Chein, J., and Steinberg, L. (2016). "The Dual Systems Model: Review, Reappraisal, and Reaffirmation," *Developmental Cognitive Neuroscience*, **17**, pp. 103–117.

[169] Simmonds, D. J., Hallquist, M. N., and Luna, B. (2017). "Protracted Development of Executive and Mnemonic Brain Systems Underlying Working Memory in Adolescence: A Longitudinal fMRI Study," *Neuroimage*, **157**, pp. 695–704.

[170] Fair, D. A., Schlaggar, B. L., Cohen, A. L., Miezin, F. M., Dosenbach, N. U. F., Wenger, K. K., Fox, M. D., Snyder, A. Z., Raichle, M. E., and Petersen, S. E. (2007). "A Method for using Blocked and Event-Related fMRI Data to Study "Resting State" Functional Connectivity," *Neuroimage*, **35**(1), pp. 396–405.

[171] Robbins, T. W., James, M., Owen, A. M., Sahakian, B. J., McInnes, L., and Rabbitt, P. (1994). "Cambridge Neuropsychological Test Automated Battery (CANTAB): A Factor Analytic Study of a Large Sample of Normal Elderly Volunteers," *Dementia and Geriatric Cognitive Disorders*, **5**(5), pp. 266–281.

[172] Schwarz, A. J. and McGonigle, J. (2011). "Negative Edges and Soft Thresholding in Complex Network Analysis of Resting State Functional Connectivity Data," *Neuroimage*, **55**(3), pp. 1132–1146.

Vita

Shlomit Gur

Education

- Ph.D., Neuroscience, Intercollege Graduate Program in Neuroscience, Pennsylvania State University (University Park, PA, USA), 2019
  - Machine-learning-based Neuroinformatics (Adviser: Prof. Dr. V.G. Honavar; Artificial Intelligence Research lab, College of Information Sciences and Technology)
- M.Sc., Neuroscience (Research), Faculty of Medicine and Health Sciences, Erasmus University (Rotterdam, the Netherlands), 2013
  - Computational Neuroscience (Adviser: Dr. A.R. Houweling)
  - Electrophysiology (Advisers: Prof. Dr. J.G.G. Borst and Dr. M. van der Heijden)
- B.Sc., Mathematics and Minor in Computer Science, Bar-Ilan University (Ramat-Gan, Israel), 2006

Selected Publications

- Gur, S., El-Manzalawy, Y., Diaz, M. T., and Honavar, V. G. "Age-Related Differences in Task-Specific Functional Connectivity in Phonological and Semantic Picture-Based Match-Mismatch Tasks in the Presence of Distractor Words," (under review).
- Gur, S. and Honavar, V. G. (2018). "PATENet: Pairwise Alignment of TimeEvolving Networks," in International Conference on Machine Learning and Data Mining in Pattern Recognition (P. Perner, ed.), vol. 10934, Springer, Cham, pp.85-98.