BLENDING GENERATIVE MODELS WITH DEEP LEARNING FOR MULTIDIMENSIONAL PHENOTYPIC PREDICTION FROM BRAIN CONNECTIVITY DATA

by

Niharika S. D'Souza

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
December 2021

© 2021 Niharika S. D'Souza
All rights reserved

Abstract

Network science has provided us with foundational machinery to study complex entities such as social networks, genomics etc. The human brain is a complex network that has garnered immense interest within data science. Connectomics or the study of the underlying brain connectivity patterns has become important for the characterization of various neurological disorders such as Autism, Schizophrenia etc.

This thesis proposes a collection of mathematical models that can fuse information from functional and structural connectivity with clinical phenotypes. Here, functional connectivity is measured by resting state functional MRI (rs-fMRI), while anatomical connectivity is captured using Diffusion Tensor Imaging (DTI). The phenotypes could be continuous measures of behavior or cognition, or may capture levels of impairment in the case of neuropsychiatric disorders and are often scored by clinicians.

We first develop a joint network optimization framework to predict clinical severity from rs-fMRI connectivity. This model couples two key terms into a unified optimization framework: a generative matrix factorization and a discriminative linear regression. We demonstrate that the proposed joint inference successfully generalizes to predicting impairments in Autism Spectrum

Disorder (ASD). Moreover, the model can extract brain biomarkers that are informative of individual clinical measures. We then present two modeling extensions to non-parametric and neural networks in lieu of linear regression.

Next, we extend our framework to incorporate multimodal information from Diffusion Tensor Imaging (DTI) and dynamic functional connectivity. Our generative matrix factorization now estimates a time-varying functional decomposition. At the same time, it is guided by anatomical connectivity priors in a graph-based regularization. This framework is coupled with a deep network that predicts multidimensional clinical characterizations and models the temporal dynamics of the functional scan. Overall, we can simultaneously explain multiple impairments, isolate stable multi-modal connectivity signatures, and study the evolution of various brain states at rest.

Lastly, we focus on end-to-end geometric frameworks which are designed to characterize the complementarity between functional and structural connectomes, while using clinical information as a secondary guide. Our representation learning scheme is a matrix autoencoder that can reflect the underlying data geometry. This is coupled with a manifold alignment model that maps from function to structure and a deep network that maps to phenotypic information. We demonstrate that the model reliably recovers structural connectivity patterns across individuals, while robustly extracting predictive yet interpretable brain biomarkers. Finally, we also present a preliminary exposition on the theoretical aspects of the representation.

Thesis Committee

Primary Readers

Dr. Archana Venkataraman (Primary Advisor)
John C. Malone Assistant Professor,
Department of Electrical and Computer Engineering.

Core Faculty,
Malone Center for Engineering in Healthcare and,
Center for Imaging Science (CIS) and,
Mathematical Institute for Data Science.

Secondary Appointment,
Department of Computer Science and,
Department of Applied Mathematics and Statistics.
Johns Hopkins Whiting School of Engineering

Dr. Amitabh Basu
Associate Professor,
Department of Applied Mathematics and Statistics

Secondary Appointment,
Department of Computer Science.
Johns Hopkins Whiting School of Engineering

Dissertation Committee Members

Dr. Rene Vidal
Herschel L. Seder Professor,
Department of Biomedical Engineering and,
Director, Mathematical Institute for Data Science.

Secondary Appointment,
Department of Electrical and Computer Engineering and,
Department of Computer Science and,
Department of Mechanical Engineering.

Faculty Member,
Center for Imaging Science (CIS) and,
Institute for Computational Medicine (ICM) and,
Laboratory for Computational Sensing and Robotics (LCSR).
Johns Hopkins Whiting School of Engineering

Dr. Stewart Mostofsky
Director,
Center for Neurodevelopmental and Imaging Research,
Kennedy Krieger Institute.

Professor,
Department of Neurology.
Johns Hopkins School of Medicine.

Dr. Kilian Pohl
Associate Professor,
Department of Psychiatry & Behavioral Sciences,
Stanford University.

Director of Biomedical Computing,
Center for Health Sciences
SRI International.

Acknowledgments

This momentous undertaking would not have been nearly as enjoyable and meaningful without the support and encouragement of several key individuals.

First and foremost, I am extremely grateful to my advisor, Prof. Archana Venkataraman. She was my first lens into the fascinating, yet (not so) random world of computational neuroimaging and machine learning. I am thankful to have had her unwavering support through this tumultuous yet rewarding pursuit of challenging problems. I am glad that she patiently yet firmly pushed me out of my comfort zone. Her profound research insights and broad vision have been foundational in shaping my identity as an independent researcher.

I express my heartfelt gratitude to my dissertation committee members, Dr. Amitabh Basu, Dr. Stewart Mostofsky, Dr. Renè Vidal, and Dr. Kilian Pohl, to Dr. Daniel Robinson and Dr. Jerry Prince for serving as a part of my Graduate Board Oral committee, and to Dr. Carey Priebe for being a part of my thesis proposal committee. Their constructive feedback has helped me hone my critical thinking ability, and has instilled a sense of technical rigour that is core of this research endeavour. I am also grateful to Dr. Tanveer Syeda-Mahmood, for giving me a chance to be a part of the IBM Research team during my graduate tenure as an intern and a second time as a staff member. Among my mentors, I am especially grateful to Dr. Debdoot Sheet, Dr. Baylon Fernandes, and Dr. Chandan Chakraborty for their support of my professional goals and the opportunities they provided me through my undergraduate years.

Along with Prof. Venkataraman, this work would not have been as impactful without the intellectual contributions of our collaborators at Kennedy Krieger Institute. Dr. Mary Beth Nebel, Dr. Mostofsky, and Dr. Nicholas Wymbs curated the resting state data used in this work and provided clinical insights into the implications and interpretation of the results. Deana Crocetti and Joshua Robinson curated the diffusion data and have been invaluable in helping me process and understand the data.

I am indebted to my lab-mates from the Neural Systems Analysis Lab: Naresh Nandakumar, Jeff Craley, Ravi Shankar and Sayan Ghosal. Thank you for sacrificing countless hours to practice talks, collaborations, research discussions, journal clubs, and proof reading. Graduate life at Hopkins would have been way more tedious if not for your engaging presence and spirited banter. I will forever be trying to find better research companionship!

I am grateful for the friendship and encouragement provided by Ruhika Darbare, Ranjani Srinivasan, Arun Nair, Sriram Sankaranarayanan, Aditya Joysula, Gayatri Pahapale, and Kaustav Bera during trying times in my graduate tenure. I am indebted to Anirud Thyagharajan and Sebin Mathew, who shaped my aspiration to work as a machine learning researcher early on. Thank you to all of you for lending a listening ear and providing insightful advice to navigate the twists and turns of life, both internal and external to graduate school.

I would not have been where I am today without the love and devotion of my family. I am grateful to my parents: Kripa and John D'Souza, Aunt, Uncle, and Cousin: Kala, Dimitrius and Ethan D'Mello, and finally to Orville Pinto. Thank you for endlessly shuttling me across academic centers, and for supporting and fostering all of my endeavours. Most of all, thank you for always believing in me through my darkest moments, and always exemplifying dedication and excellence.

Finally, this thesis is dedicated to my loving grandparents, Hilda and Late Charles D'Souza. My grandmother was my first mathematics and science teacher. Not only did she push me to excel at academics, but also encouraged me to be forever curious about, perceptive to, and respectful of the beauty of the world around me. I will forever cherish all that I have learned from you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The human brain is a complex and mysterious entity that is at the core of our existence and being. It integrates and coordinates key information received from the sense organs. It codifies instructions that control the rest of the body, thus playing a central role in decision making. From an anatomical standpoint, it can be described as a network of individual processing centers interconnected by neural axons. Functionally, the brain partitions itself into a myriad of regional hubs that specialize at complex and sometimes abstract tasks. With recent technological advancements, non-invasive imaging techniques such as MRI, CT, PET, EEG allow us to probe the organization of the brain [1]. Traditional clinical analyses have focused on characterizing localized morphometric properties such as segmenting key structures [2], tracking changes in volume or tissue properties within disease types [3], and pinpointing regions of (aberrant) functional activation [4].

Over the past decade, there has been a growing emphasis in neuroscience to analyze the human brain as a complex network of interacting entities. Connectomics is the study of such underlying connectivity patterns. In fact, connectomics studies have provided several fundamental insights into the organization of the human brain.

In this thesis, we take a close look at structural and functional connectivity between brain regions at a population level. Structural connectivity informs us about the neuronal fiber bundles that connect regions of the brain, and is often measured using Diffusion Tensor Imaging (DTI) [5]. DTI provides a proxy for these structural connections by measuring the directional (anisotropic) diffusion of water molecules within the brain. Further, a computational algorithm (tractography) [6] is employed in order to estimate and track the location and direction of these fiber bundles within the white matter in the brain.

In parallel, functional neuroimaging provides a glimpse into communication patterns in the brain. A common neuroimaging modality for measuring the functional connectivity is the functional Magnetic Resonance Imaging (fMRI) protocol. FMRI studies the co-activation patterns across different brain regions either in response to external stimuli or at rest, the underlying assumption being that two brain regions which reliably co-activate are more likely to participate in similar neural processes as opposed to two uncorrelated regions [7, 8, 9].

Task-based fMRI has been widely in order to isolate brain regions that are functionally associated with the completion of a specific task. These protocols require careful experimental design as well as subject training against the task paradigm. On the other hand, resting state fMRI (rs-fMRI) is acquired in the absence of external stimuli. In practice, the underlying functional connectivity is often quantified via temporal correlations between neural fMRI responses. Thus, rs-fMRI can be used to probe the intrinsic functional specialization of brain regions/networks in steady state. This could be particularly useful for brain mapping in individuals with motor, language or cognitive impairments, neurodevelopmental disorders, or pediatric populations.

We explore two key ideas in this thesis, namely (1) representation learning for functional and structural connectivity and (2) prediction of fine-grained multidimensional phenotypic measures. We build up a unified framework that integrates DTI (structural) and rs-fMRI (dynamic functional) connectivity data. Deriving inspiration from classical representation learning, our generative models are designed to interpretable rather than black-box in nature. This in turn helps us probe learned representations and isolate canonical connectivity signatures (referred to as subnetworks). In order to better explain subject-level differences within a cohort, we focus on the prediction of phenotypic measures in unseen patients. Our models leverage the representational flexibility of deep networks in order to map to multidimensional and diverse clinical characterizations. We also leverage key insights from the fields of numerical optimization, machine learning, graph theory, and deep learning to ensure our models are well posed and computationally tractable. Overall, these efforts help us obtain a more holistic understanding of brain connectivity and its behavioral implications.

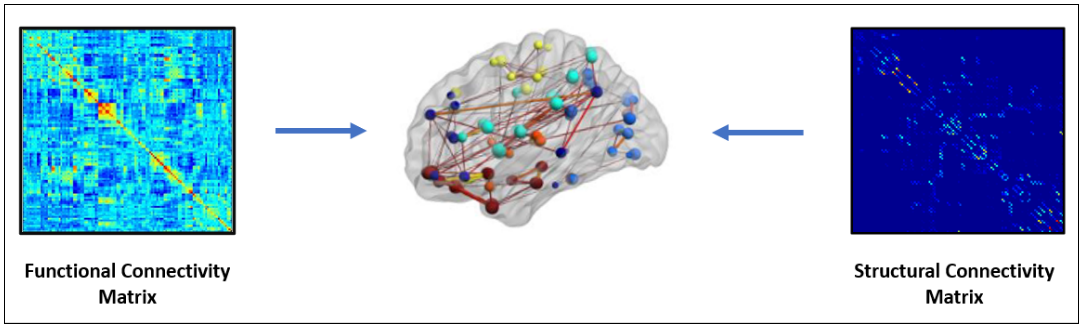

Functional Connectivity Matrix
Structural Connectivity Matrix

Figure 1.1: Dual representation of the brain as given by **(L)** functional connectivity and **(R)** structural connectivity. The former is measured by functional Magnetic Resonance Imaging (fMRI) and Diffusion Tensor Imaging (DTI)

1.1 Multimodal Integration of Functional and Structural Connectivity

The brain is increasingly being viewed as an interconnected network. Two key elements of this network are the structural pathways between brain regions and the functional signaling that rides on top. In this sense, DTI and rs-fMRI studies provide a dual representation of the brain, as can be seen in Fig. 1.1. We have functional connectivity to the left while we have structural connectivity represented on the right. One of the key contributions of this work is to develop generative frameworks that are amenable to the network based, multimodal representation of brain connectivity.

Fundamentally, function and structure are two distinct yet related viewpoints. There is strong evidence in literature of of both direct and indirect correspondences between functional and structural pathways within the brain [10, 11, 12]. Neuroimaging studies also suggest that this functional connectivity may be mediated by either direct or indirect anatomical connections [12, 13, 11, 14]. Going a step further, structural and functional connectivity have been shown to be predictive of each other at varying scales [15, 16, 17]. Thus, rs-fMRI and DTI data provide complementary information about connectivity, which when integrated together can be used to construct a more comprehensive picture of brain organization in health and disease.

Consequently, clinical research has recently been pivoting to multimodal integration with the goal of reliably inferring key properties of the brain. For example, these studies have provided us with fundamental insights into brain dysfunction in neurological disorders such as Autism [18], Schizophrenia [19], and Epilepsy [20, 21]. While very informative, hypothesis-driven discovery in this domain is nevertheless challenging due to the high data dimensionality, environmental confounds, and considerable inter-individual variability.

1.1.1 Evolution of Brain Connectivity Analysis

Traditionally, the analysis of brain connectivity has focused heavily on extracting key statistics from the scan, and quantifying variations in these statistics either between groups or across individuals. An example of such an approach in resting state functional MRI is seed based correlation analysis [22]. Here, the goal is to identify functional systems in the brain as voxels whose temporal dynamics are strongly correlated with that of a pre-specified anatomical “seed" location. Simple statistical differences in rs-fMRI and DTI connectivity between individuals have been shown to be indicative disrupted patterns of brain organization in Alzheimer’s disease [23] or Progressive Supranuclear Palsy (PSP) [24]. Similarly, classical statistical models such as multivariate analysis [25, 26] or random effects models [27] have also been employed for uncovering disease biomarkers from multimodal connectivity data. A popular alternative strategy involves the use of graph theory to summarize key connectome properties using aggregate network notions [28, 29, 30]. Nevertheless, most summary measures are typically independently computed for each modality and/or each region pair to further isolate connections that collectively differ across clinical populations and healthy individuals.

It is believed that the brain is organized as distributed network of localized and overlapping neuronal sub-systems that process and relay information. Often the full richness of this characterization may be lost in the extraction of network statistics. In this regard, model based alternatives (eg. mechanistic [31] or generative [32] models) that analyze functional and/or structural connectivity have largely focused on the identification and characterization of subsystems [33, 34] in the brain. A well known example of such a subsystem is the Default Mode Network (DMN) [35], which is ubiquitous across findings within network neuroscience studies. While these frameworks lay the foundation for modular characterizations of brain connectivity, they focus primarily on isolating group-level effects. Even studies that incorporate variability among individuals within the population [36] exhibit little generalization onto new subjects.

In a bid to expand the horizons of neuroscientific discovery, data-driven methods have been gaining prominence in connectomics [37, 38]. Often, these studies aim to automatically underscore connectivity patterns that are informative of differences between diseased and healthy cohorts. Owing to the high data dimensionality, these often follow a two stage *Feature Selection*

$\rightarrow$ *Prediction* pipeline. Common algorithms include Principal Component Analysis (PCA) and Independent Component Analysis (ICA) that concentrate connectivity information into a small set of canonical bases by leveraging key statistical properties of the feature distribution. While highly informative at diagnostic classification, group level confounds often overwhelm inter-individual variability within these models [39]. As a result, they exhibit limited generalization for predicting finer grained patient characteristics. This divide has been partially bridged by end-to-end deep learning models.

Deep learning is becoming ubiquitous for connectome analyses due to its ability to automatically extract complex representations from data that are simultaneously meaningful for a downstream prediction task. Neural Networks have the ability to efficiently learn abstractions of the input data without requiring careful feature engineering [40]. Consequently, simple models such as Multi-Layered Perceptrons (MLP), Convolutional Neural Networks (CNN) and Graph Neural Networks (GNN) have been applied for case/controls classification of neuropsychiatric disorders such as schizophrenia [41] or Autism Spectrum Disorder [42, 43]. While few models consider fine-grained prediction of clinical measures [44], a downside to these models is that the learned representations may be hard to directly interpret. In addition, they often require large amounts of training data for adequate generalization, which may not be the case with clinical neuroimaging studies.

1.1.2 Predicting Clinical Severity from Brain Connectivity

Several previously introduced frameworks have been demonstrated varying levels of success at behavioral characterization, particularly at case/control diagnosis. An important point to note is that the unification of diverse individuals under a single diagnostic umbrella may not always provide a comprehensive clinical picture. For example, differences among patients in symptomatic manifestation of complex disorders may be subtle and thus often ignored under a strict case/control distinction. These individual-specific differences tend to be subtle and often overwhelmed by group level confounds. As a result, the characterization finer-grained measures of clinical severity in the literature continues to remain an open challenge.

Modeling this dichotomy between group-level effects and subject-specific differences from connectivity data (See Fig. 1.2) is a key motivation for the frameworks introduced in this thesis.

1.1.3 Representation Learning for Connectomics

We borrow from network decomposition strategies in classical representation learning. A key advantage of this approach lies in the interpretability, as opposed to the black box nature of deep models. Typically, network decomposition models are mathematically designed to tease apart the shared structure $\mathbf{S}$ within the rs-fMRI data $\{\mathbf{X}_n\}$ from individual specific effects $\mathbf{M}_n$.

The constituent components or factors are unknowns that map to the data via a function $\mathcal{F}(\cdot)$ with a parametric form. In the simplest case, $\mathcal{F}(\cdot)$ is a matrix product. To estimate the factors, a least squares reconstruction error is

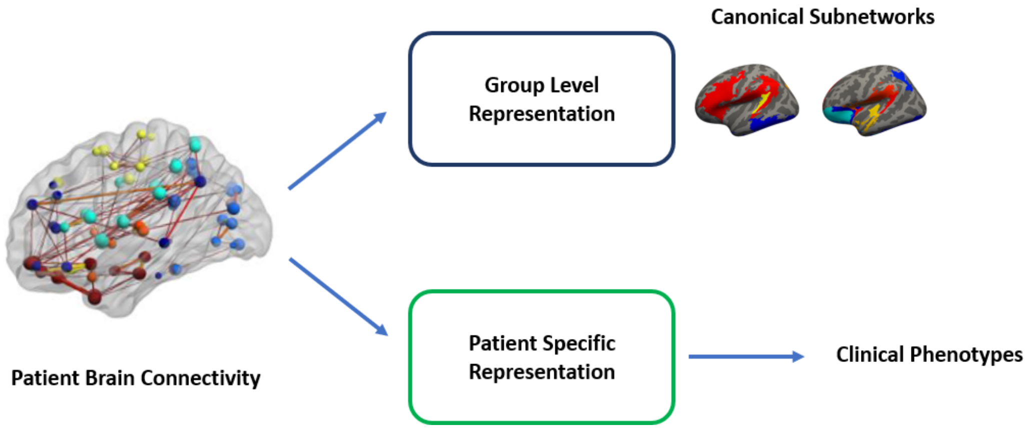

Canonical Subnetworks
Group Level Representation
Patient Brain Connectivity
Patient Specific Representation
Clinical Phenotypes

Figure 1.2: Decomposition of Brain Connectivity to tease apart group level and patient specific information

optimized:

$$\mathcal{L}_{BSS} = \sum_{n} ||\mathbf{X}_n - \mathbf{S}\mathbf{M}_n||_F^2 \tag{1.1}$$

Such algorithms are often referred to as Blind Source Separation (BSS) as they decompose the signal into constituent components that are apriori unknown. To obtain a joint solution in the individual factors, an alternating minimization [45] procedure may be utilized. Essentially, this procedure alternates through the estimation of the individual factors one at a time, fixing the estimates of the other unknowns.

A common example of a BSS algorithm is Independent Component Analysis (ICA), which has been used to pinpoint spatial [46] and temporal specialization [47], perform multi-subject analysis [48] and group comparisons [49]. ICA enforces a notion of statistical independence between the constituent signal components of the raw time series data, which translates to an explicit constraint within the optimization. While useful, strict notions of spatial or temporal independence may be too restrictive in practice, especially in the presence of considerable noise in the time-series data.

As an alternative, we choose to build generative frameworks that leverage the underlying structure within functional connectivity matrices rather than the rs-fMRI time series. Our strategy is inspired by [50], which decomposes connectivity matrices into group level and patient-specific components modeled as canonical outer product factors. These rank one components effectively leverage the low rank geometric matrix structure within the data. As opposed an unsupervised decomposition [50], our goal is to effectively explain variability among patients in the related behavioral space (clinical measures). In essence, we explicitly link the neuroimaging and behavioral spaces in the form a predictive regression model. Our model is evaluated on generalizability onto unseen subjects, rather than data-fit. Effectively, this procedure reliably uncovers neural signatures that are informative of behavioral deficits in clinical populations. Fig. 1.2 pictorially illustrates this framework.

1.2 Application to Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects an estimated 1 in 68 children in the United States. Young adults and children afflicted with ASD often face considerable social, communication and behavioral challenges, leading to significant personal and societal costs [51].

Neurologically, ASD is believed to result in aberrant inter-regional communication within the brain via impaired structural pathways [52, 53] and functional signalling [54, 55]. In fact, rather than being attributed to a single unified dysfunction, ASD is known to reflect distributed impairments across several regional networks [56, 57]. Subnetworks within the brain may be thought of as a collection of communicating regions that are associated with a specific function, for example, the visual network. Identifying such key subnetworks associated with ASD is a key link to better understanding the social and behavioral implications of the disorder.

From a behavioral standpoint, patient variability in ASD manifests as a spectrum of impairments. Typically these measures are quantified by a "behavioral score" of clinical severity that is obtained from an expert assessment. Behavioral phenotypes of ASD include communicative deficits, social and emotional reciprocity, motor impairments etc [58]. Moreover, the manifestation of these symptoms across patients is known to exhibit acute heterogeneity within a single diagnostic umbrella. These caveats render the problems of uncovering the pathogenesis of ASD and designing directions for treatment particularly challenging.

1.2.1 Multidimensional Clinical Characterizations

At the same time, there is a growing consensus in clinical psychiatry that complex disorders, such as autism and schizophrenia, are inherently multidimensional [59]. Furthermore, there is considerable patient heterogeneity within a single diagnostic umbrella that reflects subtle differences in the underlying etiology [60]. In fact, the National Institute of Mental Health (NIMH) in the United States has released the RDoc research framework [61], which advocates for a multidimensional characterization to understand the full spectrum of mental health and illness. Keeping this in mind, our generative frameworks aim to uncover connectivity signatures that in turn can explain a spectrum of diverse impairments under the ASD umbrella.

As a first exploratory step in this direction, our frameworks have a central focus on the simultaneously prediction of multiple deficits or multiscore prediction. This is a challenging and largely uncharted clinical paradigm in data-driven connectivity analysis.

1.3 Summary

Our goal is to propose flexible frameworks that are capable of representing the information within brain connectivity data effectively. We aim to leverage the extracted representations in order to map to clinical and behavioral characterizations. Through this exercise, we seek to better understand the brain and its relationship to diagnostic and clinical information.

1.3.1 Our Contributions

The main contributions of this work are four-fold:

- From an application standpoint, we build up a unified framework to integrate structural (DTI) and (dynamic) rs-fMRI connectivity together to map to behavior.
- From a clinical standpoint, our frameworks provide us with the flexibility to simultaneously explain multidimensional clinical characterizations in Autism.

- From a technical standpoint, we propose unique alternatives to black-box models (eg. end-to-end deep networks) by combining the interpretability of classical techniques with the representational power of strategically-designed deep neural networks.

- Using these principles, we develop end-to-end geometric models that probe the relationship between the complementary connectivity spaces, i.e. function and structure beyond simple phenotypic prediction.

In summary, the aforementioned frameworks carefully balance generalizability with interpretability, thus bridging the representational gap between structure, function and behavior. Additionally, using both experimental evidence and preliminary analytical insights, we demonstrate how our geometric frameworks are a first step to extracting powerful canonical connectivity representations.

1.3.2 Thesis Outline

To set the stage for subsequent chapters, Chapter 2 provides us with background information on our problem of interest and relevant literature. Chapter 3 introduces a joint network optimization model (JNO) that predicts clinical severity from rs-fMRI correlation matrices by combining a generative matrix factorization with a discriminative regression model. Chapter 4 focuses on extends the discriminative model to combine non-parametric regression and neural network predictors with the generative model. Chapter 5 extends the generative framework to model incorporate and multimodal connectivity and complement the discriminative frameworks to predict multiple clinical characterizations.

Chapter 6 discusses a technical refinement that marries classical representation learning with the simplicity of end-to-end stochastic optimization. It introduces a geometric framework (matrix autoencoder) that can robustly predict structural connectivity from functional connectivity while being guided by a secondary phenotypic prediction task. We also present some preliminary analytical and experimental results that probe the representational aspects of this framework.

Complementary to these mathematical models, Chapter 8 takes an alternate end-to-end strategy of fusing function and structure for phenotypic prediction. Specifically, this chapter explores convolutional models that treat the brain as an interconnected modular entity with rich topological properties.

Chapter 2

Background

Functional and structural neuroimaging modalities provide complementary viewpoints of brain connectivity. Therefore, in this chapter, we first give a brief overview on the acquisition of the data via the two neuroimaging modalities. We focus on the evolution of brain connectivity analysis. We describe traditional models that largely rely on domain knowledge to the recent wave of machine learning models that seek to automatically integrate multimodal information in a data-driven fashion. We then describe our Autism dataset to may the groundwork for our clinical problem of interest. Finally, we conclude this chapter with an overview of the notation for the rest of the thesis.

2.1 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive neuroimaging modality that measures changes in blood oxygenation within the brain. Haemoglobin in the blood is diamagnetic when oxygenated, in contrast with de-oxygenated haemoglobin, which is paramagnetic. Blood Oxygen Level

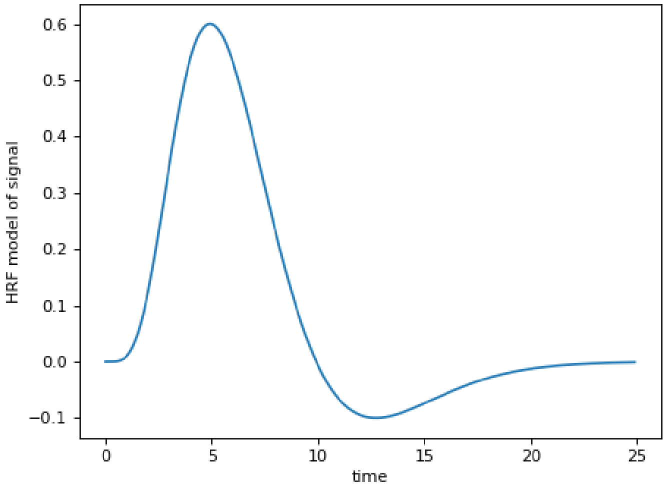


Figure 2.1: Haemodynamic Response Function Profile

Dependent (BOLD) fMRI uses a T2*-weighted protocol to measures localized changes in oxygenation over the course of the scan. Specifically, in oxygen-rich regions, the T2* relaxes slowing and results in higher signal intensity [62]. The temporal resolution of fMRI signals is limited (1-5 seconds between volumes) despite the reasonable spatial resolution (2-5mm^3). Representation wise, fMRI signals are $4-D$ with the first three dimension representing the voxel location and the fourth dimension denoting progression of time samples over the scan.

fMRI relies on the fact that neuronal activation within the brain and blood flow are coupled. When a certain region of the brain is in use, blood flow to that region also increases. It is hypothesized that the regions of the brain exhibiting increased local blood flow (and thus oxygen metabolism) are likely linked to heightened energy utilization during neurobiological processes [63]. Unfortunately, the exact relationship between the underlying neural signals and haemodynamic response is ill understood.

2.1.1 Localizing Functional Responses

Traditionally, task-evoked fMRI has been widely employed in order to localize brain regions that are involved in functional specialization. The task or event paradigm needs to be carefully designed to evoke and isolate activation responses from regions of interest. Such task-based data is typically analyzed via a Generalized Linear Model (GLM) [64]. GLMs inherently assume that each individual experimental condition elicits a linear contribution to the overall fMRI response. Let $\mathbf{z}_i \in \mathcal{R}^{T\times 1}$ represent the fMRI activation time course at the spatial location i. The experimental paradigm is encoded in a temporal design matrix $\mathbf{X} \in \mathcal{R}^{T\times M}$. The relationship between neuronal activity and the fMRI signal takes the form of a parametric transfer function. This hemodynamic response function (HRF) models neurovascular coupling and is typically convolved with the experimental protocol in order to obtain the columns of the design matrix $\mathbf{X}$. Fig. 2.1 illustrates the HRF profile. GLMs pose the following regression

$$z_i = \mathbf{X}\beta_i + \epsilon_i \tag{2.1}$$

Here, $\beta_i \in \mathcal{R}^{M\times 1}$ is an activation vector that we wish to estimate. It denotes the response to each stimulus. Finally, we assume that the corruptions ϵ_i arise as additive white Gaussian noise. Mathematically, we can solve for $\boldsymbol{\beta}$ using the least squares solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. To interpret these results, a high-valued entry in $\hat{\boldsymbol{\beta}}_i$ indicates that region i has a strong response to a particular stimulus, thus informing us of the role played by region i within the brain.

2.1.2 Exploring Functional Connectivity at Rest

On the other hand, resting state fMRI is acquired in the absence of a task paradigm to capture steady-state patterns of co-activation within brain regions. Participants are typically instructed to lie still within the scanner while being imaged at rest. While we no longer have an experimental protocol to infer the GLM activation responses from, we can instead rely on uncovering and analyzing the co-activation across pairs of regional responses. Further, it is believed that these correlation patterns within these signals illuminate the intrinsic communication between brain regions [7].

Thus, functional co-activation patterns at rest offer us insight into the "functional connectivity" (i.e. functional relationships) with brain regions. Rs-fMRI thus obviates the need for compliance with a strict task protocol. This makes it an attractive option for clinical studies, particularly in case of atypical or pediatric populations. Primary examples of such studies include characterization of neuropsychiatric disorders such as Autism Spectrum Disorder (ASD) [65], Attention Deficit Hyperactivity Disorder (ADHD) [66], and schizophrenia [67], development and assessment of behavioral therapy [68], pre-surgical planning [69] etc.

With this brief introduction, the rest of this thesis will utilize resting state fMRI acquisitions for analysis. From a clinical standpoint, we are interested in how functional connectivity relates to behavior, and in exploring the link between the functional organization of the brain and manifestation of impairments in patient cohorts.

2.2 Analysis of Functional Connectivity

Traditional rs-fMRI analysis has concentrated on comparing the statistics of the rs-fMRI data, or variation in these statistics, across individuals or between different cohorts. For example, statistical differences in rs-fMRI features between a patient cohort and neurotypical controls may be considered as biomarkers of a given disorder. However, the high dimensionality of rs-fMRI data, along with the considerable inter-patient variability, make it extremely difficult to reliably predict clinical manifestations on a individual level.

2.2.1 Exploring Functional Concordance

Rs-fMRI studies have uncovered the presence of spontaneous fluctuations within regions of the brain typically concentrated within a frequency band (< 0.08Hz) [8]. Despite the lack of external stimuli, these response signals have been found to are strongly correlated across multiple brain structures across individuals. Analysis of functional connectivity patterns aim to pinpoint and study the coherence within these response to improve the understanding of the brain and its organization.

Perhaps the earliest approach for isolating functional systems was seed-based correlation analysis, in which the functional connectivity of specific seed regions to the rest of the brain is assessed [70] and may be compared across patient cohorts [71, 72]. The seeds are typically determined by domain knowledge and fixed *a priori*. They are typically of the order of $3 - 5$ voxels in diameter and are embedded within the gray matter tissue. Once the expert specifies the location of the seed, one may wish to quantify its concordance with targeted and large subsystems within the brain (for example, the somatomotor or visual network). Denoting this target subsystem by a collection $\mathcal{S}$, we can compute the average regional time series associated with this subsystem as $\mathbf{z}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\mathbf{z}_j$. To quantify the concordance with the seed of interest i, we may quantify a correlation measure between the time series as:

$$\rho = \frac{\mathbf{z}_i^T\mathbf{z}_{\mathcal{S}}}{||\mathbf{z}_{\mathcal{S}}||||\mathbf{z}_i||} \tag{2.2}$$

In each case, the time series are mean centered before estimating the correlation value. Seed based analysis has been extremely useful in identifying brain systems reliably across subjects or discovering new systems, typically by thresholding correlations at a suitable value. A common example is the default mode network [63]. It is active at rest of when an individual performs an internally focused task and deactivates during sensory-motor experiments. The default mode network is believed to be critical for mediating cognitive processes about the environment and for memory related tasks [35].

Despite its success, seed based connectivity analysis is limited by the ability to specify the initial ROI. Further, the consistence of the results within populations may be heavily rely on the choice of threshold [73]. Hence, recent research has been pivoting towards leveraging statistical frameworks to instead focus on whole brain connectome-wide comparisons.

2.2.2 Statistical Approaches for Connectome Wide Comparisons

Connectome wide comparisons examine effects at the level of individual pairwise measures within the connectome. In this regard, there has been considerable work in developing statistical frameworks for performing comparisons of pairwise connectivity measures across populations. Examples include standard multivariate [74] or random effects models [75] that aim to capture population level differences in functional connectivity. Additionally, these analyses may involve a prohibitively large number of multiple comparisons, where a simple Bonferroni correction could prove to be too conservative for uncovering and studying the underlying effects. Several approaches have been proposed to address these problems [76, 77, 78, 79] with notable examples including the network based statistic and the Spatial Pairwise Clustering techniques [77, 79].

While straightforward to implement, these techniques do not adequately characterize distributed impairments across multiple brain systems. This characterization is believed to be crucial for understanding the complex pathologies associated with neuropsychiatric disorders [80, 55, 81]. This limitation has warranted the development of network-based models to study the inter and intra-subject variation across populations.

2.2.3 Network Models and Graph Theoretic Analysis

Network-based rs-fMRI studies typically group voxels in the brain into regions of interest (ROIs) using a standard anatomical or functional atlas. Further, the

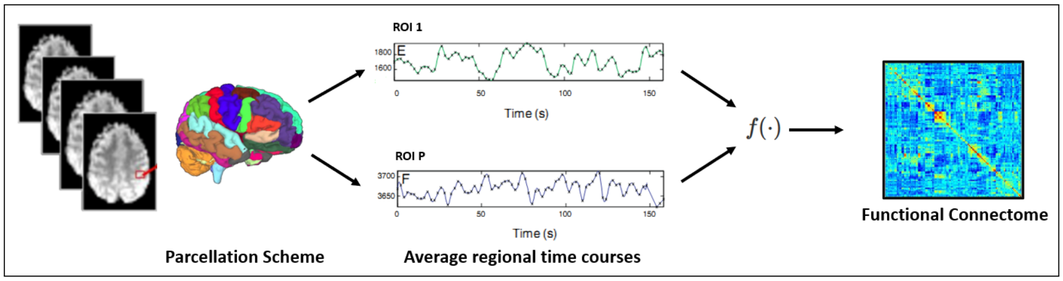


Figure 2.2: For the fMRI data, we group voxels in the brain into ROIs defined by a standard atlas and compute the average time courses for each ROI. The elements of the functional connectome capture the pairwise synchrony in these average time courses.

synchrony between the average regional time courses is summarized using a similarity matrix, which is the input for further analyses. This extraction procedure is demonstrated in Fig. 2.2. From here, we could extract global and local network properties to analyze the connectivity within the brain graph.

One may treat the brain as a complex network graph. Under this formalism, the works of [28, 29, 30] use graph theoretic notions of connectivity based on aggregate network measures, such as node degree, betweenness centrality, and eigenvector centrality to study the functional organization of the brain. These measures are extremely useful to compactly summarize the connectivity information onto a restricted set of nodes which map to brain regions.

A more global network property is small-worldedness [82], which describes an architecture of sparsely connected clusters of nodes. These characterizations are quite successful at capturing global connectivity information as well as implicating dysconnectivity within psychiatric disorders. For example, changes in small-worldedness have been implicated in many neurological disorders [83, 84] such as schizophrenia.

Constructing a comprehensive picture of brain organization requires us to simultaneously explain global as well as localized properties. Modularity analysis is the first step to examining this organization at the node level and builds up a notion of intra-modular and inter-modular connections [85]. Effectively, intra-modular connectivity measures the node's relative connectivity to other nodes within the same module, whereas inter-modular connectivity describes the way in which a node's connectivity is distributed across various modules. Together, nodes with high intra-modular connectivity (termed provincial hubs) and nodes with relatively comparable connectivity across modules (termed connector hubs) play key roles in functional specialization and functional integration respectively [86].

2.2.4 Mechanistic and Generative Models of Functional Connectivity

From the earliest work on seed-based correlation analyses, there is mounting evidence that the brain contains numerous modular sub-networks that specialize in different functionality. Within the graph based models, these sub-networks are captured as collections of densely connected nodes that interact with each other, also known as “communities".

While centrality, small-worldedness, and modularity that were introduced above have been useful in many applications, they collapse the richness of the full connectome into a few summary statistics. To address the limitations of aggregate notions, recent focus has shifted towards mechanistic network models, which are capable of incorporating hierarchy onto existing graph connectivity notions.

Community detection techniques are a class of population-level models which are used to identify highly interconnected subgraphs within a larger network. These techniques have become popular for understanding the organization of complex systems like the brain network architecture [87]. An application of this approach to identify regions having abnormal connectivity in schizophrenia patients can be found in [88]. Similarly, Bayesian community detection algorithms developed in [89] have provided valuable insights in characterizing the social and communicative deficits associated with autism. An alternative network topology is the hub-spoke model, which targets regions associated with a large number of altered connections [88, 90, 91].

Overall, these techniques have been highly successful at leveraging the underlying topology and hierarchy within brain organization for group discrimination [88] and subject-specific ROI identification [92]. However, the above methods focus on group characterizations, and even studies that consider patient variability [36] have little generalization power on new subjects.

2.2.5 Data-Driven Approaches

The functional connectivity matrices are often vectorized to convert them into connectivity features, which are the "patterns" of interest in downstream analysis. The goal of data-driven analysis is to isolate the connectivity patterns that are most discriminative across the population. However, the number of features are order of magnitudes larger than the number of subjects in the study. Thus, data-driven techniques often cast the neuroimaging prediction problem as a two stage procedure. Essentially, the first step is a feature selection or a dimensionality reduction stage, while the second stage uses the output of the first to predict the subject characteristics.

A simple representation learning framework entails a careful sub-selection of specialized biomarkers [93, 94]. On a whole brain level, data-driven approaches treat the patient connectivity information as a *feature map* and estimate lower dimensional projections, typically through PCA, kernel-PCA [95] or ICA [96]. From here, the most popular classifier (i.e. a stage two algorithm) is a Support Vector Machine (SVM) [97], which optimizes the decision boundary between patients and neurotypical controls [96]. SVMs have also been shown to identify disease sub-types [94] from the lower dimensional features with high accuracy.

While this two stage pipeline has been successful in the classification realm, characterizing finer-grained measures of clinical severity in the fMRI literature has largely been restricted to associative analysis, as opposed to an actual prediction on unseen data. For example, the work of [98] identifies key visual and motor ICA components, which are then used to compute a visuo-motor measure that is significantly correlated with social-communicative and motor deficit measures in ASD. In the context of a continuous value prediction, [99] develops a modified random forest regression algorithm for stacked multi-output score estimation from multiple ROI-voxel correlation maps.

Finally, deep learning methods have become popular for several neuroimaging data analysis. These models have the ability to efficiently learn complex abstractions of the input data without requiring careful feature engineering. As a result, they have been quite successful in a number of case/control classification tasks [43, 40, 42]. Common architectures that have been used for functional connectivity analysis include Multi-Layered Perceptrons, Convolutional Neural Networks and Graph Neural Networks. However, a downside to these models is the requirement of large amounts of training data for adequate generalization, which is rarely the case with clinical studies. Consequently, there has been limited success in predicting behavior from rs-fMRI data using neural networks.

In this thesis, we wish to isolate functional connectivity signatures that can explain variation within a clinical cohort. This variation is quantified in the clinical space, where differences among individuals may be subtle. As explained in Chapter 1, modeling group level structure coupled with patient-specific factors simultaneously is key to generalizability. Therefore, we will focus on network decomposition models that are designed specifically for functional connectivity matrices. These models will also allow us to probe canonical patterns of co-activation in the brain, offering interpretability.

2.2.6 Dynamic Functional Connectivity: Hypothesis and Models

There is now growing evidence that functional connectivity is a dynamic process that toggles between different intrinsic states evolving over a static structural connectome [100]. These states manifest over short time windows that are typically of the order of a tens of seconds to a few minutes. Several studies such as [101, 102, 103] indicate the importance of modeling this evolution for characterizing neuropsychiatric disorders such as schizophrenia and ASD.

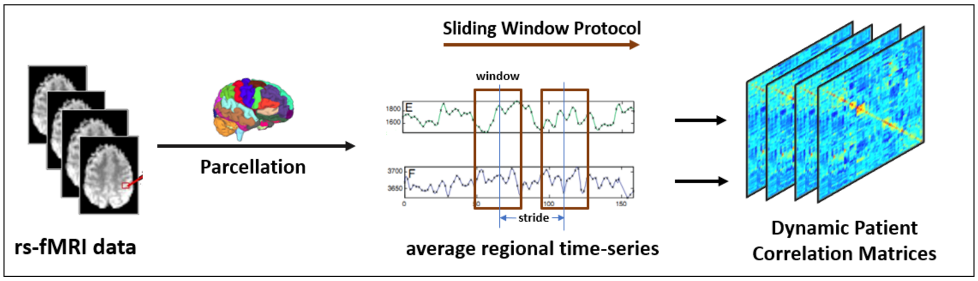


Figure 2.3: For the fMRI data, voxels in the brain are grouped into ROIs according to a standard atlas. From the average time courses for each ROI, a sliding window protocol may be used to extract time-varying functional connectivity matrices.

In the simplest case, the rs-fMRI time-series are fit to a Markov model that encodes a state transition behavior [104]. Alternatively, model based frameworks may also be used to detect dynamic changes in correlation patterns rather than working with the time series directly. Often, these measures are estimated between large-scale brain networks such as the Default Mode Network, Somatosensory Network etc. An example is the Dynamic Conditional Correlation (DCC) protocol that was initially developed in the econometrics and finance literature [105] and later adapted to the study of brain organization using rs-fMRI [106]. It poses a time-varying matrix estimation problem to explicitly model the evolution of connectivity patterns in the brain, and has shown robustness in the test-retest setting [107] with rs-fMRI. Another example is the time varying graphical lasso [103]. Unfortunately, many of these methods are unstable when scaled up [108, 109]. For example, at a whole brain ROI-level analysis of dynamic connectivity, the correlation matrices may be ill conditioned in the absence of additional regularization. Consequently, most dynamic connectivity studies continue to rely on sliding-window correlations as inputs.

In this thesis (specifically, Chapter 5), we will employ sliding window correlation patterns to capture dynamic connectivity across the scan. The sliding window protocol is defined by the window length and stride, as illustrated in Fig. 2.3. The window length defines the length of the time sequence considered by each dynamic correlation matrix, while the stride controls the overlap in successive sliding windows.

After computing the dynamic connectivity, the downstream analysis mimics the static approaches described above. For example, network decompositions may be used to isolate the intrinsic brain states. From here, typical network comparisons involve comparing key statistics of dynamic evolution, such as mean dwell time or the number of transitions between states. Differences in such statistics have been found within diseased vs neurotypical populations, acting as biomarkers of dysfunction.

In our treatment of dynamic connectivity, we pursue a dual line of inquiry. Firstly, we focus on encoding the transient brain states behavior. Essentially, we want to isolating key neural signatures in the form of subnetworks and model their individual contribution as it evolves over the course of the scan. To this end, we employ Long Short Term Memory (LSTM) networks which are capable of tracking the dynamics of the patient specific factors while implicitly enforcing temporal smoothness. The canonical bases or "states" are fixed over the scan and are shared across the cohort while their relative contribution is allowed to vary. Second, we are interested in identifying which time points of the scan are most important for clinical prediction. For this purpose, we make use of temporal attention models [110] which have been shown to be useful in a variety of domains. Overall, such carefully crafted deep learning models help us extend our representational frameworks. This in turn helps refine our understanding of functional connectivity and its relationship with behavioral characterizations.

2.3 Integration of Multimodal Connectivity Data

There is strong evidence within the neuroimaging literature of the correspondence between functional and anatomical pathways within the brain [10]. In fact, several studies suggest that the functional connectivity may be mediated by either direct or indirect anatomical connections [12, 13, 11, 14]. Diffusion Tensor Imaging (DTI) is a protocol of Magnetic Resonance Imaging which has been adopted for tracking the structural pathways within the brain.

Together, rs-fMRI and DTI data provide complementary information about function and structure respectively. Thus, it is of great interest to integrate the two views together to construct a more comprehensive picture of brain organization in health and disease. Therefore, multimodal integration has become an promising direction of study for uncovering the neurobiological underpinnings of Autism Spectrum Disorder (ASD) [111], Attention Deficit Hyperactivity Disorder (ADHD) [112], and Schizophrenia [67].

Similar to Subsection 2.2, clinical applications have motivated several key advances in the analysis of multi-modal connectivity data.

2.3.1 Diffusion Tensor Imaging

An axon or nerve fiber can be described as a long, threadlike extension of a neuron. It connects specific regions within the brain, relaying crucial information via targeted electrical impulses. The white matter within the brain is primarily composed of bundles of such myelinated nerve fibers. These fibers may be tightly packed together into fiber tracts or fiber bundles with common source and final destination.

Diffusion Tensor Imaging (DTI) characterizes the anisotropic diffusion of water as it traverses soft tissue [5]. In particular, this diffusion of water molecules happens less freely across white matter fiber bundles in the brain rather than along them. DTI protocols leverage this effect to illuminate the structural organization of the brain. From the point of view of clinical applications, DTI has been successful in discovering abnormalities within white-matter diseases such as Multiple Sclerosis [113] or tracking the progression in Alzhiemer's disease [114] or for localizing strokes [115].

A single DTI volume is obtained by applying a magnetic pulse sequence in a specific gradient direction $\mathbf{u}_k$. The resulting signal intensity $\mathbf{I}_k$ at each voxel is then given by:

$$\mathbf{I}_k = \mathbf{I}_0 \exp^{-b\mathbf{u}_k^T \mathbf{D} \mathbf{u}_k} \tag{2.3}$$

according to the Stejskal-Tanner equations [116]. $\mathbf{I}_0$ is the intensity at the corresponding voxel with no gradient pulse is applied. b is the b-value which can be pre-calculated according to the timing, amplitude and shape of the gradient pulse and is typically constant for the complete acquisition. Finally, $\mathbf{D}$ is the diffusion tensor and is symmetric and postitive semi-definite. It characterizes the directional mobility of the water molecule. $\mathbf{I}_0$ and $\mathbf{D}$ are voxel-specific .

2.3.1.1 Quantifying Anatomical Connectivity

By collecting several images from unique gradient directions, one can estimate the pointwise diffusion tensor entries $\mathbf{D}_{ij}$. Typically, DTI scans may use six or more such gradient directions to derive directional information on the underlying neuronal fiber bundles. From the diffusion tensor $\mathbf{D}$, scalar measures of anatomical connectivity may then be computed and compared across individuals. Typically, these measures are computed at the voxel level from the eigenvalues of the diffusion tensor $\mathbf{D}$, $\{\lambda_1, \lambda_2, \lambda_3\}$. A simple example is the Mean Diffusivity statistic which computes the arithmetic mean $\bar{\lambda} = \lambda_1 + \lambda_2 + \lambda_2/3$ of the eigenvalues. A more sensitive measure is the Fractional Anisotropy (FA), that describes the degree of anisotropy of a diffusion process. This is mathematically computed as

$$FA = \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \bar{\lambda})^2 + (\lambda_2 - \bar{\lambda})^2 + (\lambda_3 - \bar{\lambda})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}} \tag{2.4}$$

A number of biological processes are believed to contribute to changes in FA. For example, local FA values have been found to be significantly affected due to the inflammation of underlying white matter fiber bundles, as well as by changes in myelination [117]. Differences in the statistics of the FA value distribution have been found to be associated with neurological disorders such as Schizophrenia [118].

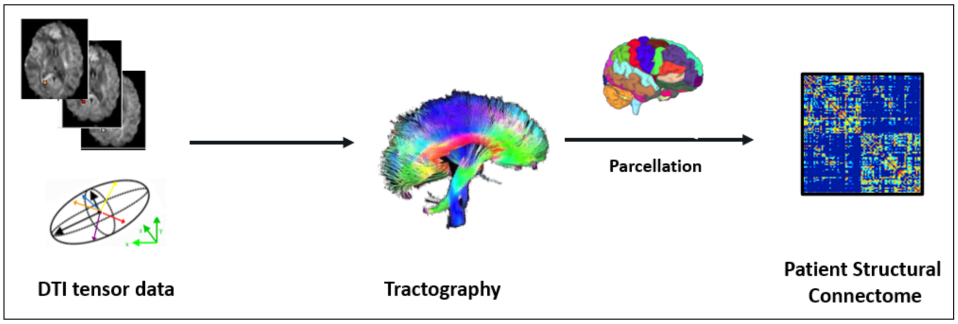


Figure 2.4: Tractography is performed on the raw DWI data to track the path of neuronal fibers in the brain. Based on the parcellation scheme, we construct a map of the fibre tracts between ROIs in the brain.

2.3.1.2 Exploring and Tracking Structural Pathways

Going one step further, we can use fiber tracking algorithms (i.e. tractography) to construct detailed 3D maps of anatomical pathways within the brain based on the diffusion tensors. This procedure is laid out in Fig. 2.4 There are two broad categories of algorithms used for tractography, namely deterministic and probabilistic approaches.

Streamline fiber tractography [119] is deterministic in the sense that it assumes each voxel to be characterized by a single primary fiber orientation as dictated by the dominant eigenvalue and eigenvector of $\mathbf{D}$ subject to additional smoothness constraints. It then infers global fiber trajectories by piecing together the local orientations. Mathematically, one may consider the set of (local) fiber orientations as a 3D vector field. A streamline can be defined as a curve that is tangent to the vector field along its trajectory. Thus, the global fiber trajectories are the streamlines. A white matter tract is the the molecule follows in this vector field from a fixed starting location referred to as the "seed".

Often, the local fiber orientation estimates from tractography may be subject to errors. Here, local imaging noise, artifacts, as well as inaccuracies in the streamline integration and local modelling errors contribute to the final global fiber track estimation. Particularly in steamline tractography, measurement uncertainty may propagate and compound in effect. Probabilistic tractography [119] algorithms aim to directly characterize this uncertainty, by generating a large collection of possible trajectories from each seed point. This effectively characterizes a "distribution" over possible fiber tracts rather than a fixed trajectory. Probabilistic tractography algorithms build upon the streamline ones, with a key difference being that the orientations for propagation of the tract are drawn at random from a orientation distribution function (ODF) defined locally. Brain regions with resulting trajectories of higher density are deemed to have a high probability of anatomical connection with the seed point. This effectively allows fiber tracking to continue in regions with high uncertainty, which may often be missed within streamline algorithms.

Given the results of tractography, one may wish to quantify anatomical connectivity along the estimated white matter pathways using various measures. For example, one may compute the probability of diffusion between two brain regions, or the number of fibers linking the regions, or the mean Functional Anisotropy (FA) along the tracts connecting them.

Traditional streamline tractography algorithms often suffer from issues with tracking multiple fibers. Particularly, the estimated diffusion tensor may be nearly isotropic even in the case where two fibers cross or merge. Consequently, the algorithm may follow an incorrect trajectory. For our in house dataset, we utilize a probabilistic tracking framework designed to handle multiple fiber orientations [120]. This utilizes a relevance determination mechanism such that at crossing regions, the algorithm maintains the orientation of the streamline while still tracking non-dominant pathways. [120] demonstrate that this offers better sensitivity in tracking non-dominant fibres, while avoiding significant changes to results within the dominant pathways.

Our frameworks utilize the results of fiber tractography to convert the tracking information into a symmetric region-to-region graph of anatomical connectivity $\mathbf{A}$. Each element of the graph $\mathbf{A}[i, j]$ indicates the number of fiber tracts connecting two region pairs. A well known issue with probabilistic tracking is that larger white matter bundles may be favored over smaller ones. In order to avoid this, the anatomical network may be represented as a binary matrix, with value 1 indicating the presence of at least l fiber tract connecting the region pairs, where l is relatively small threshold.

2.3.2 Statistical Approaches to Joint Modeling of Function and Structure

Traditional multimodal analyses of rs-fMRI and DTI data have largely focused on post-hoc statistical comparisons of features extracted from the data. For example, simple statistical differences in rs-fMRI and DTI connectivity between subjects have been used to discover disrupted patterns of brain organization in Alzheimer's disease [23] and Progressive Supranuclear Palsy (PSP) [24].

At a population level, multivariate analysis [25, 26] or random effects models [27] have been employed to first independently compute, and then combine features from both modalities. Despite their success at biomarker discovery, these techniques often fail to generalize at a patient-specific level. Furthermore, they often ignore higher-order interactions between multiple subsystems in the brain. As alluded to earlier, this characterization is believed to be critical for improving the understanding understanding complex neuropsychiatric disorders [80, 55].

Overall, such shortcomings have pushed research in multimodal integration towards adopting a multi-modal network based treatment of connectivity. The goal is to develop techniques that are capable of simultaneously accounting for both inter-subject and intra-subject variability.

2.3.3 Graph Theoretic Analysis and Mechanistic Models for Multimodal Data

Similar to the functional connectome, the structural connectivity matrix derived from tractography captures the strength of the pairwise anatomical connection between different ROIs, as seen in Fig. 2.4.

Some of the simplest approaches to analyzing network properties are rooted in the field of graph theory. For example, the works of [28, 29, 30] use aggregate network measures such as centrality or small worldedness to study the organization of the brain. Complementary changes in small-worldedness in both anatomical and functional networks have been well documented across the literature [121, 122], with concurrent disruptions of functional networks [123] or structural networks [124] implicated in neuropsychiatric disorders such as schizophrenia. One of the main limitations of these approaches is that they independently analyze the fMRI and DTI data, and as such, draw heuristic conclusions about the relationship between the two modalities.

Building on the community models described in Subsection 2.2.4 which exclusively focus on functional connectivity, a natural direction would be to incorporate structural connectivity as prior information guiding the inference. In this light, the work of [125] proposes a probabilistic framework that jointly models latent anatomical and functional connectivity to discover population-level differences in schizophrenia. Similarly, the work of [126] uses a unified Bayesian framework to identify gender-differences in multimodal connectivity patterns across different age groups.

While successful at combining multi-modal information for group differentiation, these techniques do not directly address inter-individual variability.

2.3.4 Data-Driven Multimodal Integration

Data-driven methods integrating structural and functional connectivity have also focused heavily on groupwise discrimination from the static connectomes. These methods also usually follow a two-step approach where feature selectors and discriminators are trained sequentially in a pipeline. For example, the authors in [127] combine graph theoretic features computed from rs-fMRI and DTI graphs with Support Vector Machines (SVMs) to identify individuals with Mild Cognitive Impairment. Another example is the work of [128], which employs a pipeline consisting of joint-Independent Component Analysis (j-ICA) on the two modalities followed by Canonical Correlation Analysis (CCA) to combine them and distinguish schizophrenia patients from controls.

In contrast to the pipelined approaches, end-to-end deep learning methods combining feature selection and prediction are becoming ubiquitous in multimodal neuroimaging studies. These are highly successful due to their ability to learn complex abstractions directly from input data. As an example, the work of [129] uses a Deep Belief Network (DBN) on multimodal data to disambiguate patients with ASD from healthy controls.

As mentioned previously, two main limitations of utilizing deep learning are the requirement of large amounts of data for generalization, and the lack of insight they provide into interpretation. To circumvent these issues, our frameworks perform multimodal integration using regularized generative models instead of end-to-end networks. While the frameworks are designed to incorporate the dichotomy between intra-subject and inter-subject differences, they also allow us to probe the signatures of brain connectivity that are representative across the population and predictive of clinical information.

2.4 Dataset: Acquisition and Preprocessing

In this section we outline the acquisition protocol used to collect the data and the subsequent pre-processing steps.

Our primary clinical dataset consists of 58 children with high functioning Autism Spectrum Disorder (ASD) acquired at the Kennedy Krieger Institute in Baltimore, USA. Henceforth, we refer to this as the KKI dataset. The age of the subjects from this cohort is 10.06 ± 1.26 with an IQ of 110 ± 14.03.

Social and communicative deficits in ASD are believed to arise from aberrant interactions between regions of the brain that are linked by structural and functional connectivity [130]. Thus, identifying these patterns plays a crucial role in better understanding the disorder and is a key motivating application to developing our frameworks.

2.4.1 Neuroimaging Data

2.4.1.1 rs-fMRI Acquisition and Pre-processing

rs-fMRI acquisition was performed on a Phillips $3T$ Achieva scanner with a single shot, partially parallel gradient-recalled EPI sequence with TR/TE = 2500/30ms, flip angle 70°, res $= 3.05 \times 3.15 \times 3$mm, having 128 or 156 time samples. The children were instructed to relax with eyes open and focus on a central cross-hair while remaining still. We used an in-house pre-processing pipeline pre-validated across several studies [98, 36]. This consists of slice time correction, rigid body realignment, and normalization to the EPI version of the MNI template using SPM [131], followed by temporal detrending of the time courses to remove gradual trends in the data. A CompCorr50 [132, 133] strategy was used to estimate and remove spatially coherent noise from the white matter and ventricles, along with the linearly detrended versions of the six rigid body realignment parameters and their first derivatives, followed by spatial smoothing using a 6mm FWHM Gaussian kernel and temporal smoothing via a band pass filter ($0.01 - 0.1$Hz). Lastly, the data was despiked using the AFNI package [134].

2.4.1.2 DTI Acquisition and Pre-processing

The DTI acquisition for the KKI dataset was collected on a 3T Philips scanner (EPI, SENSE factor= 2.5, TR= 6.356s, TE= $75ms$, res = $0.8 \times 0.8 \times 2.2$mm, and FOV= 212). We collected two identical runs, each with a single b0 and 32 non-collinear gradient directions at $b = 700s/mm^2$. The data was pre-processed using the standard FDT [135] pipeline in FSL consisting of susceptibility distortion correction, followed by corrections for eddy currents, motion and outliers. From here, tensor model fitting was performed to generate the transformation matrices and extract atlas based metrics. We used the BEDPOSTx tool in FSL [120] to perform a bayesian estimation of the diffusion parameters at each voxel, followed by tractography using PROBTRACKx [120].

2.4.2 Phenotypic Data

We analyzed three independent measures of clinical severity for the KKI dataset. These include:

1 Autism Diagnostic Observation Schedule, V. 2 (ADOS-2) total raw score

2 Social Responsiveness Scale (SRS) total raw score

3 Praxis total percent correct score

The ADOS consists of several sub-scores which quantify the social- communicative deficits in individuals along with the restrictive/repetitive behaviors [136]. The test evaluates the child against a set of guidelines and is administered by a trained clinician. We compute the total score by adding the individual sub-scores. The dynamic range for ADOS is between $0 - 30$, with higher score indicating greater impairment.

The SRS scale quantifies the level of social responsiveness of a subject [137]. Typically, these attributes are scored by parent/care-giver or teacher who completes a standardized questionnaire that assess various aspects of the child's behavior. Consequently, SRS reporting tends to be more variable across subjects, as compared to ADOS, since the responses are heavily biased by the parent/teacher attitudes. The SRS dynamic range is between $70 - 200$ for ASD subjects, with higher values corresponding to higher severity in terms of social responsiveness.

Finally, Praxis is assessed using the Florida Apraxia Battery (modified for children) [138]. It assesses the ability to perform skilled motor gestures on command, by imitation, and with actual tool use. Several studies [138, 139, 140, 98] reveal that children with ASD show marked impairments in Praxis a.k.a., developmental dyspraxia, and that impaired Praxis correlates with impairments in core autism social-communicative and behavioral features. Performance is videotaped and later scored by two trained research-reliable raters, with total percent correctly performed gestures as the dependent variable of interest. Scores therefore range from $0 - 100$, with higher scores indicating better Praxis performance.

2.5 Preliminaries

For the rest of the thesis, we follow a notational convention where we denote matrices by bold capital letters and vectors by bold lower case letters. Indices for both matrices and vectors are denoted using lower case letters.

Recall that network-based models often group voxels into regions of interest (ROIs) using a standard anatomical or functional atlas [141]. The choice of atlas specifies the “resolution" of information that will be the basis for subsequent analyses. The synchrony between representative (often average) regional time series quantifies the functional relationships between these regions.

Formally, let P be the number of regions in the parcellation, N be the number of subjects in the cohort, and T be the number of time-points in the rs-fMRI scan for subject n. We use $\mathbf{X}_n \in \mathcal{R}^{P\times T}$ to denote the collection of regional time series for subject n. The functional connectome $\mathbf{\Gamma}_n \in \mathcal{R}^{P\times P}$ is calculated as:

$$\mathbf{\Gamma}_n(i,j) = f(\mathbf{X}_n[i,:], \mathbf{X}_n[j,:]) \quad \forall\, i,j \in \{1,\ldots,P\} \tag{2.5}$$

In an abstract sense, the function $f(\cdot)$ captures the similarity between pairs of regional time series, denoted in Eq. (2.2) by $\mathbf{X}_n[i,:]$ and $\mathbf{X}_n[j,:]$. Most commonly, $f(\cdot)$ is a non-directed measure, and $\mathbf{\Gamma}_n$ is symmetric. Fig. 2.3 depicts the computation of dynamic or time-varying connectivity $\{\mathbf{\Gamma}_n^t\}$ (at time point t) using the sliding window protocol. In this specific case, instead of using the entire time series for estimating a single snapshot of functional connectivity, the scan is divided into individual short segments.

Our experiments rely on the Automatic Anatomical Labelling (AAL) atlas [142] parcellation for the rs-fMRI and DTI data. AAL consists of 116 cortical, subcortical and cerebellar regions. Thus, for each individual, we have correlation matrices of size 116 × 116 based on the Pearson's Correlation Coefficient between the average regional time-series.

$$\mathbf{\Gamma}_n[i,j] = \frac{(\mathbf{X}_n[i,:] - \bar{\mathbf{X}}_n[i,:])^T(\mathbf{X}_n[j,:] - \bar{\mathbf{X}}_n[j,:])}{||\mathbf{X}_n[i,:] - \bar{\mathbf{X}}_n[i,:]||_2||\mathbf{X}_n[j,:] - \bar{\mathbf{X}}_n[j,:]||_2} \tag{2.6}$$

Where $\bar{\mathbf{X}}_n[i,:] = \sum_k \mathbf{X}[i,k]$ is the average signal. This is a symmetric and non-directed measure of functional connectivity. Empirically, we observed a consistent noise component with nearly unchanging contribution from all brain regions and low predictive power for both datasets. Therefore, we subtracted out the first eigenvector contribution from each of the correlation matrices and used the residual matrices for each subject.

As mentioned previously, our frameworks decompose functional connectivity matrices $\mathbf{\Gamma}_n$ into a group level and patient specific representation (See Fig. 1.2). To lay the groundwork for chapters 3-5, we will briefly describe the representational setup.

We will use the matrix $\mathbf{B} \in \mathcal{R}^{P \times K}$ to denote the canonical basis. Effectively, it is a concatenation of K elemental vectors $\mathbf{b}_k \in \mathcal{R}^{P \times 1}$, i.e. $\mathbf{B} := [\mathbf{b}_1 \quad \mathbf{b}_2 \quad ... \quad \mathbf{b}_K]$. Since $K \ll P$, we effectively perform dimensionality reduction on the functional connectivity matrices. While the bases are common to all patients in the cohort, the combination of these subnetworks is unique to each patient and is captured by the non-negative coefficients $\mathbf{c}_{nk}$. These coefficients model the variability in the dataset along in the space of the elemental basis vectors.

On the behavioral side, $\mathbf{y}_n$ can be thought of as a vector with the individual scalar measures of clinical severity concatenated together. Our frameworks model the link between connectivity and behavior as a predictive regression model $g(\cdot)$ that takes in as input the patient specific coefficients. Effectively, we would like to approximate a mapping such that $g(\mathbf{c}_n) = \mathbf{y}_n$. By combining the estimation of the generative and discriminative setup, we seek representations which faithfully capture informative representations. We will evaluate the fidelity of this procedure via the predictive generalization in a cross validated setting.

Finally, each DTI connectivity matrix $\mathbf{A}_n$ is binary, where $[\mathbf{A}_n]_{ij} = 1$ corresponds to the presence of at least one tract between the regions i and j, 116 in total for AAL. For Chapters 5 and 8, we utilize this binary DTI graph as an anatomical prior on the functional connectivity. Effectively, this acts as a regularization that acts on the functional matrix decomposition. Specifically, we will use the graph laplacian regularizer $\mathbf{L}_n$ derived from $\mathbf{A}_n$ as $\mathbf{L}_n = \mathbf{V}_n^{-\frac{1}{2}}(\mathbf{V}_n - \mathbf{A}_n)\mathbf{V}_n^{-\frac{1}{2}}$. Here, $\mathbf{V}_n = \mathbf{diag}(\mathbf{A}_n\mathbf{1})$ is the degree matrix. In our experiments, we will use k-fold cross validation as the evaluation strategy for all our models and baselines. Hence, for the KKI dataset, we impute the missing DTI connectivity for the individuals, who do not have DTI based on the training data in each cross validation fold.

In Chapter 6, our treatment of structural connectivity is a bit more nuanced. Instead of only predicting pheotypes from connectivity, we are instead interested in how functional and structural connectivity are related to each other. Therefore, we examine whether we can reliably predict structural connectivity from functional connectivity matrices across the population. We adopt a non-binarized weighted version of $\mathbf{A}_n$. Each $\mathbf{A}_n[i,j]$ quantifies the relative strength of connection between region pairs. Mathematically, this is computed as the ratio of the number of tracts connecting the pairs to the total number of estimated tracts. Further, this normalization induces the property that $||\mathbf{A}||_1 = 1$, which defines the geometry of the structural connectivity space. We will describe in Chapter 6 how we design our framework to make use of these properties within the learned mapping.

2.6 Summary

To summarize, prior work on brain connectivity and behavior has employed a wide toolkit using simple data statistics, graph theoretic approaches, and machine learning to uncover patterns of interaction between structure and function. Several frameworks use specialized feature extraction on the rs-fMRI and DTI, then draw correspondences between them post-hoc. In addition, several techniques take a two-step approach, which effectively decouples the feature extraction from the downstream prediction/classification. From an application standpoint, several studies focus exclusively on group-wise discrimination, and do not map to more fine-grained continuous clinical measures that are important for characterizing neuropsychiatric disorders. Finally, several machine/deep learning frameworks vectorize features from the connectomes at the input, due to which they do not fully exploit the structure within the connectomes themselves.

In our work, we leverage techniques from optimization, deep learning, geometric modeling to address the aforementioned issues from a modeling standpoint. Subsequent chapters will detail how our research departs from prior work and advances the field of brain connectivity analysis for clinical applications.

Chapter 3

JNO: A Joint Network Optimization Framework for Functional Connectomics and Clinical Severity

As discussed in Chapter 2, understanding the relationship between the functional connectivity and behavioral data spaces can offer key insights into characterizing complex neuropsychiatric disorders. However, most prior work in this application area examines a case/control classification problem. A typical pipeline has two stages, a feature extraction (from statistics, graph theory, machine learning) followed by a discriminative model. Characterizing finer-grained measures of clinical severity in the fMRI literature has been restricted to associative analysis, as opposed to an actual prediction on unseen data.

Dictionary learning [143, 50] methods move away from the pipelined representations, and have recently gained traction due to their ability to simultaneously model both group level and patient specific information. The work of [144] proposed a correlation matrix decomposition strategy, in which, multiple rank one outer products capture an underlying *generative* basis. The sparse basis representation identifies meaningful co-activation patterns common to all the patients, while patient-specific coefficients combine the subnetworks and model the individual variability in the dataset. An extension of their work [145] looks at classification of young adults versus children, again, by the addition of an SVM like hinge loss. Our work builds on this representation by using the *discriminative* nature of these coefficients to predict their clinical severity via a linear regression penalty.

This Joint Network Optimization (JNO) framework combines both a generative and discriminative term, as opposed to a pipelined hyperparameter search. We employ an alternating minimization strategy to jointly infer the set of bases, coefficients and regression weights that best explain the data. The generalizability of the model is indicated by the regression performance on unseen data, instead of the correlation fit as used in [144]. This refinement demonstrates the potential of our JNO framework in identifying patient-predictive biomarkers of a given disorder.

Outline: The work presented in this chapter appeared in [146, 39]. The rest of the chapter is organized as follows. Section 3.1 introduces our generative model, while Section 3.2 describes the discriminative model to map to clinical severity. Sections 3.5 and 3.4 describes our joint objective and the alternating minimization scheme for inference. Sections 3.5, 3.6, and 3.7 present empirical validation for our framework including experiments on synthetic and real world data. Finally, Section 3.8 and 3.9 discuss the clinical significance, robustness of the framework under different settings, and the advantages and limitations.

Fig. 3.1 presents a graphical overview of our model. The two inputs to our model are the rs-fMRI similarity matrices (upper left) and the scalar clinical severity scores for each patient (lower right). As mentioned earlier, Fig. 2.2 illustrates the construction of the similarity matrix from the data. These matrices quantify the Pearson's Correlation Coefficient between the average time courses for each region of interest (ROI). The clinical scores are obtained from an expert evaluation and quantify the severity of the symptoms for the individual.

Notice that the correlation matrices in Fig. 3.1 have a dual representation.

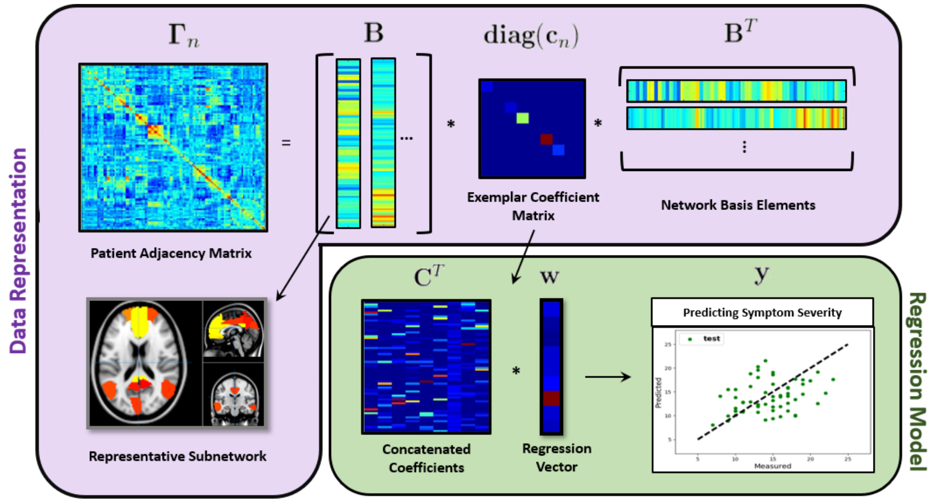


Figure 3.1: A two level joint model for connectivity and prediction. **Purple Box:** Depicts the functional data representation or 'generative' term. The correlation matrix is decomposed into a group basis term and a patient specific coefficient term. The columns of the basis matrix correspond to individual subnetworks when projected onto the brain. We stack these coefficients into a matrix. **Green Box:** Prediction of symptom severity via linear regression

The generative part of the model is indicated in the purple box. Here, we decompose the correlation matrix into a basis term and a patient coefficient term. The columns of the basis capture ROI co-activation patterns common to the entire cohort, while the coefficients differ across patients and quantify the strength of each basis column in the matrix representation. The green box indicates the discriminative part of the model. Here, we leverage the information from the patient-specific coefficients to estimate a given measure of clinical severity via a linear regression model for each individual.

3.1 Generative Model for Functional Connectomics

We define $\mathbf{\Gamma}_n \in \mathcal{R}^{P \times P}$ as the correlation matrix for patient n, where P is the number of regions given by the parcellation. As seen in Fig. 3.1, we model $\mathbf{\Gamma}_n$ using a group average basis representation and a patient-specific network strength term. The matrix $\mathbf{B} \in \mathcal{R}^{P \times K}$ is a concatenation of K elemental bases vectors $\mathbf{b}_k \in \mathcal{R}^{P \times 1}$, i.e. $\mathbf{B} := [\mathbf{b}_1 \quad \mathbf{b}_2 \quad ... \quad \mathbf{b}_K]$, where $K \ll P$. These bases capture steady state patterns of co-activation across regions in the brain. While the bases are common to all patients in the cohort, the combination of these subnetworks is unique to each patient and is captured by the non-negative coefficients $\mathbf{c}_{nk}$. We include a non-negativity constraint $\mathbf{c}_{nk} \geq 0$ on the coefficients to preserve the positive semi-definite structure of the correlation matrices $\{\mathbf{\Gamma}_n\}$. Our complete rs-fMRI data representation is:

$$\mathbf{\Gamma}_n \approx \sum_k \mathbf{c}_{nk}\mathbf{b}_k\mathbf{b}_k^T \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \tag{3.1}$$

As seen in Eq. (3.1), we model the heterogeneity in the cohort using a patient specific term in the form of $\mathbf{c}_n := [\mathbf{c}_{n1} \quad ... \quad \mathbf{c}_{nK}]^T \in \mathcal{R}^{K\times 1}$. Taking $\mathbf{diag}(\mathbf{c}_n)$ to be a diagonal matrix with the K patient coefficients on the diagonal and off-diagonal terms set to zero, Eq. (3.1) can be re-written in matrix form as follows:

$$\mathbf{\Gamma}_n \approx \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \tag{3.2}$$

Overall, this formulation strategically reduces the dimensionality of the data, while providing a patient level description of the correlation matrices.

3.2 Discriminative Model for Clinical Severity

As shown in the green box of Fig. 3.1, the patient coefficients $\{\mathbf{c}_{nk}\}$ from the representation term, are used to model the clinical severity score $\mathbf{y}_n$ using a linear regression vector $\mathbf{w} \in \mathcal{R}^{K\times 1}$

$$\mathbf{y}_n \approx \mathbf{c}_n^T\mathbf{w} \tag{3.3}$$

Concatenating the vectors $\mathbf{c}_n$ into a matrix $\mathbf{C} := [\mathbf{c}_1 \quad ... \quad \mathbf{c}_N] \in \mathcal{R}^{K\times N}$, and the severity scores into a vector $\mathbf{y} := [\mathbf{y}_1 \quad ... \quad \mathbf{y}_N]^T \in \mathcal{R}^{N\times 1}$, Eq. (3.3) can be equivalently represented in matrix form:

$$\mathbf{y} \approx \mathbf{C}^T\mathbf{w} \tag{3.4}$$

3.3 Joint Objective

We combine the two contrasting viewpoints described above into a joint objective by summing the contributions of Eq. (3.2) and Eq. (3.4) below:

$$\mathcal{J}(\mathbf{B},\mathbf{C},\mathbf{w}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 + \gamma||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2 \ \ s.t. \ \mathbf{c}_{nk} \geq 0 \quad (3.5)$$

Here, $\sum_n ||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2$ is the total error in the representation of the N patient correlation matrices, and $||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2$ is the prediction error for the behavioral data. Finally, γ is the trade-off between the rs-fMRI data-representation and score prediction terms.

3.3.1 Regularization Penalties

Since we wish to capture a compact, yet clinically informative subnetwork representations, we add an ℓ_1 penalty to encourage sparsity in **B**. Intuitively, this regularizer will sub-select a small number of nonzero entries in **B** that explain the data. From an optimization perspective, notice that scaled solution pairs $\{\mathbf{B},\mathbf{C}\}$ and $\{\alpha\mathbf{B}, \frac{1}{\alpha^2}\mathbf{C}\}$, as well as $\{\mathbf{C},\mathbf{w}\}$ and $\{\beta\mathbf{C}, \frac{1}{\beta}\mathbf{w}\}$ give rise to equivalent data representations. As a result, we introduce a quadratic penalty on **C** to act as a bound constraint. Similarly, we add an ℓ_2 regularization term to the regression vector **w** analogous to ridge regression. Mathematically, the three regularizers can be written as:

$$\lambda_1||\mathbf{B}||_1 + \lambda_2||\mathbf{C}||_F^2 + \lambda_3||\mathbf{w}||_2^2 \quad (3.6)$$

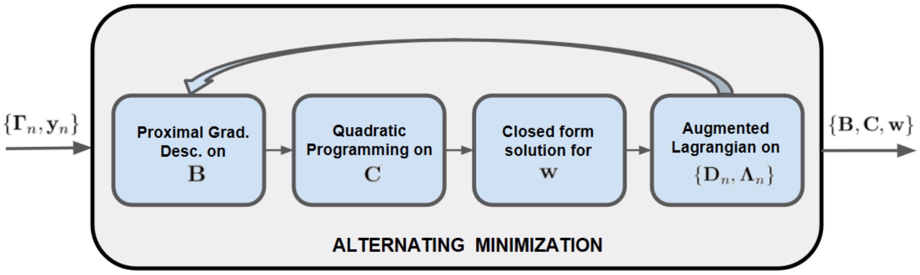

$\{\Gamma_n, \mathbf{y}_n\}$
Proximal Grad. Desc. on $\mathbf{B}$
Quadratic Programming on $\mathbf{C}$
Closed form solution for $\mathbf{w}$
Augmented Lagrangian on $\{\mathbf{D}_n, \Lambda_n\}$
$\{\mathbf{B}, \mathbf{C}, \mathbf{w}\}$
ALTERNATING MINIMIZATION

Figure 3.2: Our Optimization Strategy, we iterate through four main steps until global convergence

The penalty terms in Eq. (3.6) are added to the main objective in Eq. (3.5). The final joint objective is as follows:

$$\mathcal{J}(\mathbf{B},\mathbf{C},\mathbf{w}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 + \gamma||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2 + \lambda_1||\mathbf{B}||_1$$
$$+ \lambda_2||\mathbf{C}||_F^2 + \lambda_3||\mathbf{w}||_2^2 \quad s.t. \;\; \mathbf{c}_{nk} \geq 0 \quad (3.7)$$

The parameter λ_1 controls the number of nonzero elements in **B** by scaling the contribution of the ℓ_1 penalty. Similarly, λ_2 and λ_3 relate to element wise bounds on the entries in **C** and **w** since they scale the contribution of their respective ℓ_2 norms.

3.4 Joint Inference Strategy

We employ an alternating minimization technique in order to infer the set of latent variables $\{\mathbf{B},\mathbf{C},\mathbf{w}\}$. Here, we optimize the JNO objective function from Eq. (3.7) for each output variable, while holding the estimates of the other unknowns constant.

Proximal gradient descent [147] is an attractive algorithm to handle the non-differentiable sparsity penalty on **B** in Eq. (3.7), when the supporting terms in the variable of interest are convex. However, from Eq. (3.7), we see that the Frobenius norm terms expand to a biquadratic representation in **B**, which is non-convex. We circumvent this problem by introducing N constraints of the form $\mathbf{D}_n = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)$. We enforce these constraints using the Augmented Lagrangian [148], denoting the set of Lagrangian matrices by $\{\mathbf{\Lambda}_n\}$. The modified objective function in Eq. (3.7) takes the form:

$$\mathcal{J}(\cdot) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{D}_n\mathbf{B}^T||_F^2 + \gamma||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2 + \sum_n \text{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n))\right]$$

$$+ \sum_n \frac{1}{2}||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)||_F^2 + \lambda_1||\mathbf{B}||_1 + \lambda_2||\mathbf{C}||_F^2 + \lambda_3||\mathbf{w}||_2^2 \quad s.t.\ \mathbf{c}_{nk} \geq 0 \tag{3.8}$$

Tr[**M**] is the trace operator, which sums the diagonal elements of the argument matrix **M**. The additional Frobenius norm terms $||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)||_F^2$ act as regularizers for the trace constraints. Observe that Eq. (3.8) is now convex in both **B** and the set $\{\mathbf{D}_n\}$, which allows us to optimize them via standard procedures.

Fig. 3.2 provides an overview of the alternating minimization strategy employed. Each individual block in our optimization is described below.

3.4.1 Proximal Gradient Descent on B

We first write out the optimization problem with respect to $\mathbf{B}$ when the estimates of $\{\mathbf{C}, \mathbf{w}\}$ are held constant:

$$\mathbf{B}^{k+1} = \text{argmin}_{\mathbf{B}} \lambda_1 ||\mathbf{B}||_1 + \sum_n ||\mathbf{\Gamma}_n - \mathbf{D}_n \mathbf{B}^T||_F^2 + \sum_n \text{Tr}\left[\mathbf{\Lambda}_n^T (\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n))\right]$$

$$+ \sum_n \frac{1}{2} ||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)||_F^2$$

$$\therefore \mathbf{B}^{k+1} = \text{argmin}_{\mathbf{B}} ||\mathbf{B}^k||_1 + \frac{1}{\lambda_1} \mathcal{G}(\mathbf{B}^k)$$

We see that the proximal gradient iteration is the solution to the following fixed point problem:

$$\mathbf{0} \in \frac{1}{\lambda_1} \frac{\partial \mathcal{G}}{\partial \mathbf{B}} + \partial(||\mathbf{B}||_1)$$

Here, $-\frac{\partial \mathcal{G}}{\partial \mathbf{B}}$ is a descent direction for the $\mathbf{B}$ update, and t controls the magnitude of the step we take in this direction. In practice, we fix t at 10^{-4} for stable convergence. The derivative of $\mathcal{G}$ with respect to $\mathbf{B}$, is computed as:

$$\frac{\partial \mathcal{G}}{\partial \mathbf{B}} = \sum_n \left[2\left[\mathbf{B}\mathbf{D}_n^T \mathbf{D}_n - \mathbf{\Gamma}_n \mathbf{D}_n\right] - \mathbf{D}_n \mathbf{diag}(\mathbf{c}_n)\right]$$

$$+ \sum_n \left[\mathbf{B}\mathbf{diag}(\mathbf{c}_n)^2 - \mathbf{\Lambda}_n \mathbf{diag}(\mathbf{c}_n)\right]$$

Given the fixed learning rate parameter t, the proximal update for $\mathbf{B}$ is easily computed as:

$$\mathbf{B}^{k+1} = \mathbf{sgn}(\mathbf{X}) \circ (\mathbf{max}(|\mathbf{X}| - t, \mathbf{0})) \tag{3.9}$$

$$\mathbf{X} = \mathbf{B}^k - (t/\lambda_1) \frac{\partial \mathcal{G}}{\partial \mathbf{B}} \tag{3.10}$$

This step first estimates a locally smooth quadratic model at each iterate $\mathbf{B}^k$ and applies a step of iterative shrinkage thresholding to the compute the local solution of $\mathbf{B}$. The resulting iterative algorithm is computationally efficient compared to the counterpart sub-gradient based descent methods and arrives at a good local solution for an appropriate choice of the learning rate.

At a high level, Eq. (3.10) performs an iterative shrinkage thresholding operation to handle the non-smoothness of the $||\mathbf{B}||_1$ using a locally smooth quadratic model.

3.4.2 Optimizing C using Quadratic Programming

The objective is quadratic in $\mathbf{C}$ when $\mathbf{B}$, and $\mathbf{w}$ are held constant. Furthermore, the $\mathbf{diag}(\mathbf{c}_n)$ term decouples the updates for $\mathbf{c}_n$ across patients. Each $\mathbf{c}_n$ is the solution to the a separate optimization problem of the following form:

$$\mathbf{c}_n^{k+1} = \text{argmin}_{\mathbf{c}_n \in \mathcal{R}^{K+}} \text{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^k))\right] + \lambda_2||\mathbf{c}_n^k||_2^2$$

$$+ \frac{1}{2}||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^k)||_F^2 + \gamma((\mathbf{c}_n^k)^T\mathbf{w} - \mathbf{y}_n)^2$$

Hence, we use N quadratic programs (QP) of the form below to solve for the vectors $\{\mathbf{c}_n\}$:

$$\frac{1}{2}\mathbf{c}_n^T\mathbf{H}_n\mathbf{c}_n + \mathbf{f}_n^T\mathbf{c}_n \;\; s.t. \;\; \mathbf{A}_n\mathbf{c}_n \leq \mathbf{b}_n$$

The QP parameters for our problem are given by:

$$\mathbf{H}_n = \mathcal{I}_K \circ (\mathbf{B}^T\mathbf{B}) + 2\gamma\mathbf{w}\mathbf{w}^T + 2\lambda_2\mathcal{I}_K \tag{3.11}$$

$$\mathbf{f}_n = -2\left[\mathcal{I}_K \circ (\mathbf{D}_n^T + \mathbf{\Lambda}_n^T)\mathbf{B}\right]\mathbf{1} - 2\gamma\mathbf{y}_n\mathbf{w}; \tag{3.12}$$

$$\mathbf{A}_n = -\mathcal{I}_K \quad \mathbf{b}_n = \mathbf{0} \tag{3.13}$$

The non-negativity constraint requires us to project the quadratic programming solution to the space of positive reals in K dimensions for each $\mathbf{c}_n$ through $\mathbf{A}_n$ and $\mathbf{b}_n$. Since the Hessians $\{\mathbf{H}_n\}$ for our problem are positive definite, there exist polynomial time algorithms for solving the bound constrained QPs to the global optimum value. The decoupling of the $\{\mathbf{c}_n\}$ allows us to solve for each coefficient vector in parallel.

3.4.3 Closed Form Update for w

The global minimizer of $\mathbf{w}$ is computed at the first order stationary point of the convex objective, which is:

$$\mathcal{J}(\mathbf{w}) = \lambda_3||\mathbf{w}||_2^2 + \gamma||\mathbf{C}^T\mathbf{w} - \mathbf{y}||_2^2$$

$$\frac{\partial\mathcal{J}}{\partial\mathbf{w}} = 0 = 2\lambda_3\mathbf{w} + 2\gamma(\mathbf{C}\mathbf{C}^T\mathbf{w} - \mathbf{C}\mathbf{y})$$

The closed form update can be expressed as:

$$\mathbf{w} = (\mathbf{C}\mathbf{C}^T + \frac{\lambda_3}{\gamma}\mathcal{I}_K)^{-1}(\mathbf{C}\mathbf{y})$$

Thus, the ratio $\frac{\lambda_3}{\gamma}$ acts as a regularizer for the matrix inversion in our estimate, ensuring that the update for $\mathbf{w}$ is well defined at each iterate. This is analogous to a regularized linear regression update for $\mathbf{w}$.

3.4.4 Optimizing the Constraint Variables $\mathbf{D}_n$ and $\mathbf{\Lambda}_n$

A closed form solution for the primal variables $\{\mathbf{D}_n\}$ can be obtained by setting their first derivatives to zero:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{D}_n} = 0 = \mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n\mathbf{B} - \mathbf{\Lambda}_n - \mathbf{D}_n - 2\mathbf{D}_n\mathbf{B}^T\mathbf{B}$$

$$\therefore \quad \mathbf{D}_n = (\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n\mathbf{B} - \mathbf{\Lambda}_n)(\mathcal{I}_K + 2\mathbf{B}^T\mathbf{B})^{-1}$$

The gradient ascent update on $\{\mathbf{\Lambda}_n\}$ is as follows:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{\Lambda}_n} = \mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)$$

$$\mathbf{\Lambda}_n^{k+1} = \mathbf{\Lambda}_n^k + \eta_k \frac{\partial \mathcal{J}}{\partial \mathbf{\Lambda}_n}$$

Similar to the case of the coefficients $\mathbf{c}_n$, each of the N pairs of updates $\{\mathbf{D}_n, \mathbf{\Lambda}_n\}$ are decoupled from each other, and can be solved in parallel.

Overall, the sets of $\mathbf{\Lambda}_n$ gradient ascent updates ensure that the respective set of constraints $\mathbf{D}_n = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)$ is satisfied with increasing certainty at each iteration. The Augmented Lagrangian construct $||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)||_F^2$ prevents trivial Lagrangian $\mathbf{\Lambda}_n$ solutions.

The updates for $\mathbf{D}_n$ and $\mathbf{\Lambda}_n$ ensure that the proximal constraints are satisfied with increasing certainty at each iteration. The learning rate parameter η_k for the gradient ascent step of the augmented Lagrangian is chosen to guarantee sufficient decrease for every iteration of alternating minimization.

In practice, we initialize this value to 10^{-3}, and scale it by 0.5 at each iteration.

3.4.5 Prediction on an unseen patient

In order to estimate the coefficients $\hat{\mathbf{c}}$ for a new patient, we re-solve the quadratic program in Eq. (3.13) using the $\{\mathbf{B}^*, \mathbf{w}^*\}$ computed from the training data via the procedure outlined in Section 3.4. We explicitly set the contribution from the data term in Eq. (3.8) to 0, since the corresponding value of $\hat{\mathbf{y}}$ is unknown for the new patient. We also implicitly assume that the conditions for the proximal operator hold, i.e. the constraint $\hat{\mathbf{D}} = \mathbf{B}^*\mathbf{diag}(\hat{\mathbf{c}})$ is exactly satisfied. The estimation of the unseen patient's coefficients are thus mathematically formulated as follows:

$$\hat{\mathbf{c}} = \text{argmin}_{\mathbf{c}} ||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c})\mathbf{B}^T||_F^2 + \lambda_2||\mathbf{c}||_2^2 \quad s.t. \;\; \mathbf{c}_k \geq 0 \tag{3.14}$$

Once again, Eq. (3.14) can be formulated as a quadratic program. The parameters from Eq. (3.13) correspond to:

$$\mathbf{H}_n = 2(\mathbf{B}^T\mathbf{B}) \circ (\mathbf{B}^T\mathbf{B}) + 2\lambda_2\mathcal{I}_K$$

$$\mathbf{f}_n = -2\left[\mathcal{I}_K \circ (\mathbf{B}^T\mathbf{\Gamma}_n\mathbf{B})\right]\mathbf{1};$$

$$\mathbf{A}_n = -\mathcal{I}_K \qquad \mathbf{b}_n = \mathbf{0}$$

The estimate for the behavioral score for the test patient is given by the vector product $\hat{\mathbf{y}} = \hat{\mathbf{c}}^T\mathbf{w}^*$.

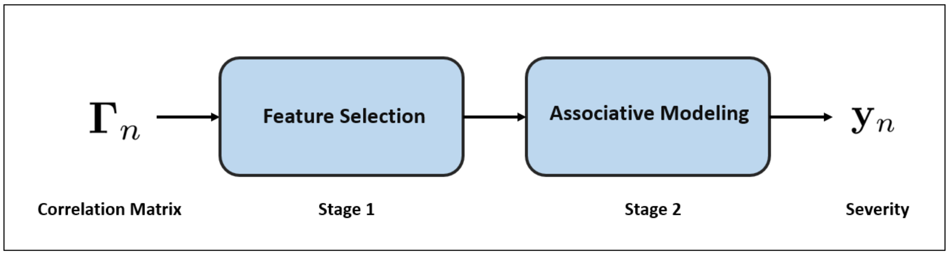


Figure 3.3: A typical two stage baseline. We input the correlation matrices to Stage 1, which performs Feature Extraction on the raw correlations. This step could be a technique from machine learning, graph theory or a statistical measure. Stage 2 fits an associative regression model to the output representation of Stage 1

3.5 Model Evaluation

3.5.1 Baseline Methods

We evaluate the performance of our method against a set of well established statistical, graph theoretic, and data-driven frameworks that have been used to provide rich feature representations. Fig. 3.3 describes a general two stage pipeline for our task. The first stage is a representation learning step used for feature extraction. Stage 2 is a regression model to map the learned features to behavioral data. We evaluate our method against several choices of linear and non-linear algorithms for Stage 1. These are combined with a regularized linear regression in Stage 2, similar to our method. Additionally, we evaluate the performance obtained by omitting a Stage 1 and training a deep neural network end-to-end on the input correlation features. Lastly, we demonstrate the advantage provided by combining the neuroimaging and behavioral representations in the JNO framework. For this, we present a comparison where the feature learning and prediction stages are decoupled, similar to the baselines.

3.5.1.1 Machine Learning Approach (PCA)

We start with the $P \times P$ correlation matrix $\mathbf{\Gamma}_n$ for each patient. Since this matrix is symmetric, we have $M = \frac{P\times(P-1)}{2}$ distinct rs-fMRI correlation pairs between various communicating sub-regions. Accordingly, the features from every individual are composed into a descriptor matrix $\mathbf{X} \in \mathcal{R}^{M\times N}$. We further concentrate these feature into a small number of representative bases. The basis extraction procedure in Stage 1 corresponds to a linear mapping in the original correlation space via a **Principal Component Analysis (PCA)**. In Stage 2, we construct a **regularized linear regression (ridge regression)** on the projected features to predict the clinical severity. PCA projects the observations onto a set of uncorrelated *principal component basis* by means of an orthogonal linear decomposition. Mathematically, PCA poses the following dimensionality reduction problem:

$$\mathcal{F}(\cdot) = \text{argmin}_{\mu,\mathbf{U},\mathbf{Y}} ||\mathbf{X} - \boldsymbol{\mu}\mathbf{1}^T - \mathbf{U}\mathbf{Z}||_F^2 \;\; s.t. \;\; \mathbf{U}^T\mathbf{U} = \mathcal{I}_d, \;\; \mathbf{Z}\mathbf{1} = \mathbf{0} \qquad (3.15)$$

Here, $\mathbf{U} \in \mathcal{R}^{N\times d}$ is the d dimensional subspace basis which best approximates the information from $\mathbf{X}$ in the Frobenius norm sense, computed by calculating the eigenvectors of the sample covariance matrix $\mathbf{X}\mathbf{X}^T$. Thus, $\mathbf{Z} \in \mathcal{R}^{d\times N}$ is a compact d dimensional representation of $\mathbf{X}$, where $d \ll M$. $\mathbf{1}$ is a d dimensional vector of ones. The constraint $\mathbf{Z}\mathbf{1} = \mathbf{0}$ centers $\mathbf{Z}$.

3.5.1.2 Statistical Approach (ICA)

Here, we use **Independent Component Analysis (ICA)** as the Stage 1 algorithm combined with **ridge regression**. ICA operates on the raw time series data to extract representative spatial patterns that explain rs-fMRI connectivity. ICA has become ubiquitous for identifying group level as well as individual-specific connectivity signatures. It decomposes a multivariate signal into 'independent' non-Gaussian components based on the statistics of the data. Mathematically, ICA models the components $\{\mathbf{y}_k\}$ of the observed signal $\mathbf{y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_m\right]$ as a sum of n independent components $\mathbf{S} = \left[\mathbf{s}_1, \ldots, \mathbf{s}_n\right]$ combined via the mixing matrix $\mathbf{A} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_n\right]$

$$\mathbf{y} = \sum_{i=1}^{n} \mathbf{s}_i \mathbf{a}_i \quad \textit{i.e.} \;\; \mathbf{Y} = \mathbf{AS} \tag{3.16}$$

$\mathbf{s}$ can be recovered by multiplying the observed signals $\mathbf{Y}$ with the inverse of the mixing matrix $\mathbf{W} = \mathbf{A}^{-1}$. We adaptively estimate both the mixing matrix $\mathbf{A}$ and the components $\mathbf{s}$ by setting up a cost function that maximizes the non-gaussianity of $\mathbf{s}_i = \mathbf{w}_i^T \mathbf{y}$ or minimizes the mutual information.

Group ICA extends this algorithm to a multi-subject analysis for extracting independent spatial patterns common across patients, but combined via individual time courses. We use the GIFT [149] software to perform Group-ICA to derive independent spatial maps for each patient. The correlation values between the identified components are fed to the regression model.

3.5.1.3 Graph Theoretic Approach (Node Degree)

Each correlation matrix $\mathbf{\Gamma}_n$ can be thresholded and considered a graph adjacency matrix, which we denote by $\mathbf{\Psi} \in \mathcal{R}^{P \times P}$. The element $\mathbf{\Psi}_{ij}$ gives the strength of association between two communicating sub-regions i and j. The underlying graph topology can be summarized using node/edge based importance measures [28, 82]. Again, we use a regularized linear regression technique to estimate the severity score from the reduced representation. This treatment closely parallels the machine learning approach, as we can view the graph measures as a dimensionality reduction. We compute **Node Degree (D_N)** from the adjacency graph followed by a **ridge regression** on the features.

Given the adjacency matrix $\mathbf{\Psi}$, the degree of region v is equal to the number of edges incident on v, with loops counted twice. Mathematically, the degree $\mathbf{D}_N(v)$ is computed as follows:

$$\mathbf{D}_N(v) = \sum_{j \neq v} \mathbb{1}(\mathbf{\Psi}_{jv} > 0) \tag{3.17}$$

where, $\mathbb{1}(.)$ is the indicator function, which takes the value 1 if the condition is satisfied, and 0 otherwise. This metric captures the importance of each node in explaining the graph, which in our case, corresponds to the average connectivity strength of each region in the brain.

3.5.1.4 A Neural Network Approach

Recently, there has been an upsurge in using neural networks to investigate neuroimaging correlates of developmental disorders [80]. Here, we test the efficacy of a simple **Artificial Neural Network (ANN)** for predicting the severity score from the correlation feature matrix **X** defined above. The network architecture encodes a series of non-linear transformations of the input correlations to approximate the severity score. Recall that the size of the input is dependent on our choice of parcellation, which could be of considerable width

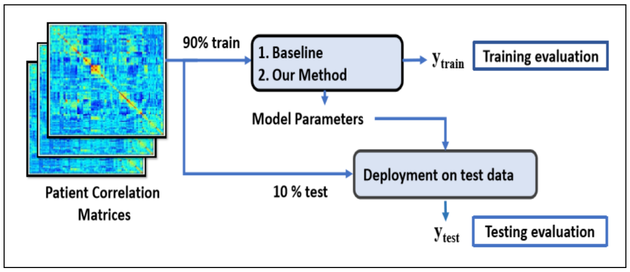


Figure 3.4: A ten-fold cross validation for evaluating performance

(of the order of $\approx$ 5000 connections for $P = 100$). After evaluating several architectures, we employ a two hidden layer network with widths 8000 and 10 respectively. We use a Rectified Linear Unit (ReLU) non-linearity after the first hidden layer and a Tanh non-linearity after the second hidden layer. We used the ADAM optimizer with an initial learning rate of 10^{-4}, scaled by 0.9 per 10 epochs, and a momentum of 0.9 to train the network.

3.5.2 Predictive Performance

We characterize the performance of each method using a 10 fold cross validation as illustrated in Fig. 3.4. For a given parameter setting, we first split the data set into 10 training and test folds. For each of the folds, we train the models on a 90 percent training set split of the data. We report the score prediction on the held out 10 percent, which constitutes the testing set for that fold. Note that each datapoint is in the test set in exactly one of the 10 folds.

We report two quantitative measures of performance. Median Absolute Error (MAE) quantifies the absolute distance between the measured and predicted scores:

$$\text{MAE} = \text{median}(|\hat{\mathbf{y}} - \mathbf{y}|),$$

where the median is computed across the set of patients. We report MAE values with the standard deviation of the error. Lower MAE indicates better testing performance.

Normalized Mutual Information (NMI) assesses the similarity in the distribution of the predicted and observed score distributions across test patients. NMI is computed as follows:

$$\text{NMI}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{H(\mathbf{y}) + H(\hat{\mathbf{y}}) - H(\mathbf{y}, \hat{\mathbf{y}})}{\min\{H(\mathbf{y}), H(\hat{\mathbf{y}})\}}$$

where $H(\mathbf{y})$ denotes the entropy of $\mathbf{y}$ and $H(\mathbf{y}, \hat{\mathbf{y}})$ is the joint entropy between $\mathbf{y}$ and $\hat{\mathbf{y}}$. NMI ranges from $0 - 1$ with higher values indicating a better agreement between predicted and measured score distributions, and thus characterizing improved performance.

3.5.3 Implementation Details

Our method has five user-specified parameters $\{\gamma, \lambda_1, \lambda_2, \lambda_3, K\}$. Recall that K is the number of basis networks, γ is the penalty tradeoff between the representation and regression terms, λ_1 is the sparsity penalty, while λ_2 and λ_3 are the regularization penalties on the coefficients $\mathbf{C}$ and regression weights $\mathbf{w}$ respectively.

We use the knee point of the eigenspectrum of the correlation matrices $\mathbf{\Gamma}_n$ to select the number of bases ($K = 8$). Empirically, the JNO model is insensitive to the choice of λ_3 and γ, so we fix both at one. Effectively, we are left with two free parameters, which we optimize by performing a bivariate grid search. We note that the generalization accuracy is dependent on the dynamic range of the scores and is sensitive to λ_1 and λ_2. Based on the cross validation results, we finally use the following settings in our experiments: For the KKI dataset, $\{\lambda_2 = 0.2, \lambda_1 = 30\}$ for ADOS, $\{\lambda_2 = 0.9, \lambda_1 = 50\}$ for SRS and $\{\lambda_2 = 0.6, \lambda_1 = 20\}$ for Praxis. We will discuss the parameter sensitivity in Subsection 3.9.

To provide a fair comparison with our JNO framework, we use a joint grid search on the Stage 1 hyperparameters and the Stage 2 ridge penalty to optimize these values for every baseline method. Again, we report the best performance in a ten fold cross validation setting.

We select 10 PCA components for the KKI dataset, and 15 for the NYU dataset. For ICA, we obtained good performance for 35 spatial maps obtained from GIFT [149]. For the graph theoretic baseline, we threshold the correlation matrices $\{\mathbf{\Gamma}_n\}$ at 0.2 to obtain valid adjacency matrices $\{\mathbf{\Psi}_n\}$. In conjunction with these, the ridge penalty parameter was swept across four orders of magnitude. Finally, we include the performance upon decoupling the ridge regression and the matrix decomposition in Eq. (3.5) as a sanity check. This is akin to the two stage treatment in the baselines where the two terms are not explicitly coupled as in the JNO objective.

3.6 Experiments on Synthetic Data

As a sanity check, we first sample data from the generative model in Eq. (3.5) and use the optimization outlined in Section 3.4 to estimate the unknowns $\{\mathbf{B}, \mathbf{C}, \mathbf{w}\}$. This procedure helps us analyze the performance of the algorithm under different noise scenarios.

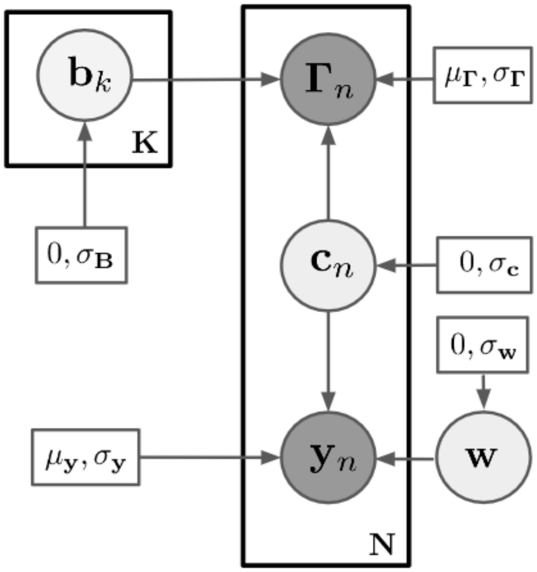


Figure 3.5: The graphical model for the joint objective. For our synthetic experiments, we fix the model parameters $\sigma_\mathbf{C} = 2, \sigma_\mathbf{w} = 0.2$

The inputs to our model are the correlation matrices $\{\mathbf{\Gamma}_n\}$ and the clinical scores $\{\mathbf{y}_n\}$. We note that the model gives a complete description of each $\mathbf{\Gamma}_n$ in terms of the basis vectors $\{\mathbf{b}_k\}$ and the patient coefficients $\{\mathbf{c}_n\}$. Since the data representation terms for each patient are coupled solely through the basis representation, the coefficient descriptors are independent of each other. In a similar observation, each score $\mathbf{y}_n$ is explained by the corresponding $\mathbf{c}_n$, independent of the remaining subjects, when we fix the regression vector $\mathbf{w}$. We use this information to describe the *observed data* $\{\mathbf{\Gamma}_n, \mathbf{y}_n\}$ using a generative model with the likelihood model based on the *hidden variables* $\{\mathbf{B}, \mathbf{C}, \mathbf{w}\}$.

Notice that, when treated as a Bayesian log-likelihood (i.e. taking a negative exponent of the objective), the ℓ_2 norms in Eq. (3.5) translate into Gaussian distributions, and the ℓ_1 norm is equivalent to a Laplacian prior. The corresponding graphical model is shown in Fig. 3.5. The observed variables are indicated by the shaded circles. The white circles contain the hidden variables. The distribution parameters for the hidden variables are indicated in the corresponding rectangle pointing to the variable. The Laplacian parameter σ_B controls the overlap in the patterns of sparsity in $\mathbf{B}$, which relates to λ_1. $\mathbf{C}$ and $\mathbf{w}$ are described by Gaussians with means zero (i.e. ℓ_2 norm offset). The variances $\sigma^2_{\mathbf{C}}$ and $\sigma^2_{\mathbf{w}}$ are related to the penalty parameters λ_2 and λ_3 respectively. The non-negativity constraint on $\mathbf{c}_n$ is handled by folding (i.e. taking the absolute value of) the normal distribution to restrict the $\mathbf{c}_n$ values to be positive reals. The observed variable $\{\mathbf{y}_n\}$, translates to a Gaussian with mean $\mu_{\mathbf{y}_n} = \mathbf{c}_n^T\mathbf{w}$, and variance parameters $\sigma_{\mathbf{y}_n}$. This is again folded to reflect positive values of $\mathbf{y}_n$. The correlation matrices $\{\mathbf{\Gamma}_n\}$ are drawn from a Gaussian distribution with mean $\mu_{\mathbf{\Gamma}_n} = \mathbf{Bdiag}(\mathbf{c}_n)\mathbf{B}^T$ (which is positive

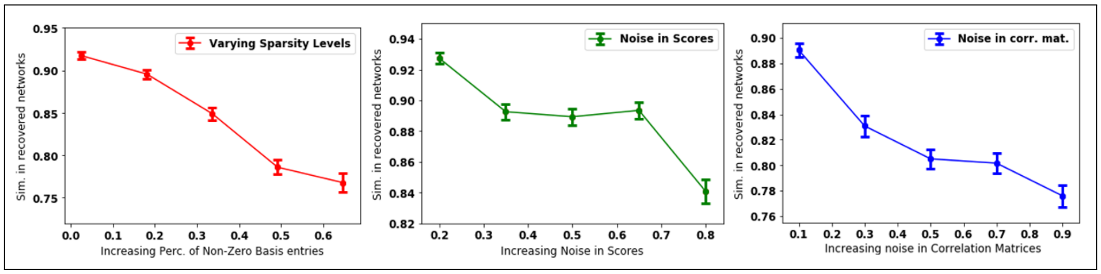


Figure 3.6: Performance on synthetic experiments. (**L**): Varying the level of sparsity ($\sigma_{\mathbf{\Gamma}_n} = 0.4$, $\sigma_{\mathbf{y}_n} = 0.2$), (**M**): Varying the level of noise in $\mathbf{y}_n$ ($\sigma_{\mathbf{B}} = 0.2$, $\sigma_{\mathbf{\Gamma}_n} = 0.4$) , (**R**): Varying the level of noise in $\mathbf{\Gamma}_n$ under ($\sigma_{\mathbf{B}} = 0.2$, $\sigma_{\mathbf{y}_n} = 0.2$) Values on the x-axis have been normalized to reflect a $[0-1]$ range by dividing by the maximum value of the variable. Deviations from the mean recovered similarity for each parameter setting is indicated in the figure and have been reported as a standard error value. The reported x-axis range reflects the regimes within which the algorithm converges to a local solution

semi-definite by construction) and variance σ_{Γ_n}.

There are two sources of noise for the observed variables, which include the error in the correlation matrices $\mathbf{\Gamma}_n$, and the error in the severity scores $\mathbf{y}_n$. These scenarios can be directly related to controlling the variance parameters $\sigma_{\mathbf{\Gamma}_n}$ and $\sigma_{\mathbf{y}_n}$ respectively. Additionally, we are interested in the performance of the algorithm under varying levels of overlap in the sparsity patterns in $\mathbf{B}$.

We evaluate the performance using an average inner-product measure of similarity S between each recovered network, $\hat{\mathbf{b}}_k$, and its corresponding best matched generating network, $\mathbf{b}_k$, both normalized to unit norm, i.e.:

$$S = \frac{1}{K}\sum_{k} \frac{|\mathbf{b}_k^T \hat{\mathbf{b}}_k|}{||\mathbf{b}_k||_2||\hat{\mathbf{b}}_k||_2}. \tag{3.18}$$

Fig. 3.6 depicts the performance of the algorithm in these three cases. The x-axis corresponds to increasing the levels of noise, while the y-axis indicates the similarity metric S computed for the particular setting. In the leftmost plot, an x-axis value close to 0 indicates high percentage of sparsity in $\mathbf{B}$, while increasing values correspond to denser basis matrices. Throughout this experiment, the values of the other free parameters in the generative model were held constant. The middle plot evaluates subnetwork recovery when the noise in the scores, i.e. $\sigma_{\mathbf{y}_n}$ is increased. The x-axis reports normalised values of $\sigma_{\mathbf{y}_n}$ while the remaining free parameters were held constant. Similarly, the rightmost plot in Fig. 3.6 indicates performance under varying noise in the correlation matrices $\mathbf{\Gamma}_n$. Again, normalized $\sigma_{\mathbf{\Gamma}_n}$ values are reported on the x-axis. Numerical results have been aggregated over 100 independent trials.

As expected, increasing the noise in the correlation matrices and scores

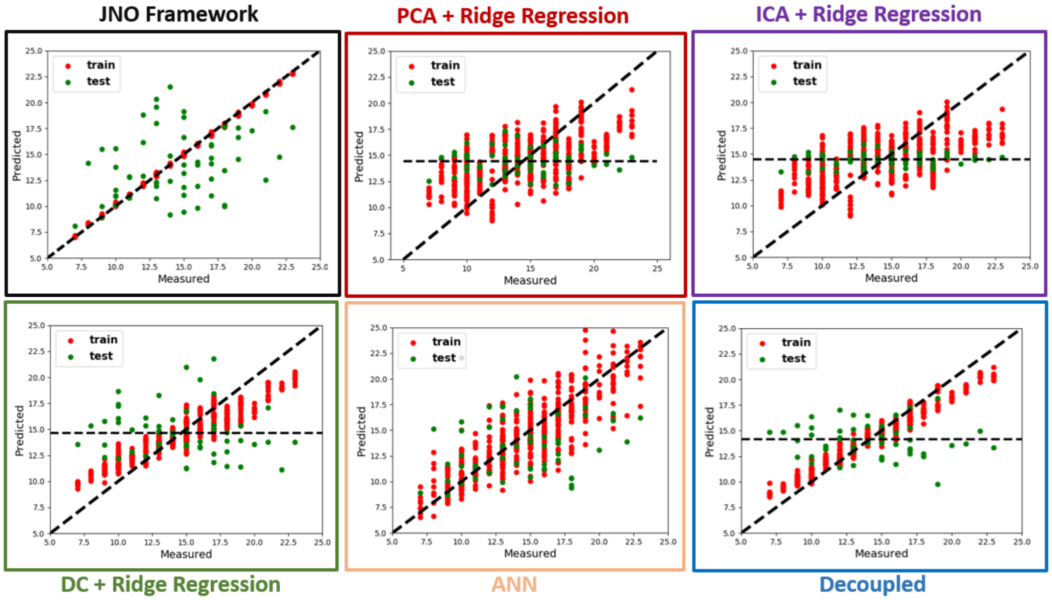


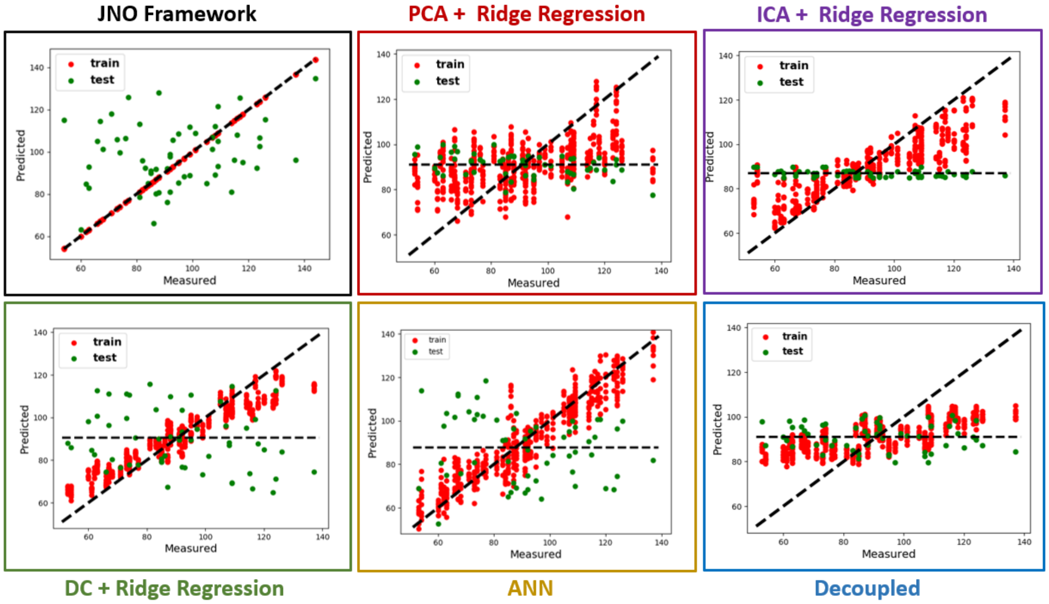
Figure 3.7: KKI dataset: Prediction performance for the ADOS score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange Box**: ANN on correlation features **Blue Box:** Decoupled matrix cactorization and ridge Regression

worsens the recovery performance of the algorithm. This is indicated by the decay in the similarity measure with increasing noise parameters as well as an increase in the corresponding variance. Additionally, the algorithm performs better when there is lesser overlap in the columns of **B**, i.e. when the generating basis is sparse. However, we observe that our algorithm is robust in the noise regime estimated from the real-world rs-fMRI data $(0.01 - 0.2)$ and recovered sparsity levels $(0.1 - 0.4)$. In addition, we identify the stable parameter settings for the algorithm which guide our real world experiments.

3.7 Population Studies on Autism

Figs. 3.7−3.9 compare the performance of our method against the baselines described in Section 3.5.1 for the prediction of ADOS, SRS and Praxis respectively for the KKI dataset. We plot the score predicted by the algorithm on the y-axis against the measured ground truth score on the x-axis. The bold $x = y$ line indicates ideal performance. The red points correspond to training data, while the green points represent the held out testing data for all the folds in the cross validation. Our method is indicated at the top left corner.

We observe that, although the training performance of the baselines is good (i.e. the red points follow the $x = y$ line), the JNO achieves the best training performance in all cases. Furthermore, we notice that all the two stage baseline testing performances track the mean value of the held out data (indicated by the black horizontal line). Our method clearly outperforms the baselines and is able to capture a trend in the data, beyond a mean value estimation in case of both datasets for all scores. This can be observed by the spread of the green points about the $x = y$ line in the case of the JNO method. Through our experiments, we noticed that the testing performance of the ANN is dependent on the choice of architecture. For example, the architecture chosen in Section 3.5 performs well on predicting ADOS for the KKI dataset, but performs poorly on all other comparisons. Our empirical evaluations could not identify a single architecture that performed well in all cases, like our JNO framework. The failure of the two stage decomposition in the bottom right comparison figures strengthens our hypothesis that a joint modeling of the neuroimaging and behavioral data is necessary in the context



JNO Framework

PCA + Ridge Regression

ICA + Ridge Regression

DC + Ridge Regression

ANN

Decoupled

Figure 3.8: KKI dataset: Prediction performance for the SRS score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange Box**: ANN on correlation features **Blue Box:** Decoupled matrix factorization and ridge regression

of generalization onto unseen data. The lackluster generalization performance of the baselines is testament to the difficulty of the task at hand. The number of connections or features available to us are of the order of a 6670 dimensional vector representation for ≈ 60 patients. Both the machine learning and graph theoretic techniques we selected for a comparison are well known in literature for being able to robustly provide compact characterizations for high dimensional datasets. However, we see that PCA and ICA are unable to estimate a reliable projection of the data that is particularly indicative of clinical severity. Similarly, the node degree measure heavily rely on being able to accurately identify informative network topologies from the observed correlation matrices. However, its aggregate nature captures general trends and

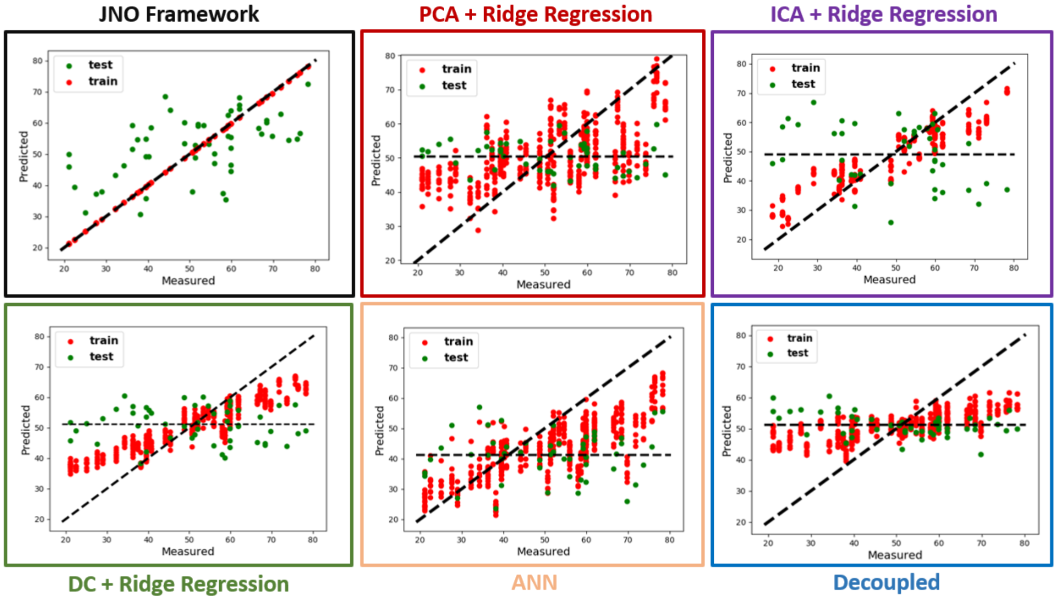


Figure 3.9: KKI dataset: Prediction performance for the Praxis score for **Black Box:** JNO Framework. **Red Box:** PCA and ridge regression **Purple Box:** ICA and ridge regression **Green Box:** Node degree centrality and ridge regression **Orange Box**: ANN on correlation features **Blue Box:** Decoupled matrix factorization and ridge regression

is not successful in characterizing subtle patient level differences. The failure of the decoupled matrix factorization and ridge regression makes a strong case for including the regression term as a part of our JNO objective. The basis obtained in this case are not indicative of clinical severity, due to which the regression performance suffers. Despite sweeping parameters across several orders of magnitude, we observe that the baselines are only good at capturing group level information, as is indicated by the training fit. However, they fail to characterize patient level differences for an unseen subject and simply predict the mean of the given cohort. On the other hand, the generalization power of the ANN is contingent on the model order choice. This is demonstrated by its inability to perform well on comparisons outside of ADOS. Said another way, we have to change the network architecture for different severity measures across datasets. This is a major computational disadvantage when compared with our method.

A key difference between the JNO framework and the baselines is that we utilize the structure of the correlation matrices to guide the predictive model. In essence, we optimize for the tradeoff between the neuroimaging and behavioral data representations jointly, instead of posing it as a two stage problem. The matrix decomposition we employ explicitly models the group information through the basis, and the patient differences through the coefficients. The limited number of basis elements we employ to decompose the data provides us with compact representations which explain the connectivity information well. The regularization terms and constraints ensure that the problem is well posed, while providing clinically meaningful and informative representations about the data. We also quantify the performance indicated in these figures in Tables 3.1

3.8 Clinical Significance

Figs. 3.10−3.12 illustrate the subnetworks in **B**, as trained on the ADOS, SRS and Praxis in the KKI dataset, respectively. Since each column of the basis corresponds to a set of co-activated subregions, we plot the normalized values stored in these columns onto the corresponding AAL ROIs. The colorbar indicates subnetwork contribution to the AAL regions. Regions colored as negative values are anticorrelated with regions storing positive ones. We rank the 8 subnetworks obtained from SRS and Praxis according to their overlap

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test
ADOS	PCA & ridge	2.18 ± 2.2	2.99 ± 1.71	0.22	0.18
	ICA & ridge	2.13 ± 1.1	3.01 ± 1.90	0.31	0.23
	D_N & ridge	1.22 ± 0.91	3.68 ± 2.53	0.45	0.39
	ANN	2.68 ± 2.21	**2.28 ± 1.30**	0.91	**0.58**
	Decoupled	2.36 ± 2.33	2.63 ± 1.90	0.15	0.30
	JNO Framework	**0.088 ± 0.13**	2.53 ± 1.86	**0.99**	0.52
SRS	PCA & ridge	12.92 ± 10.48	19.09 ± 12.48	0.64	0.39
	ICA & ridge	7.96 ± 6.35	20.8 ± 17.3	0.83	0.63
	D_N & ridge	5.77 ± 4.88	19.63 ± 17.23	0.85	0.59
	ANN	4.77 ± 4.09	21.25 ± 14.63	0.81	0.56
	Decoupled	12.06 ± 10.04	18.5 ± 16.4	0.74	0.37
	JNO Framework	**0.13 ±0.07**	**13.27 ± 10.85**	**0.99**	**0.78**
Praxis	PCA & ridge	9.44 ± 6.83	12.83 ± 8.84	0.64	0.37
	ICA & ridge	4.79 ± 4.17	13.08 ± 13.07	0.73	0.63
	D_N & ridge	4.78 ± 3.24	13.93 ± 8.14	0.68	0.56
	ANN	9.34 ± 7.21	14.90 ± 10.06	0.69	0.39
	Decoupled	10.17 ± 7.96	13.24 ± 10.38	0.68	0.29
	JNO Framework	**0.11 ± 0.065**	**10.18± 6.58**	**0.99**	**0.79**

Table 3.1: Performance evaluation using **Median Absolute Error (MAE)** and **Normalized Mutual Information (NMI)** fit, both for testing & training. Lower MAE & higher NMI score indicate better performance. We have highlighted the best performance in bold. Near misses have been underlined.

with the subnetworks from ADOS. As seen from these figures, corresponding subnetworks show considerable overlap in regional co-activation patterns. The individual variations can arise from the fundamental differences in the behavioral traits that each score is trying to capture.

From a clinical standpoint, Subnetwork 7 includes competing i.e. anti-correlated contributions from regions of the default mode network (DMN) and somatomotor network (SMN). Abnormal connectivity within the DMN and SMN has been previously reported in ASD [150, 98]. Subnetwork 5 comprises of competing contributions from SMN regions. Additionally, it includes higher order visual processing areas in the occipital and temporal lobes, which

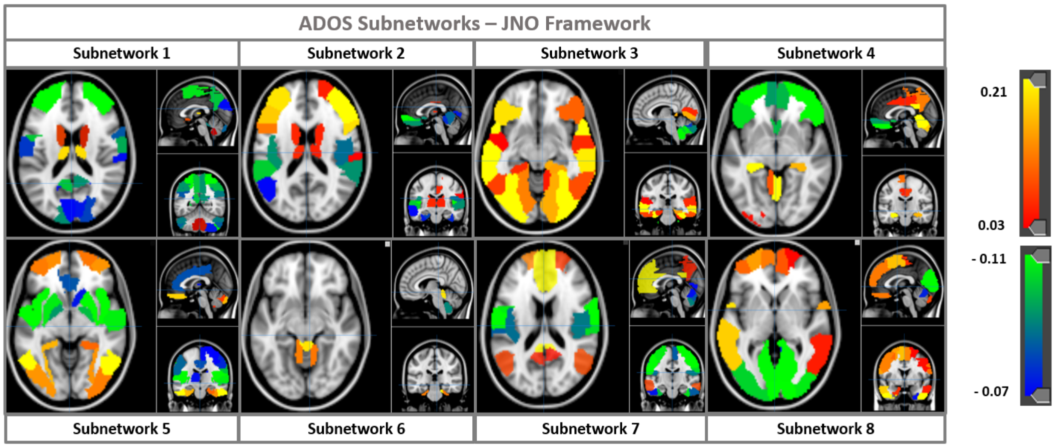


Figure 3.10: Subnetworks estimated to predict the ADOS score by the JNO. Regions having negative contributions are anti-correlated with areas having positive values

is consistent with behavioral reports of reduced visual-motor integration in the ASD literature [98]. Subnetwork 1 has competing from prefrontal and subcortical contributions, mainly the thalamus, amygdala and hippocampus. The thalamus is responsible for relaying sensory and motor signals to the cerebral cortex in the brain. The hippocampus is known to play a crucial role in the consolidation of long and short term memory, along with spatial memory to aid navigation. Altered memory functioning has been shown to manifest in children diagnosed with ASD [151]. Along with the amygdala, which is known to be associated with emotional responses, these areas may be crucial for social-emotional regulation in ASD. Finally, Subnetwork 2 is comprised of competing contributions from the central executive control network and the insula, which is thought to be critical for switching between self-referential and goal-directed behavior [152].

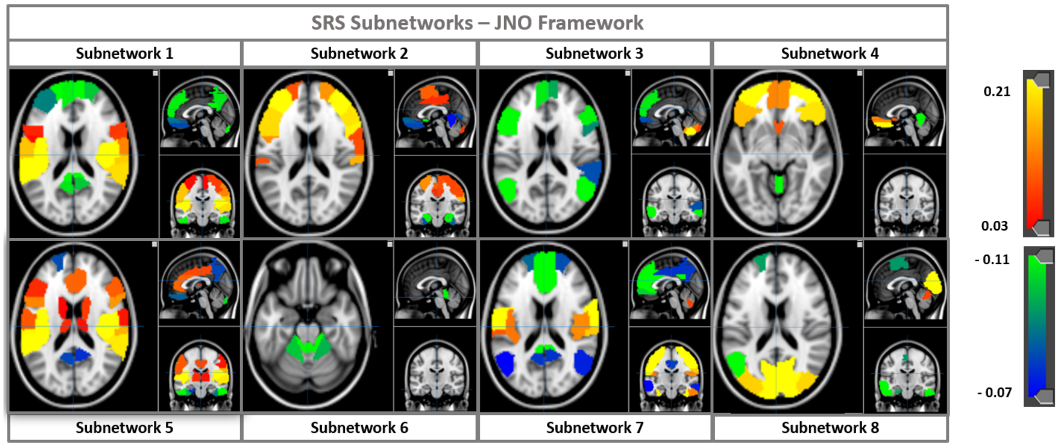


Figure 3.11: Subnetworks estimated to predict the SRS score. Regions having negative contributions are anti-correlated with areas having positive values

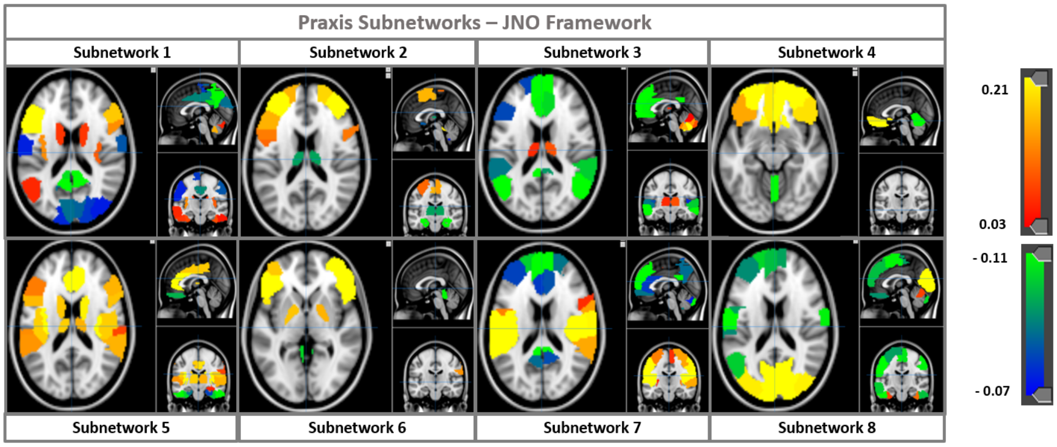


Figure 3.12: Subnetworks estimated to predict the Praxis score. Regions having negative contributions are anti-correlated with areas having positive values

3.8.1 Robustness in Subnetwork Recovery

Notice that we estimate a different basis matrix **B** for each cross validation fold. Therefore, one important property to verify is that these subnetworks are similar across different cohorts of the data.

We observed an average similarity of 0.79±0.06 for the ADOS networks, 0.86±0.04 for the SRS networks, and 0.76±0.06 for the Praxis networks across their cross validation runs. Additionally, upon a cross comparison between the ADOS and SRS networks, we obtained an average similarity of 0.82±0.07. Similarly, the overlap between ADOS and Praxis is 0.79±0.04, and between SRS and Praxis is 0.77±0.06. For a convenient visual inspection, we have arranged the networks in Fig. 3.11 (SRS) and Fig. 3.12 (Praxis) in the order of their inner product similarity with the ADOS networks in Fig. 3.10. This finding strengthens the hypothesis that our model is successful at capturing the stable underlying mechanisms which explain the different sets of deficits of the disorder.

3.8.2 Comparing Subnetwork Representations

In this section, we compare the subnetworks identified by the JNO to the representations learned by the baseline methods. Recall that we have used a regularized linear regression as the Stage 2 predictor for the baselines. Therefore, we can probe the learned regression weights to characterize the baseline network representations.

Degree centrality looks at the relative importance of each brain region or 'node' to the overall representation. To visualize the pattern identified by

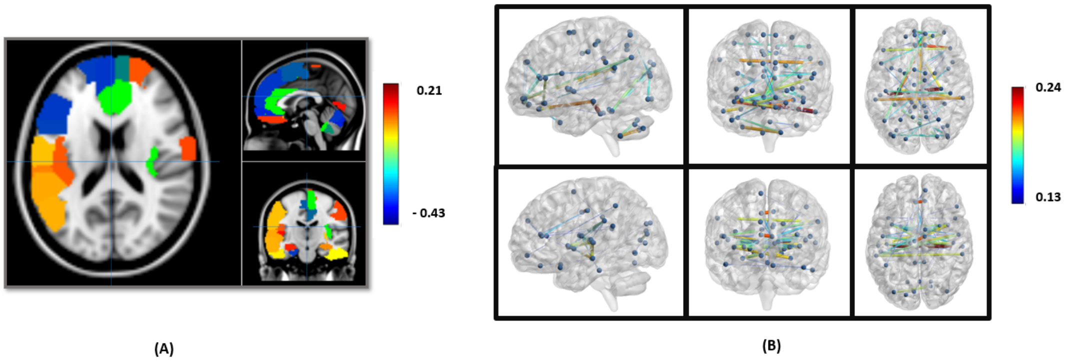


Figure 3.13: (A): Representation learned from the prediction of ADOS by Node degree centrality + ridge regression. The colorbar indicates the weight of the ROI assigned by the ridge regression. **(B):** Top two subnetworks identified by the prediction of the ADOS score by PCA + ridge regression. The colorbar indicates the weight of the connection.

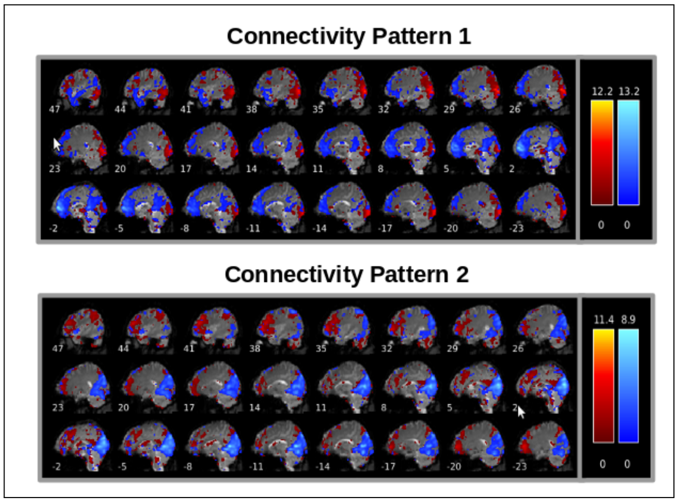


Figure 3.14: Connectivity patterns identified as important in the prediction of the ADOS score by ICA + ridge regression. Each plot displays 2 spatial components contributing to the correlation feature. The colorbar indicates the weight of the connection.

the degree centrality + ridge regression baseline, we display the regression weights on the brain surface plots in Fig. 3.13 (A), normalized to unit norm. The colorbar indicates the strength of co-activation. Regions storing negative values are anticorrelated with regions storing positive weights. We again

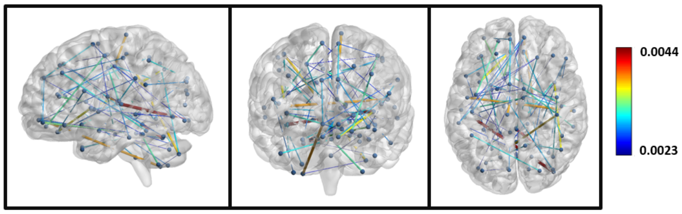


Figure 3.15: Connectivity patterns identified in the prediction of the ADOS score by the ANN. The colorbar indicates the weight of the connections. The narrow range of values are indicative that the ANN assigns equal weighting to most connections on an average

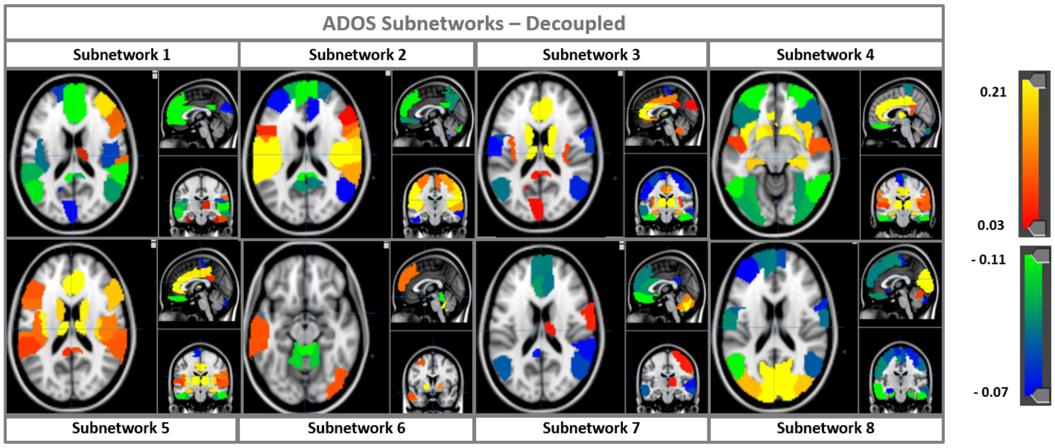


Figure 3.16: Subnetworks estimated to predict ADOS score by decoupling the matrix decomposition and ridge regression. Regions having negative contributions are anti-correlated with areas having positive values

observe patterns from the DMN in the subnetwork plot. Note that the DMN was also a key connectivity pattern identified by the JNO. However, several other subnetworks identified by the JNO do not figure in this representation.

On the other hand, for the PCA + ridge regression baseline, the regression weights inform us of the relative importance of the principal components in prediction. Since the features fed into PCA are the $M = (P \times (P-1))/2$ correlation values, we are left with a 6670 dimensional edge connectivity representation for the AAL per component. We first examine the absolute value of the regression weights learned, and then display the connectivity in the top 2 basis components in Fig. 3.13 (B). We render this connectivity measure using the BrainNet Viewer [153] software. For clarity, we have chosen to display the top 5 percent of the connections obtained. The solid edges signify retained connections, while the blue spheres correspond to nodes of the AAL regions. The colorbar to the right indicates the strength of the connections. We notice that the components consist of several crossing connections spread across different regions of the brain. As compared to our model, which pinpoints key subnetworks already known to be associated with ASD, the representation obtained is not immediately interpretable.

In the ICA + ridge regression baseline, the input to the regression model are the correlation values between the components identified by ICA. After the model is fit, we sort the input correlations based on the learned regression weights. This helps us identify the features important for prediction. In Fig. 3.14 , we display the spatial maps of the top 2 connections identified by the algorithm. We again, observe patterns from the DMN and visual areas. However, it fails to capture several other subnetwork patterns that the JNO identifies as important for ASD.

For the ANN, we use the weight matrix learned at the input layer to inform us of the subnetwork connectivity. Recall that this matrix is of dimension $M \times D$, where $M = (P \times (P-1))/2 = 6670$ for the AAL atlas. For our application, D is of width 8000. We first take the absolute values of these weights, and then normalize the columns of this matrix to unit norm. We then average across the rows to obtain a single 6670 dimensional edge-edge connectivity vector.

Again, we use the BrainNet connectivity plots to display this information in Fig. 3.15. We have chosen to display the top 1 percent of the connections obtained. The solid edges signify retained connections, while the blue spheres correspond to nodes of the AAL regions. The colorbar to the right indicates the strength of the connections. We observe several overlapping connectivity patterns spread across the entire brain despite applying a stringent threshold. Additionally, the narrow range of values indicates that the ANN assigns nearly equal weight to all connections on an average. Similar to the PCA baseline, this representation is unable to capture interpretable connectivity patterns which explain behavior.

Finally, we examine the representation learned by performing the matrix decomposition and prediction separately, i.e. the decoupled case. Note that the learned basis matrix **B** follows the same interpretation as that of the JNO. We display the corresponding co-activation patterns in Fig. 3.16. Again, the colorbar indicates the strength of activation of the AAL ROIs. Negative regions are anticorrelated with the positive regions. For convenience, we have ordered the 8 subnetworks according to their similarity with the ADOS subnetworks identified in Fig. 3.10. Since we use the same matrix decompotion as the JNO, we observe several similarities in the learned representations. We also notice subtle differences in the patterns on account of the coupling with the predictive term in the JNO. We conjecture that these learned differences are what gives the JNO the leverage to generalize to unseen data.

3.9 Discussion

We have shown both predictive power and interpretabilitiy of our model thus far. Furthermore, characterizing model generalizability is important for future application of our framework. Here, we first examine the sensitivity of our prediction results with respect to the model hyperparameters. We discuss mitigation strategies to handle hyperparameter sensitivity that make our framework more robust.

3.9.1 Mitigating Parameter Sensitivity

As initially described in Section 3.5, our JNO framework is insensitive to the regression tradeoff γ and ridge penalty λ_3. We also have a natural way to set the number of subnetworks K. However, we observe that our JNO framework is fairly sensitive to the sparsity on $\mathbf{B}$ and the ridge penalty on the coefficients $\mathbf{c}_n$, i.e. λ_1 and λ_2 respectively. Fig. 3.17 represents the MAE recovery performance of the algorithm for varying settings of λ_1 and λ_2, holding the remaining parameter settings constant when evaluated on the KKI dataset. The red plots in each case indicate the performance of the JNO framework. The $\mathbf{x}$-axis denotes the parameter value, while the $\mathbf{y}$-axis quantifies the MAE from cross validation. Observe that the best λ_1 and λ_2 settings for the individual scores are different, i.e ADOS-$\{\lambda_1 = 30, \lambda_2 = 0.2\}$, SRS-$\{\lambda_1 = 50, \lambda_2 = 0.9\}$, and Praxis-$\{\lambda_1 = 20, \lambda_2 = 0.6\}$. Additionally, the kinks in the plots (shown by the black arrow) also indicate that small changes in the sparsity and coefficient regularization lead to a dramatic change in performance, i.e. the operating points for these two parameters are narrow. We suspect that the hyperparameter differences can be partially attributed to the different dynamic range of each clinical score. Specifically, these differences impact the tradeoff between the representation learning and prediction terms in the JNO optimization. This in turn affects the generalization performance at a particular hyperparameter setting. These observations further illustrate the difficulty of the problem we are trying to address.

We propose two main modifications to tackle the observed hyperparameter sensitivity in λ_1 and λ_2. Given that the dynamic ranges of the scores are quite different and potentially impact generalization, our first mitigation strategy is to rescale the measures to a fixed interval. Since ADOS is the most widely accepted observational measure of clinical autism severity, we have scaled and offset the remaining scores to have a range of 0–30 (similar to ADOS). To mitigate the narrow 'operating point', we include an extra template average correlation term in Eq. (3.5). We now model the residual outer-product terms as deviations around a mean template correlation matrix $\mathbf{B}_{avg}$. The rationale behind this additional term is that it encourage sparsity in the basis matrix along with the explicit ℓ_1 penalty. The modified objective is as follows:

$$\begin{aligned}\mathcal{J}(\mathbf{B}, \mathbf{B}_{avg}, \mathbf{C}, \mathbf{w}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{B}_{avg} - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 \\ + \gamma||\mathbf{y} - \mathbf{C}^T\mathbf{w}||_2^2 + \lambda_1||\mathbf{B}||_1 + \lambda_2||\mathbf{C}||_F^2 + \lambda_3||\mathbf{w}||_2^2 \;\; s.t. \;\; \mathbf{c}_{nk} \geq 0,\end{aligned} \quad (3.19)$$

Notice that $\mathbf{B}_{avg}$ has a closed form update, which does not add much computational overhead. The updates for the remaining variables follow the same procedure as described in Section. 3.4, except that the term, $\{\mathbf{\Gamma}_n\}$ is replaced with $\{\mathbf{\Gamma}_n - \mathbf{B}_{avg}\}$ in every update.

The green plots in Fig. 3.17 illustrate the cross validated performance of the modified JNO framework from Eq. (3.19). The operating point $\{\lambda_1, \lambda_2\}$ for the modified framework is fairly consistent across the scores. Moreover, the green plots exhibit a larger stable range (highlighted in yellow) compared to the red plots. Accordingly, we identify the settings $\{\lambda_1 = 10\text{–}30, \lambda_2 = 0.08\text{–}0.6\}$ as the operating range for the modified JNO objective, which is roughly an order of magnitude larger than the original formulation and does not exhibit any kinks. Fig. 3.18 and Fig. 3.19 illustrate the best generalization performance for SRS and Praxis using the two algorithms.

Notice that the modified JNO has a slight tradeoff in regression performance at the expense of the gain in parameter stability. We highlight the importance of this exploration, as future applications of our work include applying our method to rs-fMRI and severity scores from a variety of neurological disorders. Our modified formulation provides additional flexibility in this sense, and extends the overall generalizability of our model.

3.9.2 Evaluating Generalizability

Finally, notice that the training examples (red points) in Figs. 3.7$-$3.9 follow the $\mathbf{x} = \mathbf{y}$ line nearly perfectly. Here, we explain this (potentially misleading) phenomenon in terms of the parametrizatization of our joint objective in Eq. (3.7).

Recall that Section 3.4.5 describes the procedure for calculating the coefficients for an unseen patient $\bar{\mathbf{c}}_n$ from the training solution set $\{\mathbf{B}^*, \mathbf{w}^*\}$. Recall

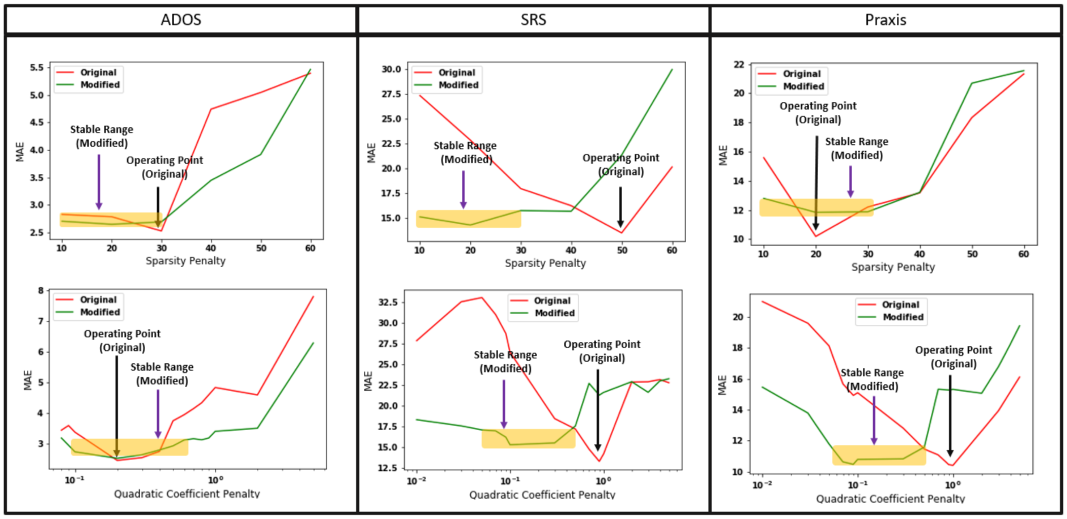


Figure 3.17: Comparing the sensitivity of the JNO framework with the modified objective in Eq. (3.19). Prediction performance with varying **Top** λ_1 for **(L-R):** ADOS, SRS and Praxis **Bottom** λ_2 for **(L-R):** ADOS, SRS and Praxis

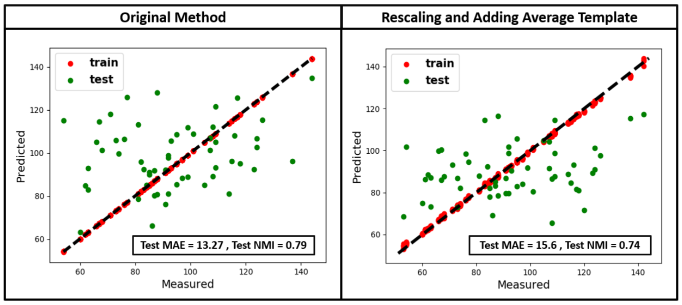


Figure 3.18: A performance comparison for SRS prediction after modifying the objective according to Eq. (3.19). **(L)** Original Method **(R)** After re-scaling and average template addition

that we explicitly set the contribution from the data term in Eq. (3.5) to 0. Since the patient is not a part of the training set, the corresponding value of $\hat{\mathbf{y}}$ is unknown. In contrast, the training performance is computed based on the estimated coefficients $\mathbf{c}_n$, which have access to the severity scores. Here, we examine the effect of removing the severity information when calculating the coefficients for the training patients. In other words, we estimate the

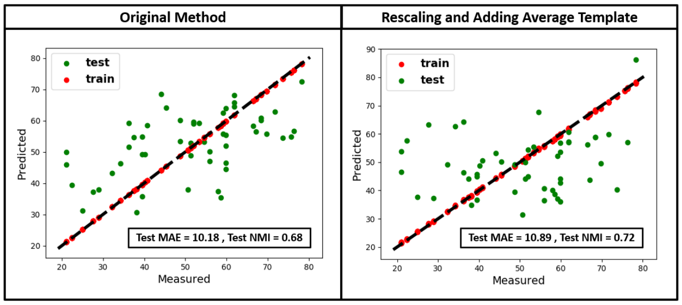


Figure 3.19: A performance comparison for Praxis prediction after modifying the objective according to Eq. (3.19). **(L)** Original Method **(R)** After re-scaling and average template addition

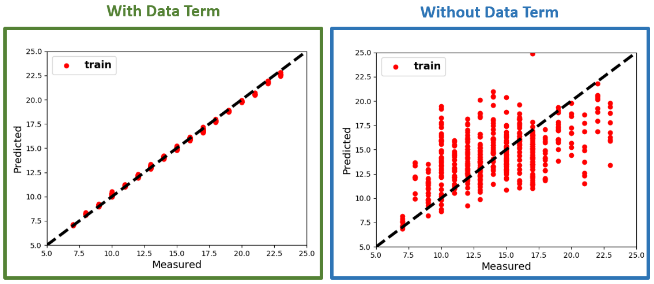


Figure 3.20: Prediction Performance of the JNO for ADOS on training data when **(L)** The data term is included in computing $\mathbf{c}_n$ **(R)** The data term is excluded from the computation of $\mathbf{c}_n$

corresponding severity $\mathbf{y}$ excluding the ridge regression term. Accordingly, Fig. 3.20 highlights the differences in training fit with and without this term is not included in estimating $\mathbf{c}_n$. Notice that in the latter, the training accuracy has the same distribution as the testing points in Figs. 3.7—3.9. Taken together, we conclude that, the linear predictive term overparamterizes the search space of solutions for $\mathbf{c}_n$ to yield a near perfect fit. We use this observation to emphasize that the subnetworks and regression model learned by our JNO framework are capturing the underlying data distribution and not simply 'overfitting' the training data.

Lastly, our paper [39] evaluates our model on a second ASD cohort acquired at the NYU site within the ABIDE database [154]. It also evaluates the effect of changing the parcellation scheme of choice, i.e. effect of changing resolution of the functional connectivity data. In each comparison, we observe that the JNO framework provides consistent improvements over several baselines. More importantly, [39] presents additional results on test-retest reliability by examining cross-site generalizability and robustness to the choice of parcellation scheme. All of these comparisons allude to the efficacy of the JNO at reliably extracting representations from rs-fMRI connectivity data that are explanative of clinical measures.

3.10 Summary and Conclusion

Our JNO model cleverly exploits the structure intrinsic to rs-fMRI correlation matrices through an outer product representation. The regression term further guides the basis decomposition to explain the group level and patient specific information. The compactness of our representation serves as a dimensionality reduction step that is related to the clinical score of interest, unlike the pipelined treatment commonly found in the literature. As seen from the results, our JNO framework outperforms a wide range of well established baselines from the machine learning and graph theoretic methods ubiquitous in fMRI analysis on two separate real world datasets.

We conjecture that the baseline techniques fail to extract representative patterns from the correlation data, and learn only the group level representation for the cohort. Consequently, they overfit the training set, despite sweeping the parameters across several orders of magnitude. Any patient level symptomatic and connectivity level differences are lost due to the restrictive pipelined procedure and the group level confounds.

Our Joint Network Optimization Framework is also agnostic to the choice of parcellation scheme. We have demonstrated this by our additional experiments on the KKI dataset, where we chose the 246 region Brainnetome parcellation to extract correlation matrices (see Section 3.6 in [39]). We further emphasize that our framework makes minimal assumptions on the data. Provided we have access to a valid behavioral and network similarity measure, this analysis can be easily adapted to other neurological disorders and even predictive network models outside the medical realm. This greatly broadens the scope of the method to numerous potential applications.

3.10.1 Limitations and Scope for Refinement

From the behavioral standpoint, the JNO focuses on predicting scalar measures of severity individually rather than a collective. However, it is known that complex disorders such as ASD are inherently multi-dimensional in manifestation.

A natural direction of exploration would thus be a simple multi-score extension which can incorporate data from different behavioral domains. Unfortunately, a naive multi-output modification of Eq. (3.5) performs poorly on this task. An alternative would be to replace the linear regression term in Eq. (3.5) with more powerful non-linear counterparts, thus providing us with the flexibility to model more complex decision functions which can better map the behavioral space. This paves way for our discussion in Chapter 4, where we refine our discriminative models to explore non-parametric and neural network regressors.

Another avenue for refinement is to incorporate structural connectivity information in the form of anatomical priors from Diffusion Tensor Imaging (DTI). As mentioned previously, these scans are used to define and track existing anatomical pathways in the brain. Furthermore, this work analyzes functional connectivity as a static snapshot, rather than an evolving process over the scan duration. Incorporating this information into the network optimization model could be an important step towards unifying anatomical, functional and behavioral domains to better understand altered brain functioning in the context of neurological disorders such as Autism, ADHD, and Schizophrenia. This would require us to refine the generative model we have developed for the neuroimaging data. We reserve this discussion to Chapter 5, which builds on the ideas we just presented to develop a model capable of parsing multimodal and dynamic connectivity simultaneously.

Chapter 4

Beyond Linear Regression Models

This chapter explores two mathematical extensions to the JNO framework from Chapter 3. Our main motivation is to improve the representational flexibility of the discriminative model. In essence, we would like to have the ability to model more complex relationships between the low dimensional neuroimaging space and the behavioral space.

Outline: Section 4.1 presents a coupled manifold optimization framework that combines non-parametric regression with the dictionary learning on the rs-fMRI correlation matrices. Going a step further, Section 4.2 presents a technique that marries classical representation learning with neural network predictors into the same optimization.

4.1 CMO: A Coupled Manifold Optimization Framework for Connectomics and Behavior

In this section, we borrow ideas from non-parametric models and manifold learning to extend our discriminative model.

Numerous non-parametric approaches have been employed to study complex brain topologies, especially in the context of disease classification. For supervised learning, the most popular classifier is a support vector machine (SVM) [97], which optimizes a separating hyperplane between two [96] or more [60] classes. These hyperplanes may be defined either in the native space (linear SVM) of rs-fMRI features or in a contrived higher dimensional space (kernel SVM) non-parametrically. Alternatively, the work of [155] used graph kernels on the spatio-temporal fMRI time series dynamics to distinguish between the autistic and healthy groups. Going one step further, [156] used higher order morphological kernels to classify ASD sub-populations.

While these methods are computationally efficient and simple in formulation, their generalization power is limited by the input data features. Often, subtle individual level changes are overwhelmed by group level confounds. To this end, we again take the approach integrate the feature learning step directly into our framework. We simultaneously optimize both the embeddings and the projection onto the behavioral space. Since this optimization is also coupled to the brain basis, it helps us model the behavioral and neuroimaging data space jointly by reliably capturing individual variability. We leverage the kernel trick to provide both the representational flexibility and computational tractability to outperform a variety of baselines. The rest of this section is based on work which appeared in this conference paper [157].

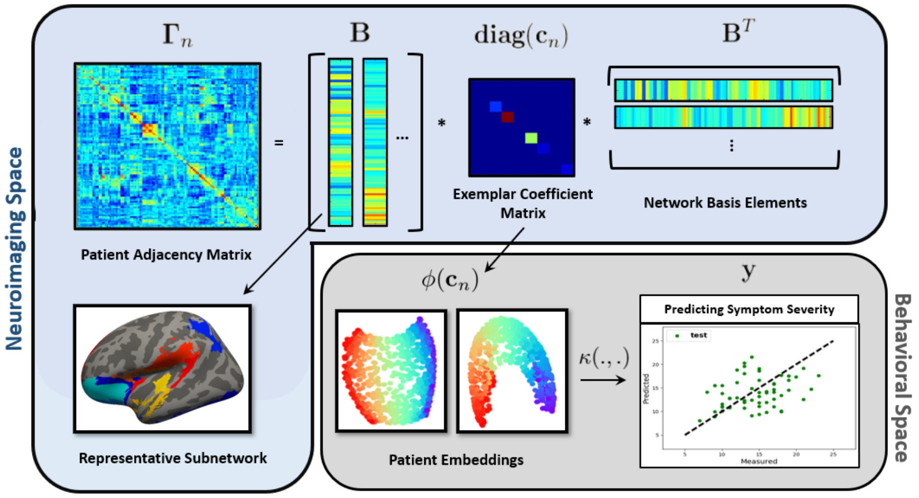


Figure 4.1: Joint Model for the Functional Connectomics and Behavioral Data. **Blue Box:** Matrix Manifold Representation **Gray Box:** Non-Linear kernel Ridge Regression

4.1.1 Non-Parametric Regression Model

Fig. 4.1 presents our Coupled Manifold Optimization (CMO) framework. The blue box represents our neuroimaging term. Since we group voxels into P ROIs, this yields the $P \times P$ input correlation matrices $\{\mathbf{\Gamma}_n\}_{n=1}^{N}$ for N patients. As seen, the correlation matrices are projected onto a low rank subspace spanned by the group basis. The loadings are related to severity via a non-linear manifold and the associated kernel map, as indicated in the gray box.

Recall that $\mathbf{\Gamma}_n$ is positive semi-definite by construction. Again, we employ a patient specific low rank decomposition $\mathbf{\Gamma}_n \approx \mathbf{Q}_n\mathbf{Q}_n^T$ to represent the correlation matrix. Each rank K factor $\{\mathbf{Q}_n \in \mathcal{R}^{P\times K}\}$, where $K \ll P$, projects onto a low dimensional subspace spanned by the columns of a group basis $\mathbf{B} \in \mathcal{R}^{P\times K}$. The vector $\mathbf{c}_n \in \mathcal{R}^{K\times 1}$ denotes the patient specific loading coefficients as follows:

$$\mathbf{\Gamma}_n \approx \mathbf{Q}_n\mathbf{Q}_n^T = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T \tag{4.1}$$

where $\mathbf{diag}(\mathbf{c}_n)$ is a matrix with the entries of $\mathbf{c}_n$ on the leading diagonal, and the off-diagonal elements as 0. Eq. (4.1) resembles a joint eigenvalue decomposition for the set $\{\mathbf{\Gamma}_n\}$, similar to one provided by Common Principal Components [158]. The bases $\mathbf{b}_k \in \mathcal{R}^{P\times 1}$ capture co-activation patterns common to the group, while the coefficient loadings $\mathbf{c}_{nk}$ capture the strength of basis column k for patient n.

We use these coefficients to predict clinical severity via a non-linear manifold. We define an embedding map $\boldsymbol{\phi}(\cdot) : \mathcal{R}^K \rightarrow \mathcal{R}^M$, which maps the native space representation of the coefficient vector $\mathbf{c}$ to an M dimensional embedding space, i.e. $\boldsymbol{\phi}(\mathbf{c}_n) \in \mathcal{R}^{M\times 1}$. If $\mathbf{y}_n$ is the clinical score for patient n, we have the non-linear regression:

$$\mathbf{y}_n \approx \boldsymbol{\phi}(\mathbf{c}_n)^T\mathbf{w} \tag{4.2}$$

with weight vector $\mathbf{w} \in \mathcal{R}^{M\times 1}$. Once again, our joint objective combines Eq. (4.1) and Eq. (4.2)

$$\mathcal{J}(\mathbf{B}, \{\mathbf{c}_n\}, \mathbf{w}) = \sum_n \left[||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 + \lambda||\mathbf{y}_n - \boldsymbol{\phi}(\mathbf{c}_n)^T\mathbf{w}||_2^2 \right] \tag{4.3}$$

along with the constraint $\mathbf{c}_{nk} \geq 0$ to maintain positive semi-definiteness of $\{\mathbf{\Gamma}_n\}$. Here, λ controls the trade-off between the two representations. We include an ℓ_1 penalty on $\mathbf{B}$ to promote sparse solutions for the basis. We also regularize both the coefficients $\{\mathbf{c}_n\}$ and the regression weights $\mathbf{w}$ with

ℓ_2 penalties to ensure that the objective is well posed. We add the terms $\gamma_1||\mathbf{B}||_1 + \gamma_2\sum_n ||\mathbf{c}_n||_2^2 + \gamma_3||\mathbf{w}||_2^2$ to $\mathcal{J}(\cdot)$ in Eq. (4.3) with the penalties γ_1,γ_2 and γ_3 respectively.

4.1.2 Joint Inference Strategy

Again, we use alternating minimization to estimate the hidden variables $\{\mathbf{B}, \{\mathbf{c}_n\}, \mathbf{w}\}$. This procedure iteratively optimizes each unknown variable in Eq. (4.3) by holding the others constant until global convergence is reached.

Proximal gradient descent [147] is an efficient algorithm which provides good convergence guarantees for the non-differentiable ℓ_1 penalty on $\mathbf{B}$. However, it requires the objective to be convex in $\mathbf{B}$, which is not the case due to the bi-quadratic Frobenius norm expansion in Eq. (4.3). Hence, we introduce N constraints of the form $\mathbf{D}_n = \mathbf{Bdiag}(\mathbf{c}_n)$, similar to our work in [146, 39]. We enforce these constraints using the Augmented Lagrangians $\{\mathbf{\Lambda}_n\}$:

$$\mathcal{J}(\mathbf{B}, \{\mathbf{c}_n\}, \mathbf{w}, \{\mathbf{D}_n\}, \{\mathbf{\Lambda}_n\}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{D}_n\mathbf{B}^T||_F^2 + \lambda\sum_n ||\mathbf{y}_n - \boldsymbol{\phi}(\mathbf{c}_n)^T\mathbf{w}||_2^2$$
$$+ \sum_n \left[\mathrm{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n))\right] + \frac{1}{2}||\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n)||_F^2\right] \quad (4.4)$$

with $\mathbf{c}_{nk} \geq 0$. The additional terms $||\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n)||_F^2$ regularize the trace constraints. Eq. (4.4) is now convex in both $\mathbf{B}$ and the set $\{\mathbf{D}_n\}$, which allows us to optimize them via standard procedures.

We iterate through the following four update steps till global convergence:

4.1.2.1 Proximal Gradient Descent on B

The gradient of $\mathcal{J}$ with respect to $\mathbf{B}$ is:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{B}} = \sum_{n} 2\left[\mathbf{B}\mathbf{D}_n^T - \mathbf{\Gamma}_n\right]\mathbf{D}_n - \mathbf{D}_n\mathbf{diag}(\mathbf{c}_n) + \mathbf{B}\mathbf{diag}(\mathbf{c}_n)^2 - \mathbf{\Lambda}_n\mathbf{diag}(\mathbf{c}_n)$$

With a learning rate t, the proximal update with respect to $||\mathbf{B}||_1$ is given by:

$$\mathbf{B}^k = \mathbf{prox}_{||\cdot||_1}\left[\mathbf{B}^{k-1} - \left[\frac{t}{\gamma_1}\right]\frac{\partial \mathcal{J}}{\partial \mathbf{B}}\right] \; s.t. \; \mathbf{prox}_t(\mathbf{L}) = \mathbf{sgn}(\mathbf{L}) \circ (\mathbf{max}(|\mathbf{L}| - t, \mathbf{0})) \tag{4.5}$$

Effectively, this update performs an iterative shrinkage thresholding on a locally smooth quadratic model of $||\mathbf{B}||_1$.

4.1.2.2 Kernel Ridge Regression for w:

We denote $\mathbf{y}$ as the vector of the clinical severity scores and stack the patient embedding vectors i.e. $\boldsymbol{\phi}(\mathbf{c}_n) \in \mathcal{R}^{M\times 1}$ into a matrix $\mathbf{\Phi}(\mathbf{C}) \in \mathcal{R}^{M\times N}$. The portion of $\mathcal{J}(\cdot)$ that depends on $\mathbf{w}$ is:

$$\mathcal{F}(\mathbf{w}) = \lambda||\mathbf{y} - \mathbf{\Phi}(\mathbf{C})^T\mathbf{w}||_2^2 + \gamma_3||\mathbf{w}||_2^2 \tag{4.6}$$

Setting the gradient of Eq. (4.6) to 0, and applying the matrix inversion lemma, the closed form solution for $\mathbf{w}$ is similar to kernel ridge regression:

$$\mathbf{w} = \mathbf{\Phi}(\mathbf{C})\left[\mathbf{\Phi}(\mathbf{C})^T\mathbf{\Phi}(\mathbf{C}) + \frac{\gamma_3}{\lambda}\mathcal{I}_N\right]^{-1}\mathbf{y} = \mathbf{\Phi}(\mathbf{C})\boldsymbol{\alpha} = \sum_j \boldsymbol{\alpha}_j\boldsymbol{\phi}(\mathbf{c}_j) \tag{4.7}$$

where $\mathcal{I}_N$ is the identity matrix. Let $\kappa(\cdot,\cdot) : \mathcal{R}^M \times \mathcal{R}^M \rightarrow \mathcal{R}$ be the kernel map for $\boldsymbol{\phi}$, i.e. $\kappa(\mathbf{c}, \hat{\mathbf{c}}) = \boldsymbol{\phi}(\mathbf{c})^T\boldsymbol{\phi}(\hat{\mathbf{c}})$. The dual variable $\boldsymbol{\alpha}$ can be expressed as $\boldsymbol{\alpha} = (\mathbf{K} + \frac{\gamma_3}{\lambda}\mathcal{I}_N)^{-1}\mathbf{y}$, where $\mathbf{K} = \mathbf{\Phi}(\mathbf{C})^T\mathbf{\Phi}(\mathbf{C})$ is the Gram matrix for the kernel $\boldsymbol{\kappa}(\cdot,\cdot)$. Eq. (4.7) implies that $\mathbf{w}$ lies in the span of the coefficient embeddings defining the manifold. We use the form of $\mathbf{w}$ in Eq. (4.7) to update the loading vectors in the following step, without explicitly parametrizing the vector $\boldsymbol{\phi}(\mathbf{c}_n)$.

4.1.2.3 Trust Region Update for $\{\mathbf{c}_n\}$

The objective function for each patient loading vector $\mathbf{c}_n$ decouples as follows when the other variables are fixed:

$$\mathcal{F}(\mathbf{c}_n) = \lambda||\mathbf{y}_n - \phi(\mathbf{c}_n)^T\mathbf{w}||_2^2 + \gamma_2||\mathbf{c}_n||_2^2 + \text{Tr}\left[\boldsymbol{\Lambda}_n^T(\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n))\right] \\ + \frac{1}{2}||\mathbf{D}_n - \mathbf{Bdiag}(\mathbf{c}_n)||_F^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \quad (4.8)$$

We now substitute this form into Eq. (4.8) and use the kernel trick, to write:

$$||\mathbf{y}_n - \boldsymbol{\phi}(\mathbf{c}_n)^T\mathbf{w}||_2^2 = ||\mathbf{y}_n - \sum_j \boldsymbol{\phi}(\mathbf{c}_n)^T\boldsymbol{\phi}(\hat{\mathbf{c}}_\mathbf{j})\boldsymbol{\alpha}_j||_2^2 = ||\mathbf{y}_n - \sum_j \kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)\boldsymbol{\alpha}_j||_2^2$$

where $\{\hat{\mathbf{c}}_\mathbf{n}\}$ denotes the coefficient vector estimates from the previous step to compute $\mathbf{w}$. Notice that the kernel trick buys a second advantage, in that we only need to optimize over the first argument of $\kappa(\cdot,\cdot)$. Since kernel functions typically have a nice analytic form, we can easily compute the gradient $\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)$ and hessian $\nabla^2\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)$ of $\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_j)$ with respect to $\mathbf{c}_n$.

Given this, the gradient of $\mathcal{F}(\cdot)$ with respect to $\mathbf{c}_n$ takes the following form:

$$\mathbf{g}_n = \frac{\partial \mathcal{F}}{\partial \mathbf{c}_n} = \mathbf{c}_n \circ \left[\left[\mathcal{I}_R \circ (\mathbf{B}^T \mathbf{B}) \right] \mathbf{1} \right] - \left[\mathcal{I}_R \circ (\boldsymbol{\Lambda}_n^T \mathbf{B} + \mathbf{D}_n^T \mathbf{B}) \right] \mathbf{1} + 2\gamma_2 \mathbf{c}_n$$

$$- \lambda \sum_i \boldsymbol{\alpha}_i \left[2\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_i)\mathbf{y}_i - \sum_k \boldsymbol{\alpha}_k \left[\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_i)\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_k) + \kappa(\mathbf{c}_n, \hat{\mathbf{c}}_k)\nabla\kappa(\mathbf{c}_n, \hat{\mathbf{c}}_i) \right] \right]$$

where $\mathbf{1}$ is the vector of all ones. Notice that the top line of the gradient term is from the matrix decomposition and regularization terms, and the bottom line corresponds to the kernel regression. The Hessian $\mathbf{H}_n = \partial^2 \mathcal{F} / \partial \mathbf{c}_n^2$ can be similarly computed.

Given the low dimensionality of $\mathbf{c}_n$, we derive a trust region optimizer for this variable. The trust region algorithm provides guaranteed convergence, like the popular gradient descent method, with the speedup of second-order procedures. The algorithm iteratively updates $\mathbf{c}_n$ according to the descent direction $\mathbf{p}_k$, i.e. $\mathbf{c}_n^{(k+1)} = \mathbf{c}_n^{(k)} + \mathbf{p}_k$. The vector $\mathbf{p}_k$ is computed via the following quadratic objective, which is a second order Taylor expansion of $\mathcal{F}$ around $\mathbf{c}_n^k$:

$$\mathbf{p}^* = \text{argmin}_{\mathbf{p}} \mathcal{F}(\mathbf{c}_n^k) + \mathbf{g}_n^k(\mathbf{c}_n^k)^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \mathbf{H}_n^k(\mathbf{c}_n^k)\mathbf{p} \;\; s.t. \; ||\mathbf{p}||_2 \leq \delta_k \;, \; \mathbf{c}_{nr}^k + \mathbf{p}_r \geq 0$$

where $\mathbf{g}_n(\cdot)$ and $\mathbf{H}_n(\cdot)$ are the gradient and Hessian referenced above evaluated at the current iterate $\mathbf{c}_n^k$. We recursively search for a suitable trust region radius δ_k such that we are guaranteed sufficient decrease in the objective at each iteration. This algorithm has a lower bound on the function decrease per update, and with an appropriate choice of the δ_k, converges to a local minimum of $\mathcal{F}(\cdot)$ [159].

4.1.2.4 Augmented Lagrangian Update for $\mathbf{D}_n$ and $\mathbf{\Lambda}_n$

Each $\{\mathbf{D}_n\}$ has a closed form solution, while the dual variables $\{\mathbf{\Lambda}_n\}$ are updated via gradient ascent:

$$\mathbf{D}_n = (\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + 2\mathbf{\Gamma}_n\mathbf{B} - \mathbf{\Lambda}_n)(\mathcal{I}_K + 2\mathbf{B}^T\mathbf{B})^{-1} \tag{4.9}$$

$$\mathbf{\Lambda}_n^{k+1} = \mathbf{\Lambda}_n^k + \eta_k(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)) \tag{4.10}$$

We cycle through the updates in Eqs. (4.9-4.10) to ensure that the proximal constraints are satisfied with increasing certainty at each step. We choose the learning rate parameter η_k for the gradient ascent step of the Augmented Lagrangian to guarantee sufficient decrease for every iteration of alternating minimization.

4.1.2.5 Prediction on unseen data

We use the estimates $\{\mathbf{B}^*, \mathbf{w}^*, \{\mathbf{c}_n^*\}\}$ obtained from the training data to compute the loading vector $\bar{\mathbf{c}}$ for an unseen patient. We must remove the data term in Eq. (4.4), as the corresponding value of $\bar{\mathbf{y}}$ is unknown for the new patient. Hence, the kernel terms in the gradient and hessian disappear. We also assume that the conditions for the proximal operator hold with equality; this eliminates the Augmented Lagrangians in the computation. The objective in $\bar{\mathbf{c}}$ reduces to the following quadratic form:

$$\frac{1}{2}\bar{\mathbf{c}}^T\bar{\mathbf{H}}\bar{\mathbf{c}} + \bar{\mathbf{f}}^T\bar{\mathbf{c}} \;\; s.t. \;\; \bar{\mathbf{A}}\bar{\mathbf{c}} \leq \bar{\mathbf{b}} \tag{4.11}$$

Note that the formulation is similar to the trust region update we used previously. For an unseen patient, the parameters from Eq. (4.11) are:

$$\bar{\mathbf{H}} = 2(\mathbf{B}^T\mathbf{B}) \circ (\mathbf{B}^T\mathbf{B}) + 2\gamma_2\mathcal{I}_K$$

$$\bar{\mathbf{f}} = -2\mathcal{I}_K \circ (\mathbf{B}^T\mathbf{\Gamma}_n\mathbf{B})\mathbf{1};\;\; \bar{\mathbf{A}} = -\mathcal{I}_K \;\; \bar{\mathbf{b}} = \mathbf{0}$$

The Hessian $\bar{\mathbf{H}}$ is positive definite, which leads to an efficient quadratic programming solution to Eq. (4.11). The severity score for the test patient is $\bar{\mathbf{y}} = \boldsymbol{\phi}(\bar{\mathbf{c}})^T\mathbf{w}^* = \sum_j \kappa(\bar{\mathbf{c}}, \mathbf{c}_j^*)\boldsymbol{\alpha}_j^*$, where $\boldsymbol{\alpha}^* = \left[\mathbf{\Phi}(\mathbf{C}^*)^T\mathbf{\Phi}(\mathbf{C}^*) + \frac{\gamma_3}{\lambda}\mathcal{I}_N\right]^{-1}\mathbf{y}$.

4.1.3 Model Evaluation

Our evaluation strategy for the CMO is similar in spirit to Chapter 3, except that we focus on non-parametric regression models in Stage 2. We again employ a ten fold cross validation strategy, and employ the Median Absolute Error and Mutual Information metric for numerical comparison.

4.1.3.1 Baselines

We compare our algorithm with the standard manifold learning pipeline to predict the target severity score. We consider two classes of representation learning techniques motivated from the machine learning and graph theoretic literature. We construct a non-linear regression model similar to manifold learning term in Eq. (4.3). Our five baseline comparisons are as follows:

1. Principal Component Analysis (PCA) on the stacked $\frac{P\times(P-1)}{2}$ correlation coefficients followed by a kernel ridge regression (kRR) on the projections

2. Kernel Principal Principal Component Analysis (kPCA) on the correlation coefficients followed by a kRR on the embeddings

3. Node Degree computation (D_N) based on the thresholded correlation matrices followed by a kRR on the P node features

4. Betweenness Centrality (C_B) on the thresholded correlation matrices followed by a kRR on the P node features

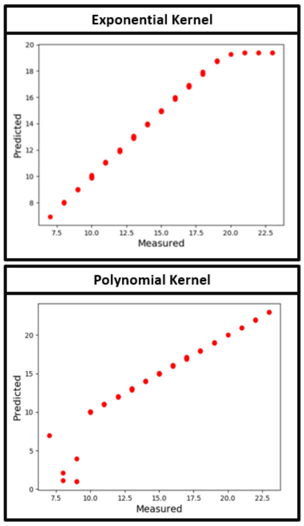
5. Decoupled Matrix Decomposition (Eq.(4.3)) and kRR on the loadings $\{\mathbf{c}_n\}$.

Baseline 5 helps us evaluate and quantify the advantage provided by our joint optimization approach as opposed to a pipelined prediction of clinical severity.

4.1.3.2 Implementation Details

We evaluate every algorithm in a ten fold cross validation setting, i.e. we train the model on a 90 percent split of our data, and report the performance on the unseen 10 percent.

The number of components was fixed at 15 for PCA and at 10 for k-PCA. For k-PCA, we use an RBF kernel with the coefficient parameter 0.1. There are two free parameters for the kRR, namely, the kernel parameter C and ℓ_2 parameter β. We obtain the best performance for the following settings: ADOS $\{C = 0.1, \beta = 0.2\}$, SRS $\{C = 0.1, \beta = 0.8\}$, and Praxis $\{C = 0.01, \beta = 0.2\}$. For the graph theoretic baselines, we obtained the best performance by thresholding the entries of $\{\mathbf{\Gamma}_n\}$ at 0.2. We fixed the parameters in our CMO



Exponential Kernel

Polynomial Kernel

Figure 4.2: Recovery **Top:** Exponential **Bottom:** Polynomial Kernel

framework using a grid search for $\{\lambda, \gamma_1, \gamma_2, \gamma_3\}$. The values were varied between $(10^{-3} - 10)$. The performance is insensitive to λ and γ_3, which are fixed at 1. The remaining parameters were set at $\{\gamma_1 = 10, \gamma_2 = 0.7, \gamma_3 = 1\}$ for all the scores. We fix the number of networks, K, at the knee point of the eigenspectrum of $\{\mathbf{\Gamma}_n\}$, i.e. $(K = 8)$.

Based on simulated data, we observed that the standard exponential kernel provides a good recovery performance in the lower part of the dynamic range, while polynomial kernels are more suited for modeling the larger behavioral scores, as shown in Fig 4.2. Thus, we use a mixture of both kernels to capture the complete behavioral characteristics:

$$\kappa(\mathbf{c}_i, \mathbf{c}_j) = \exp\left[-\frac{||\mathbf{c}_i - \mathbf{c}_j||_2^2}{\sigma^2}\right] + \frac{\rho}{l}\left(\mathbf{c}_j^T \mathbf{c}_i + 1\right)^l$$

We vary the kernel parameters across 2 orders of magnitude and select the settings: ADOS $\{\sigma^2 = 1, \rho = 0.8, l = 2.5\}$, SRS $\{\sigma^2 = 1, \rho = 2, l = 1.5\}$ and

Praxis $\{\sigma^2 = 1, \rho = 0.5, l = 1.5\}$. The varying polynomial orders reflect the differences in the dynamic ranges of the scores.

4.1.4 Experiments on Real Data

Fig. 4.3, Fig. 4.4, and Fig. 4.5 illustrate the regression performance for ADOS, SRS, and Praxis respectively. The bold $\mathbf{x} = \mathbf{y}$ line indicates ideal performance. The red points denote the training fit, while the blue points indicate testing performance. Note that baseline testing performance tracks the mean value of the data (indicated by the horizontal black line). In comparison, our method not only consistently fits the training set more faithfully, but also generalizes much better to unseen data. We emphasize that even the pipelined treatment

Table 4.1: Performance evaluation using **Median Absolute Error (MAE)** & **Mutual Information (MI)**. Lower MAE & higher MI indicate better performance.

Score	Method	MAE Train	MAE Test	MI Train	MI Test
ADOS	PCA & kRR	1.29	3.05	1.46	0.87
	k-PCA & kRR	1.00	2.94	1.48	0.38
	C_B & kRR	2.10	2.93	1.03	0.95
	D_N & kRR	2.09	3.03	0.97	0.96
	Decoupled	2.11	3.11	0.82	1.24
	CMO Framework	**0.035**	**2.73**	**3.79**	**2.10**
SRS	PCA & kRR	7.39	19.70	2.78	3.30
	k-PCA & kRR	5.68	18.92	2.85	1.74
	C_B & kRR	11.00	17.72	2.32	3.66
	D_N & kRR	11.46	17.79	2.24	3.60
	Decoupled	15.9	18.61	2.04	3.71
	CMO Framework	**0.09**	**13.28**	**5.28**	**4.36**
Praxis	PCA & kRR	5.33	12.5	2.50	2.68
	k-PCA & kRR	4.56	11.15	2.56	1.51
	C_B & kRR	8.17	12.61	1.99	3.05
	D_N & kRR	8.18	13.14	2.00	3.20
	Decoupled	10.11	13.33	3.28	1.53
	CMO Framework	**0.13**	**9.07**	**4.67**	**3.87**

using the matrix decomposition in Eq. (4.3), followed by a kernel ridge regression on the learnt projections fails to generalize. This finding makes a strong case for coupling the two representation terms in our CMO strategy. We conjecture that the baselines fail to capture representative connectivity patterns that explain both the functional neuroimaging data space and the patient behavioral heterogeneity. On the other hand, our CMO framework leverages the underlying structure of the correlation matrices through the basis manifold representation. At the same time, it seeks those embedding directions that are predictive of behavior. As reported in Table 4.1, our method quantitatively outperforms the baselines approaches, in terms of both the Median Absolute Error (MAE) and the Mutual Information (MI) metrics. The

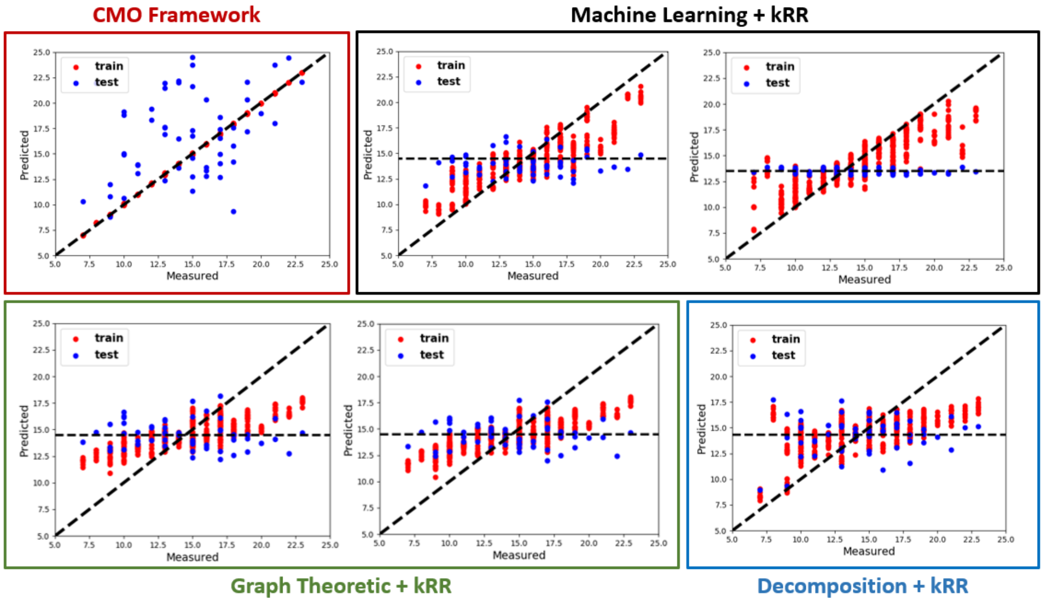


Figure 4.3: Prediction performance for the ADOS score for **Red Box:** CMO Framework. **Black Box: (L)** PCA and kRR **(R)** k-PCA and kRR, **Green Box: (L)** Node Degree Centrality and kRR **(R)** Betweenness Centrality and kRR **Blue Box:** Matrix Decomposition from Eq. (4.3) followed by kRR

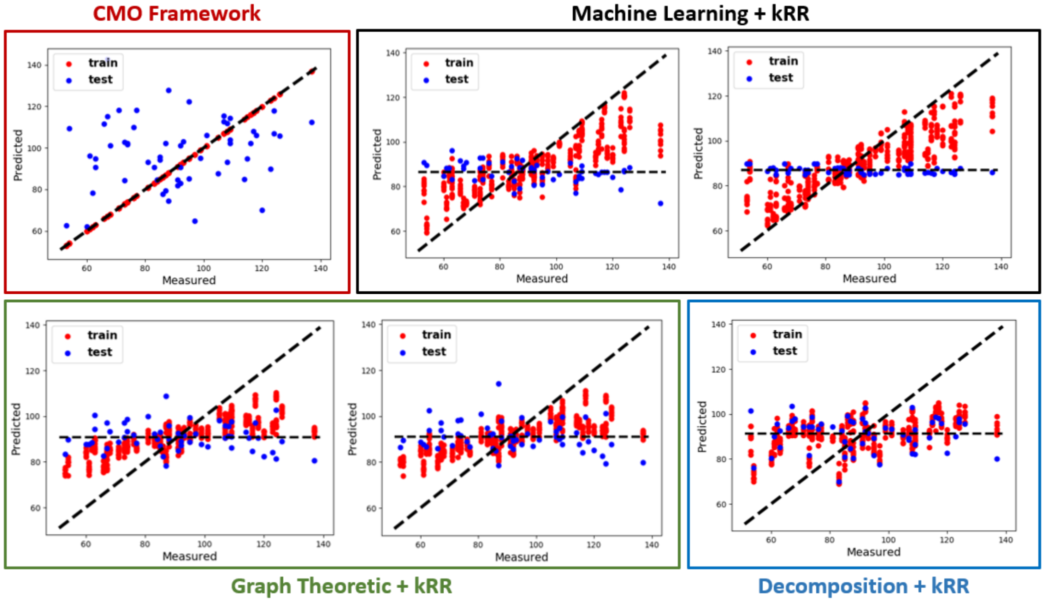


Figure 4.4: Prediction performance for the SRS score for **Red Box:** CMO Framework. **Black Box: (L)** PCA and kRR **(R)** k-PCA and kRR, **Green Box: (L)** Node Degree Centrality and kRR **(R)** Betweenness Centrality and kRR **Blue Box:** Matrix Decomposition from Eq. (4.3) followed by kRR

CMO also provides comparable performance with the JNO on all performance measures.

4.1.5 Clinical Interpretation

Fig. 4.6 illustrates the subnetworks $\{\mathbf{B}_k\}$ trained on ADOS. The colorbar indicates subnetwork contributions to the AAL regions. Regions storing negative values are anticorrelated with positive regions. From a clinical standpoint, Subnetwork 4 includes the somatomotor network (SMN) and competing i.e. anticorrelated contributions from the default mode network (DMN), previously reported in ASD [98]. Subnetwork 8 comprises of the SMN and competing contributions from the higher order visual processing areas in the

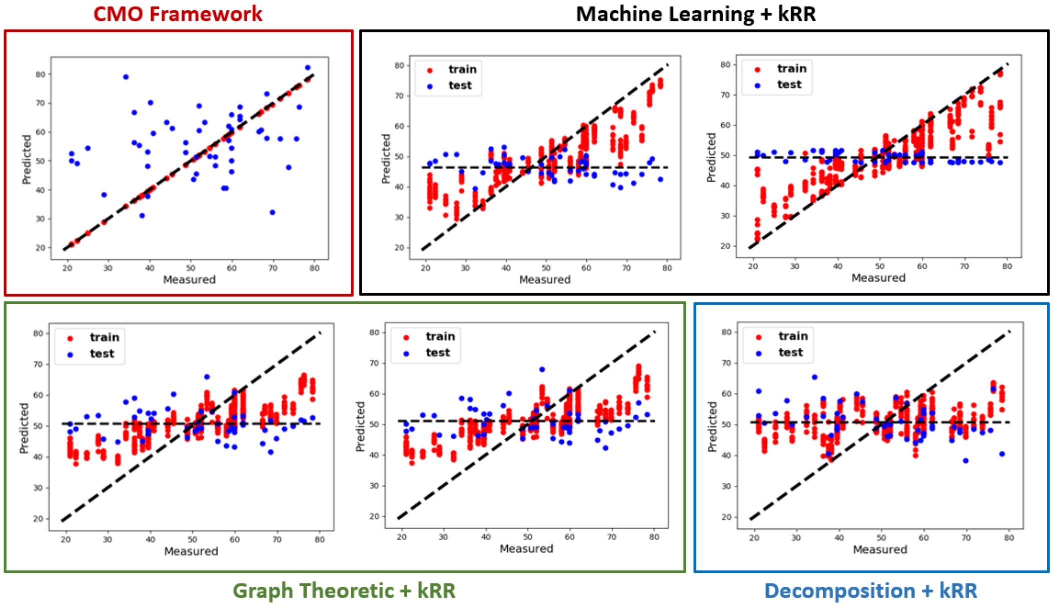


Figure 4.5: Prediction performance for the Praxis score for **Red Box:** CMO Framework. **Black Box:** **(L)** PCA and kRR **(R)** k-PCA and kRR, **Green Box:** **(L)** Node Degree Centrality and kRR **(R)** Betweenness Centrality and kRR **Blue Box:** Matrix Decomposition from Eq. (4.3) followed by kRR

occipital and temporal lobes. These findings are in line with behavioral reports of reduced visual-motor integration in ASD [98]. Though not evident from the surface plots, Subnetwork 5 includes anticorrelated contributions from subcortical regions, mainly, the amygdala and hippocampus, believed to be important for socio-emotional regulation in ASD. Finally, Subnetwork 6 has competing contributions from the central executive control network and insula, which are critical for switching between self-referential and goal-directed behavior [152]. Fig. 4.7 compares Subnetwork 2 obtained from ADOS, SRS and Praxis prediction. There is a significant overlap in the bases subnetworks obtained by training across the different scores. Additionally, several subnetworks that we extract are shared across the JNO and CMO. This strengthens the hypothesis that our method is able to identify representative, as well as

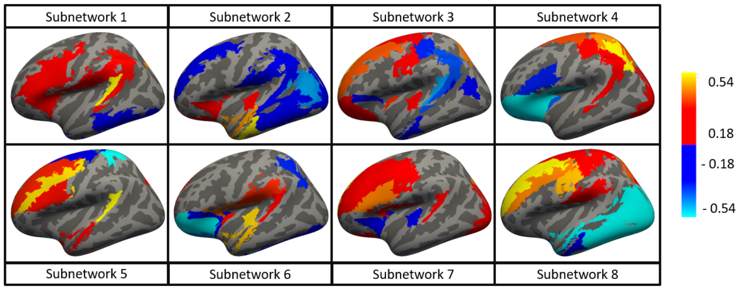


Figure 4.6: Eight subnetworks identified by our model from the prediction of ADOS. The blue & green regions are anticorrelated with the red & orange regions.

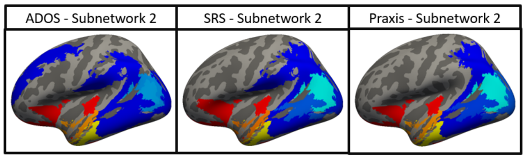


Figure 4.7: Subnetwork 2 obtained from **L:** ADOS **M:** SRS and **R:** Praxis prediction

predictive connectivity patterns.

4.1.6 Discussion

Our Coupled Manifold Optimization strategy jointly analyzes data from two distinct, but related, domains through its shared projection. In contrast to conventional manifold learning, it optimizes for the relevant embedding directions that are predictive of clinical severity. Consequently, the method captures representative connectivity patterns that are important for quantifying and understanding the spectrum of clinical severity among ASD patients. Again, this framework makes very few assumptions about the data and can be adapted to work with different similarity matrices and clinical scores.

4.1.6.1 Limitations and Scope

As alluded to earlier, mapping to multiple clinical measures is a first step to obtaining a more holistic understanding of complex brain disorders. We explored adapting Eq. (4.6) to perform multiscore regression. Accordingly, in Table 4.2, we compare against the single score regression. For the single output regression task, six of the eight hypeparameters were the same for the three scores, namely: the basis sparsity penalty basis γ_1, the tradeoff between the KR and the dictionary learning λ, the ℓ_2 penalty on the coefficients γ_2 and the regression weights γ_3, the dispersion for the exponential kernel σ^2 and the number of networks K. For the *multi-score prediction task* in Table 4.2, we fix these. The other two hyperparameters for the KR, i.e. the tradeoff between kernels ρ and the polynomial kernel order l are score-specific. Since ADOS is the most widely accepted measure of ASD clinical severity, we report the multi-score performance using the ADOS settings. In fact, altering these hyperparameters, at best, enables us to predict one of the three scores well, but always at the expense of the generalization on the other two measures.

This modeling limitation motivates the next section of this chapter where

Score	**Method**	**MAE Test**	**NMI Test**
ADOS	Single Score	2.73 ± 2.63	0.54
	Multi Score	3.17 ± 2.00	0.35
SRS	Single Score	13.28 ± 14.94	0.89
	Multi Score	33.11 ± 28.07	0.51
Praxis	Single Score	9.07 ± 11.91	0.82
	Multi Score	30.11 ± 26.47	0.61

Table 4.2: Comparing the CMO Framework on Single vs Multi-Target Regression

we move away from non-parametric regression in the quest to develop more powerful discriminative frameworks. We will show that leveraging the representational flexibility of neural networks helps us generalize to multiscore prediction.

4.2 Blending Model Based Representations with Neural Networks

This work first appeared in [160]. In this paper, we proposed one of the first end-to-end frameworks that embeds a traditional model-based representation (dictionary learning) with deep networks into a single optimization. This model derives inspiration from [146, 157] to project the patient correlation matrices onto a shared basis. However, in a notable departure from prior work, we couple the patient projection onto the dictionary with a neural network for multi-score behavioral prediction.

Specifically, we *jointly optimize for the basis, patient representation, and neural network weights* by combining gradient information from the two objectives. We demonstrate that our unified framework provides us with the necessary representational flexibility to model complex interactions in the brain, and to learn effectively from limited training data. Our optimization strategy outperforms state-of-the-art baseline methods at estimating a generalizable multi-dimensional patient characterization.

4.2.1 An Integrated Framework for Dictionary Learning and Neural Networks

Fig. 4.8 illustrates our framework. The blue box denotes our dictionary learning representation, while the gray box is the neural network architecture. Again, N is the number of patients and P is the number of regions in our brain parcellation. We decompose the correlation matrix $\mathbf{\Gamma}_n \in \mathcal{R}^{P \times P}$ for each patient n, via K dictionary elements of a shared basis $\mathbf{B} \in \mathcal{R}^{P \times K}$, and a subject-specific loading vector $\mathbf{c}_n \in \mathcal{R}^{K \times 1}$. Thus, our dictionary learning objective $\mathcal{D}$ is as follows:

$$\mathcal{D}(\cdot) = \sum_n \Big[||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 + \gamma_2 ||\mathbf{c}_n||_2^2 \Big] + \gamma_1 ||\mathbf{B}||_1 \tag{4.12}$$

where $\mathbf{diag}(\mathbf{c}_n)$ denotes a matrix with the entries of $\mathbf{c}_n$ on the leading diagonal and the non-diagonal entries as 0. Since $\mathbf{\Gamma}_n$ is positive semi-definite, we add the constraint $\mathbf{c}_{nk} \geq 0$. The columns of $\mathbf{B}$ capture representative patterns of co-activation common to the cohort. The loadings $\mathbf{c}_{nk}$ capture the network strength of basis k in patient n. We add an ℓ_1 penalty to $\mathbf{B}$ to encourage sparsity, and an ℓ_2 penalty to $\{\mathbf{c}_n\}$ to ensure that the objective is well posed. γ_1 and γ_2 are the corresponding regularization weights.

The loadings $\mathbf{c}_n$ are also the input features to a neural network. The network parameters $\mathbf{\Theta}$ encode a series of non-linear transformations that map the input features to behavior. $\mathbf{Y}_n \in \mathcal{R}^{M \times 1}$ is a vector of M concatenated clinical measures, which describe the location of patient n on the behavioral spectrum. $\hat{\mathbf{Y}}_n$ is estimated using the latent representation $\mathbf{c}_n$. We employ the

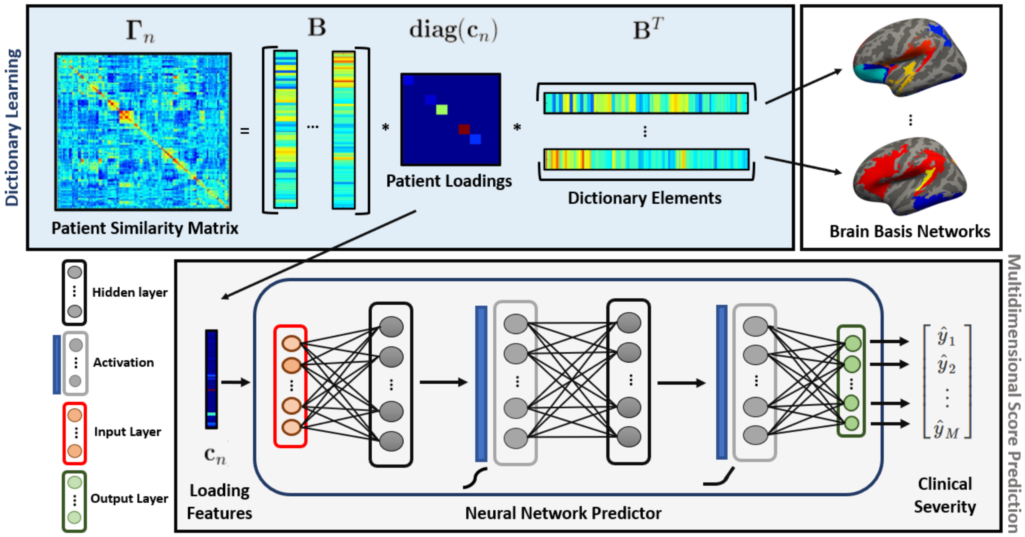


Figure 4.8: A unified framework for integrating neural networks and dictionary learning. **Blue Box:** Dictionary Learning from correlation matrices **Gray Box:** Neural Network architecture for multidimensional score prediction

Mean Square Error (MSE) to define the network loss $\mathcal{L}$:

$$\mathcal{L}(\{\mathbf{c}_n\}, \mathbf{\Theta}; \{\mathbf{Y}_n\}) = \sum_n \ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n) = \lambda \sum_n ||\hat{\mathbf{Y}}_n - \mathbf{Y}_n||_F^2 \tag{4.13}$$

Since $\mathcal{L}(\cdot)$ is added to $\mathcal{D}(\cdot)$ defined in Eq. (4.12), λ balances the contribution of the dictionary learning and neural network terms to the objective.

Our proposed network architecture is highlighted in the gray box. Our modeling choices require us to carefully control for two key network design aspects: representational capacity, and convergence of the optimization. Given the low dimensionality of the input $\mathbf{c}_n$, we opt for a simple fully connected Artificial Neural Network (ANN) with two hidden layers and a width of 40 and ReLU activation. Experimentally, we found that these modeling choices are robust to issues with saturation and vanishing gradients, which commonly confound neural network training.

4.2.2 Joint Inference Strategy

We use alternating minimization to iteratively optimize for the dictionary elements $\mathbf{B}$, the patient projections $\{\mathbf{c}_n\}$, and ANN parameters $\boldsymbol{\Theta}$. Here, we sequentially optimize for each hidden variable in the objective by fixing the rest, until global convergence.

We use Proximal Gradient Descent to handle the non-differentiable ℓ_1 penalty in Eq. (4.12), which requires the rest of the objective to be convex in $\mathbf{B}$. We circumvent this issue by the strategy in [146]. Namely, we introduce N constraints of the form $\mathbf{D}_n = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)$, and substitute them into the Frobenious norm terms in Eq. (4.12). These constraints are enforced using the Augmented Lagrangians $\{\boldsymbol{\Lambda}_n\}$. We add N terms of the form $\text{Tr}\left[\boldsymbol{\Lambda}_n^T(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n))\right] + \frac{1}{2}\,||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)||_F^2$ to Eq. (4.12). We then iterate through the following four steps until convergence.

4.2.2.1 Proximal Gradient Descent on B

Each step of the proximal algorithm constructs a a locally smooth quadratic model of $||\mathbf{B}||_1$ based on the gradient of $\mathcal{D}$ with respect to $\mathbf{B}$. Using this model, the algorithm iteratively updates $\mathbf{B}$ through shrinkage thresholding. We fix the learning rate for this step at 10^{-4}. The updates are similar in form to Eq. (4.5).

4.2.2.2 Updating the Neural Network Weights Θ

We optimize the weights $\boldsymbol{\Theta}$ according to the loss function $\mathcal{L}$ using backpropagation to estimate gradients. There are several obstacles in training a neural network to generalize and few available theoretical guarantees to guide design considerations. We pay careful attention to this, since the global optimization procedure couples the updates between $\mathbf{\Theta}$ and $\{\mathbf{c}_n\}$.

We employ the ADAM optimizer [161], which is robust to small datasets. We randomly initialize at the first main update. We found a learning rate of 10^{-4}, scaled by 0.9 every 5 epochs to be sufficient for encoding the training data, while avoiding bad local minima and over-fitting. We train for 50 epochs with a batch-size of 12. Finally, we fix the obtained weights to update $\{\mathbf{c}_n\}$.

4.2.2.3 L-BFGS update for $\{\mathbf{c}_n\}$

The objective for each $\mathbf{c}_n$ decouples as follows:

$$\mathcal{J}(\mathbf{c}_n) = \ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n) + \gamma_2 ||\mathbf{c}_n||_2^2 + \mathrm{Tr}\left[\mathbf{\Lambda}_n^T(\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n))\right] \\ + \frac{1}{2}||\mathbf{D}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)||_F^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0 \quad (4.14)$$

Notice that we can use a standard backpropagation algorithm to compute the gradient of $\ell_{\mathbf{\Theta}}(.)$ with respect to $\mathbf{c}_n$, denoted by $\nabla \ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n)$. The gradient of $\mathcal{J}$ with respect to $\mathbf{c}_n$, denoted $\mathbf{g}(\mathbf{c}_n)$, can then be computed as follows:

$$\mathbf{g}(\mathbf{c}_n) = \nabla \ell_{\mathbf{\Theta}}(\mathbf{c}_n, \mathbf{Y}_n) + \mathbf{c}_n \circ \left[\left[\mathcal{I}_K \circ (\mathbf{B}^T\mathbf{B})\right]\mathbf{1}\right] - \left[\mathcal{I}_K \circ (\mathbf{\Lambda}_n^T\mathbf{B} + \mathbf{D}_n^T\mathbf{B})\right]\mathbf{1} + 2\gamma_2\mathbf{c}_n$$

where $\mathbf{1}$ is the vector of all ones. The first term is from the ANN, while the rest are from the modified dictionary learning objective. The gradient combines information from the ANN function landscape with that from the correlation matrix estimation. For each iteration r, the BFGS [159] algorithm recursively constructs a positive-definite Hessian approximation $\mathbf{H}(\mathbf{c}_n^r)$ based on the gradients estimated. Next, we iteratively compute a descent direction $\mathbf{p}$ for $\mathbf{c}_n^r$ using the following bound-constrained objective:

$$\mathbf{p}^* = \text{argmin}_{\mathbf{p}} \mathcal{J}(\mathbf{c}_n^r) + \mathbf{g}(\mathbf{c}_n^r)^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T \mathbf{H}(\mathbf{c}_n^r)\mathbf{p} \;\; s.t. \;\; \mathbf{c}_{nk}^r + \mathbf{p}_k \geq 0 \tag{4.15}$$

We then update $\mathbf{c}_n$ as: $\mathbf{c}_n^{r+1} = \mathbf{c}_n^r + \delta \mathbf{p}^*$, repeating this procedure until convergence. Effectively, the BFGS update leverages second-order curvature information through each $\mathbf{H}(\mathbf{c}_n)$ estimation. In practice, δ is set to 0.9.

4.2.2.4 Augmented Lagrangian Update for the Constraint Variables.

We have a closed form solution for computing the constraint argument $\{\mathbf{D}_n\}$. The dual Lagrangians, i.e. $\{\mathbf{\Lambda}_n\}$ are updated via gradient ascent. We cycle through the collective updates for these two variables until convergence. We use a learning rate of 10^{-4}, scaled by 0.75 at each iteration of gradient ascent. The variable updates are similar in form to Eq. (4.9-4.10).

4.2.2.5 Prediction on Unseen Data

We use cross validation to assess our framework. For a new patient, we compute the loading vector $\bar{\mathbf{c}}$ using the estimates $\{\mathbf{B}^*, \mathbf{\Theta}^*\}$ obtained during training. We remove the contribution of the ANN term from the joint objective, as we do not know the corresponding value of $\bar{\mathbf{Y}}$ for a new patient. The proximal operator conditions are assumed to hold with equality, removing the Lagrangian terms. The optimization in $\bar{\mathbf{c}}$ takes the following form:

$$\frac{1}{2}\bar{\mathbf{c}}^T\bar{\mathbf{H}}\bar{\mathbf{c}} + \bar{\mathbf{f}}^T\bar{\mathbf{c}} \;\; s.t. \;\; \bar{\mathbf{A}}\bar{\mathbf{c}} \leq \bar{\mathbf{b}} \tag{4.16}$$

$$\bar{\mathbf{H}} = 2(\mathbf{B}^T\mathbf{B}) \circ (\mathbf{B}^T\mathbf{B}) + 2\gamma_2\mathcal{I}_K$$

$$\bar{\mathbf{f}} = -2\mathcal{I}_K \circ (\mathbf{B}^T\mathbf{\Gamma}_n\mathbf{B})\mathbf{1}; \;\; \bar{\mathbf{A}} = -\mathcal{I}_K \;\; \bar{\mathbf{b}} = \mathbf{0}$$

This formulation is similar to Eq. (4.15) from the BFGS update for $\{\mathbf{c}_n\}$. $\bar{\mathbf{H}}$ is also positive definite, thus giving an efficient quadratic programming solution to Eq. (4.16). We estimate the score vector $\bar{\mathbf{Y}}$ by a forward pass.

4.2.3 Model Evaluation

4.2.3.1 Baseline Models

We compare against two baselines that predict severity scores from correlation matrices $\mathbf{\Gamma}_n \in \mathcal{R}^{P\times P}$. The first has a joint optimization flavor similar to this work, while the second uses a CNN to exploit the structure in $\{\mathbf{\Gamma}_n\}$:

1. The Joint Network Optimization Framework in [146, 39]

2. BrainNet Convolutional Neural Network (CNN) from [44]

4.2.3.2 Implementation Details

The model in [146, 39] adds a linear predictive term $\gamma||\mathbf{C}^T\mathbf{w} - \mathbf{y}||_2^2 + \lambda_3||\mathbf{w}||_2^2$ to the dictionary learning objective in Eq.(4.12). This estimates a single regression vector $\mathbf{w}$ to compute a scalar measure $\mathbf{y}_n$ from the loading matrix $\mathbf{C} \in \mathcal{R}^{K\times N}$. To provide a fair comparison, we modify this discriminative term to $\gamma||\mathbf{C}^T\mathbf{W} - \mathbf{Y}||_2^2 + \lambda_3||\mathbf{W}||_2^2$, to predict the vectors $\{\mathbf{Y}_n \in \mathcal{R}^{M\times 1}\}_{n=1}^N$ using the weight matrix $\mathbf{W} \in \mathcal{R}^{K\times M}$. According to [146, 39], we fixed λ_3 and γ at 1, and swept the other parameters over a suitable range. We set number of networks to $K = 8$, which is the knee point of the eigenspectrum for $\{\mathbf{\Gamma}_n\}$.

The network architecture in [44] predicts two cognitive measures from correlation matrices. In our case, $\{\mathbf{\Gamma}_n\}$ are of size $P \times P$. For our comparison, we modify the output layer to be of size M. We use the recommended guidelines in [44] for setting the learning rate, batch-size and momentum during training.

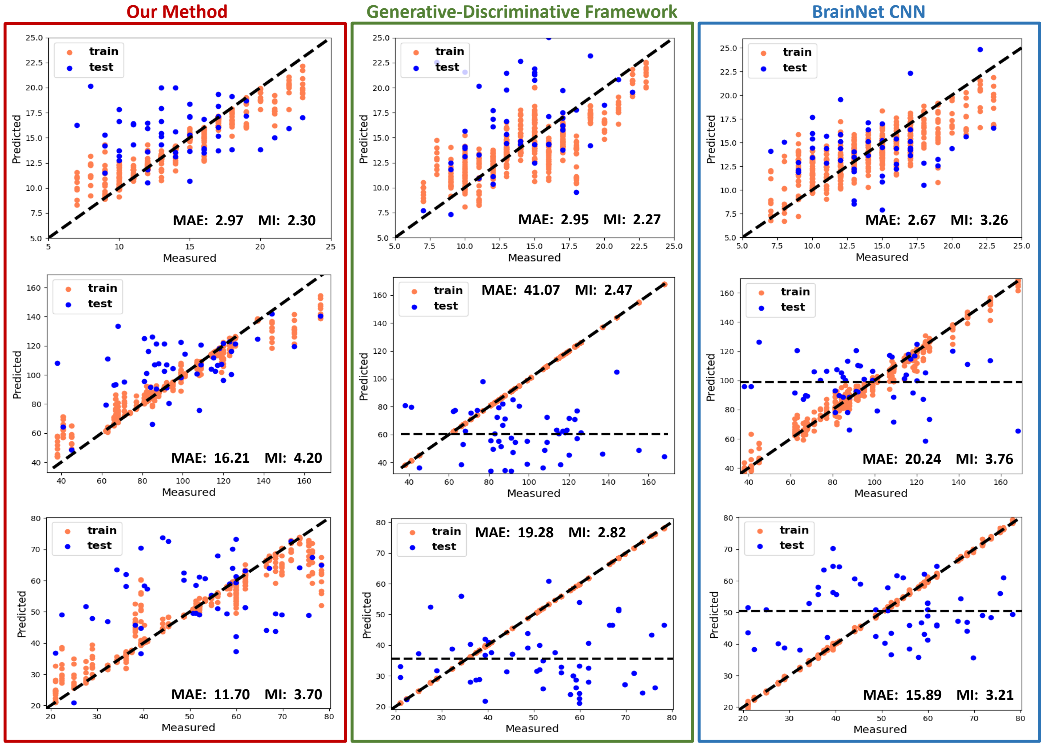


Figure 4.9: Multi-Score Prediction performance for **Top:** ADOS **Middle:** SRS **Bottom:** Praxis by **Red Box:** Our Framework. **Green Box:** Generative-Discriminative Framework from [146]. **Blue Box:** BrainNet CNN from [44]

4.2.4 Experiments on Autism Dataset

Fig. 4.9 illustrates the *multi-score regression* performance of each method based on ten fold cross validation. Our quantitative metrics are median absolute error (MAE) and mutual information (MI) between the actual and computed scores. Lower MAE and higher MI indicate better performance. The orange points indicate training fit, while the blue points denote performance on held out samples. The $\mathbf{x} = \mathbf{y}$ line indicates ideal performance. We restrict our comparison to patients with all three scores, i.e. ADOS, SRS, and Praxis.

Observe that both the Generative-Discriminative model (JNO) and the BrainNet CNN perform comparably to our model for predicting ADOS. However, our model outperforms the baselines in terms of MAE and MI for SRS and Praxis, with the blue points following the $\mathbf{x} = \mathbf{y}$ line more closely. Generally, we find that as we vary the free parameters, the baselines predict *one of the three scores well (in Fig. 4.9, ADOS), but fit the rest poorly*. In contrast, only our framework learns a representation that predicts all three clinical measures *simultaneously*, and hence overall outperforms the baselines. We believe that the representational flexibility of neural networks along with our joint optimization is key to generalization.

4.2.4.1 Subnetwork Identification

Fig. 4.10 illustrates the subnetworks in $\{\mathbf{B}_k\}$. Regions storing positive values are anticorrelated with negative regions. From a clinical standpoint, Subnetwork 8 includes the somatomotor network (SMN) and competing, i.e.

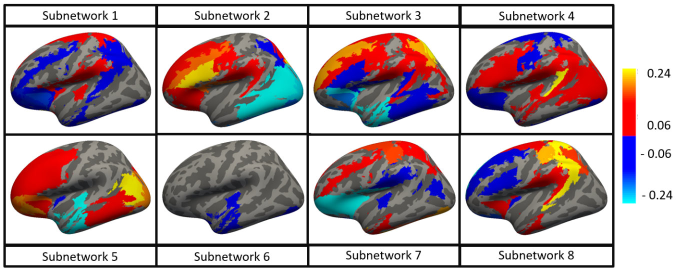


Figure 4.10: Eight subnetworks identified by our model from multi-score prediction. The blue and green regions are anticorrelated with the red and orange regions.

anticorrelated, contributions from the default mode network (DMN). Subnetwork 3 also has contributions from the DMN and SMN, both of which have been widely reported in ASD [98]. Along with the DMN, Subnetworks 5 and 2 contain positive and competing contributions from the higher order visual processing areas (i.e. occipital and temporal lobes) respectively. These findings concur with behavioral reports of reduced visual-motor integration in ASD [98]. Finally, Subnetworks 2, 3, and 8 exhibit central executive control network and insula contributions, believed to be critical for switching between self-referential and goal-directed behavior [152].

4.2.5 Discussion

This work introduces one of the first unified framework to combine classical optimization with the modern-day representational power of neural networks. This integrated strategy allows us to characterize and predict multidimensional behavioral severity from rs-fMRI connectomics data. Namely, our dictionary learning term provides us with interpretability in the brain basis for clinical impairments. The predictive term cleverly exploits the ability of neural networks to learn rich representations from data. The joint optimization helps learn informative connectivity patterns from limited training data.

This framework makes very few assumptions about the data and can be adapted to work with complex clinical score prediction scenarios. An example is an extension of this model to handle case/control severity prediction using a mixture density network (MDN) [162] in lieu of a regression network. The MDN models a mixture of Gaussians to fit a target multi-modal distibution. Accordingly, the network loss function is a negative log-likelihood, which differs from conventional formulations. This is another scenario that may advance our understanding of neuropsychiatric disorders. For example, this can be used for case/control ASD severity prediction or to underscore differences among sub-types within ASD or ADHD.

So far, we have looked at the brain through the lens of static functional connectivity profiles. The next chapter provides a more holistic exploration where we extend these ideas to incorporate dynamic and multimodal (i.e. structural) connectivity.

Chapter 5

Deep sr-DDL - A Deep Generative Hybrid Model for Multidimensional Phenotypic Prediction from Multimodal and Dynamic Connectomics Data

As mentioned in Chapter 2, techniques integrating structural and functional connectivity focus heavily on groupwise discrimination from the static connectomes. Methods include statistical tests on the node or edge biomarkers [10], data-driven representations [128], and neural networks [129] for classification. However, none of these methods tackle continuous-valued prediction, e.g., quantifying level of deficit.

On the other hand, deep learning is becoming increasingly popular for continuous prediction. The work of [44] proposes a specialized end-to-end convolutional network that predicts clinical outcomes from DTI connectomes. In [160] we combined a dictionary learning on the rs-fMRI correlations with a neural network to predict clinical severity in ASD patients. However, these methods focus on a single neuroimaging modality and do not leverage complementary information between structure and function.

There is also mounting evidence that functional connectivity between regions is a dynamically evolving process [100]. Modeling this evolution is believed to be important for understanding disorders like ASD [102, 101]. In this regard, recent methods have been proposed that use either a sparse decomposition of the rs-fMRI connectomes [103], or a temporal clustering for ASD/control discrimination [163]. While promising, these approaches focus exclusively on rs-fMRI and ignore structural information.

In this chapter, we describe a deep-generative model that integrates structural and dynamic functional connectivity with behavior into a unified optimization framework. Our generative component is a structurally-regularized Dynamic Dictionary Learning (sr-DDL) model, which uses anatomical priors from DTI to regularize a time-varying decomposition of the rs-fMRI correlation matrices. Here, the connectivity profiles are explained by shared basis networks and time-varying patient-specific loadings. Simultaneously, these loadings are input to a deep network which uses an LSTM (Long Short Term Memory Network) to model temporal trends and an ANN (Artificial Neural Network) to predict clinical severity. Our optimization procedure learns the bases, loadings, and neural network weights most predictive of behavioral deficits in ASD. We obtain a representation which is both interpretable and generalizes to unseen patients, thus providing a comprehensive characterization of the disorder.

Outline: This chapter is based on work that appeared in [164, 165]. Section 5.1

introduces the extension of our generative framework to dynamic connectivity, while Section 5.2 explains our structural regularization as guided by anatomical priors and Section 5.3 describes our deep network for multioutput prediction. Section 5.5 describes our joint optimization strategy, while Sections 5.6-5.8 describes our evaluation strategy. Finally, Section 5.9 discusses the clinical significance of the results.

Fig. 5.1 presents a graphical overview of our framework. We have two sets of inputs to the model for each individual namely, the dynamic individual-specific correlation matrices, and the DTI structural connectome graph (upper left). Our outputs are the scalar clinical scores (bottom right). We use the sliding window approach in Fig. 2.3 to extract dynamic rs-fMRI correlation matrices and tractography to extract the DTI connectomes as shown in Fig. 2.4. The DTI input to our model is the Graph Laplacian obtained from a binary DTI adjacency matrix capturing the presence/absence of a fiber between regions. Finally, the behavioral scores for each individual are obtained from an expert assessment. This score can correspond to either cognitive outcomes or severity of symptoms in case of neurodevelopmental diseases.

The green box in Fig. 5.1 describes the generative component of our framework. Here, the dynamic rs-fMRI correlation matrices are decomposed using a structurally regularized dynamic dictionary learning (sr-DDL). The columns in the bases subnetworks capture representative patterns common to the cohort. The loading coefficients differ across subjects, and evolve over time. At each timepoint, they determine the contribution of each basis to the dynamic functional connectivity profile of the individual. Finally, the DTI Graph

Laplacians re-weight the decomposition to focus on the functional connectivity between anatomically linked regions. The gray box denotes the deep networks part of our model. This network combines a Long Short Term Memory (LSTM) module with an Artificial Neural Network (ANN) to predict multiple behavioral scores. The LSTM models the temporal trends in the subject-specific loading coefficients giving rise to a hidden representation. The ANN then uses this representation to predict the corresponding behavioral outcomes.

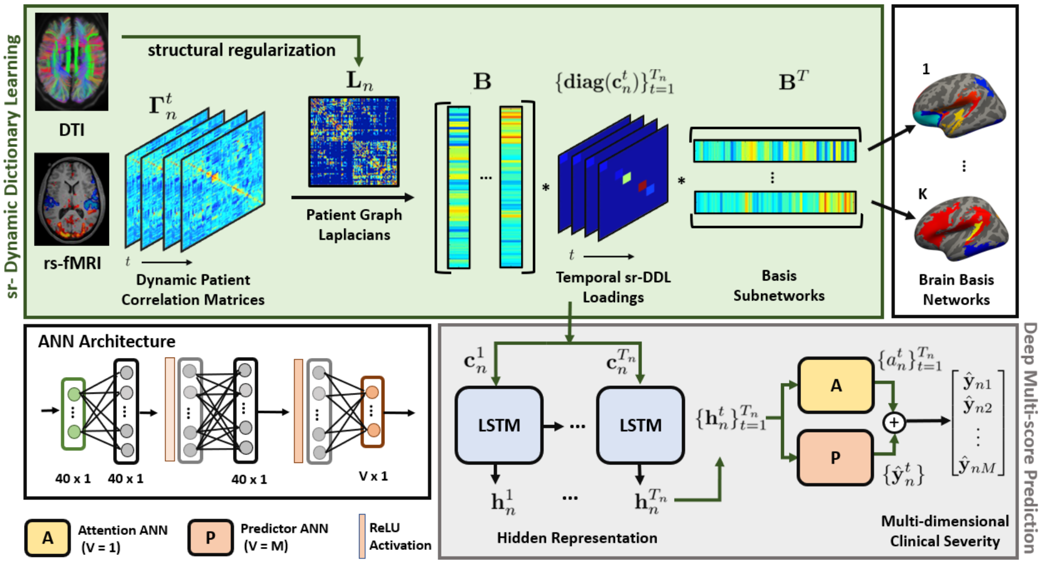


Figure 5.1: Framework to integrate structural and dynamic functional connectivity for clinical severity prediction **Green Box:** The generative sr-DDL module. The rs-fMRI dynamic correlation matrices are decomposed into the subnetwork basis and time-varying subject-specific loadings. The DTI connectivity regularizes this decomposition. **Gray Box:** Deep LSTM-ANN module for multi-score prediction. The sr-DDL coefficients are input into the LSTM to generate a hidden representation. The predictor ANN (P-ANN) generates a time varying estimate for the scores, while the attention ANN (A-ANN) weights the predictions across time to generate the final clinical severity estimate.

5.1 Dynamic Dictionary Learning for Time-Varying Functional Connectivity

We denote the set of time varying functional correlation matrices for individual n by the set $\{\mathbf{\Gamma}_n^t\}_{t=1}^{T_n} \in \mathcal{R}^{P \times P}$. Here, T_n denotes the number of sliding windows applied to the rs-fMRI scan, and P is the number of ROIs in the parcellation scheme. As seen in Fig. 5.1 (green box), we model this information using a group average basis, and subject-specific temporal loadings. The dictionary $\mathbf{B} \in \mathcal{R}^{P \times K}$ is a concatenation of K elemental bases vectors $\mathbf{b}_k \in \mathcal{R}^{P \times 1}$, i.e. $\mathbf{B} := [\mathbf{b}_1 \quad \mathbf{b}_2 \quad ... \quad \mathbf{b}_K]$, where $K \ll P$. This basis captures representative brain states which each subject cycles through over the course of the scan. We further constrain the basis vectors to be orthogonal to each other. This constraint acts as an implicit regularizer, ensuring that the learned subnetworks are uncorrelated, yet explain the rs-fMRI data well.

While the bases are shared across the cohort, the strength of their combination differs across individuals and varies over time. These loadings are denoted by the set $\{\mathbf{c}_n^t\}_{t=1}^{T_n}$ and combine the basis subnetworks uniquely to best explain each subject's functional connectivity. We introduce an explicit non-negativity constraint $\mathbf{c}_{nk}^t$ to ensure that the positive semi-definiteness of $\mathbf{\Gamma}_n^t$ is preserved. The complete rs-fMRI data representation takes the form:

$$\mathbf{\Gamma}_n^t \approx \sum_k \mathbf{c}_{nk}^t \mathbf{b}_k \mathbf{b}_k^T \quad s.t. \quad \mathbf{c}_{nk} \geq 0, \;\; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K, \tag{5.1}$$

where $\mathcal{I}_K$ is the $K \times K$ identity matrix. As seen in Eq. (5.1), the subject-specific loading vector at time t, $\mathbf{c}_n^t := [\mathbf{c}_{n1}^t \quad ... \quad \mathbf{c}_{nK}^t]^T \in \mathcal{R}^{K \times 1}$ models the heterogeneity in the cohort. Denoting $\mathbf{diag}(\mathbf{c}_n^t)$ as a diagonal matrix with the K

subject-specific coefficients on the diagonal and off-diagonal terms set to zero, Eq. (5.1) can be re-written in the following matrix form:

$$\mathbf{\Gamma}_n^t \approx \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T \quad s.t. \quad \mathbf{c}_{nk}^t \geq 0, \;\; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K \tag{5.2}$$

This matrix factorization serves to reduce the dimensionality of the data, while simultaneously modeling group-level and subject-specific information.

5.2 Structural DTI Regularization Using Anatomical Priors

Let $\mathbf{A}_n \in \mathcal{R}^{P \times P}$ be a binary adjacency matrix derived from the structural connectome of subject n. For example, $\mathbf{A}_n$ can be constructed by thresholding the number of fibers estimated between two regions via tractography. Let $\mathcal{E}$ denote the set of edges in this graph. We compute the corresponding Normalized Graph Laplacian [166] as $\mathbf{L}_n = \mathbf{V}_n^{-\frac{1}{2}}(\mathbf{V}_n - \mathbf{A}_n)\mathbf{V}_n^{-\frac{1}{2}}$, where $\mathbf{V}_n = \mathbf{diag}(\mathbf{A}_n\mathbf{1})$ is the degree matrix and $\mathbf{1}$ is the vector of all ones. Intuitively, the Graph Laplacian is a discrete analog of the Laplace difference operator in Euclidean space. The Laplace difference operator has been used to characterize local properties of functions in Euclidean space (for example, to easily identify and characterize local optima). The Graph Laplacian generalizes this notion to discrete graphs and functions that are defined on graphs. Specifically, the Graph Laplacian has become a popular spatial regularizer in computer vision [167], genetics [168], and neuroimaging [14, 169]. This regularization implicitly assumes that there is a data signal associated with each node of the graph, and it encourages these signals to be similar for nodes of the graph that have an edge between them.

We use a matrix analog to Graph Laplacian regularization via the weighted Frobenius norm i.e. $||.||_{\mathbf{L}_n}$ [170, 171], which we use in place of the isotropic ℓ_2 penalty in Eq. (5.2). In this case, the graph “signal" corresponds to the vector (i.e., profile) of approximation errors given in Eq. (5.2) between the node in question and all other nodes in the graph. The underlying anatomical connectivity graph is defined by the DTI Graph Laplacian $\mathbf{L}_n$ for each patient. Mathematically, our dictionary learning loss takes the following form:

$$||\mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T||_{\mathbf{L}_n}$$

$$= \mathrm{Tr}\left[(\mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T)\mathbf{L}_n(\mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T)\right] \quad (5.3)$$

Let $\mathbf{E}_n^t = \mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T$ denote the element-wise approximation error of the the correlation matrix $\mathbf{\Gamma}_n^t$. Similarly, we define $\tilde{\mathbf{E}}_n^t = \mathbf{V}_n^{-\frac{1}{2}}\mathbf{E}_n^t$ as a weighted version of this error based on the degree matrix. For notational convenience, we will drop the subscripts n and t from the following computation.

$$||\mathbf{E}||_{\mathbf{L}} = \mathrm{Tr}[\mathbf{E}^T\mathbf{L}\mathbf{E}] = \mathrm{Tr}[\mathbf{E}^T\mathbf{V}^{-\frac{1}{2}}(\mathbf{V}-\mathbf{A})\mathbf{V}^{-\frac{1}{2}}\mathbf{E}]$$

$$= \mathrm{Tr}[\tilde{\mathbf{E}}^T(\mathbf{V}-\mathbf{A})\tilde{\mathbf{E}}] \ \text{ where } \ \tilde{\mathbf{E}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{E}$$

$$= \sum_i\sum_j\sum_k \tilde{\mathbf{E}}(i,j)[\mathbf{V}(i,k) - \mathbf{A}(i,k)]\tilde{\mathbf{E}}(k,j)$$

$$= \sum_{i,j,k}\mathbf{V}(i,k)\tilde{\mathbf{E}}(i,j)\tilde{\mathbf{E}}(k,j) - \sum_{i,j,k}\mathbf{A}(i,k)\tilde{\mathbf{E}}(i,j)\tilde{\mathbf{E}}(k,j)$$

$$= \sum_{i,j} \mathbf{V}(i,i)\tilde{\mathbf{E}}(i,j)\tilde{\mathbf{E}}(i,j) - \sum_{i,j,k} \mathbf{A}(i,k)\tilde{\mathbf{E}}(i,j)\tilde{\mathbf{E}}(k,j)$$

$$= \sum_{j} \sum_{(i,k)\in\mathcal{E}} 2[\tilde{\mathbf{E}}(i,k)]^2 - \sum_{j} \sum_{(i,k)\in\mathcal{E}} 2[\tilde{\mathbf{E}}(i,j)\tilde{\mathbf{E}}(k,j)]$$

$$= \sum_{j} \Big[\sum_{(i,k)\in\mathcal{E}} [\tilde{\mathbf{E}}(i,k)]^2 + \sum_{(i,k)\in\mathcal{E}} [\tilde{\mathbf{E}}(k,j)]^2 \Big] - \sum_{j} \sum_{(i,k)\in\mathcal{E}} 2[\tilde{\mathbf{E}}(i,j)\tilde{\mathbf{E}}(k,j)]$$

$$= \sum_{j} \sum_{(i,k)\in\mathcal{E}} \Big[\tilde{\mathbf{E}}(i,j) - \tilde{\mathbf{E}}(k,j) \Big]^2$$

$$= \sum_{(i,k)\in\mathcal{E}} ||\tilde{\mathbf{E}}(i,:) - \tilde{\mathbf{E}}(k,:)||_2^2$$

$$= \sum_{(i,k)\in\mathcal{E}} ||[\mathbf{V}(i,i)]^{-\frac{1}{2}}\mathbf{E}(i,:) - [\mathbf{V}(k,k)]^{-\frac{1}{2}}\mathbf{E}(k,:)||_2^2$$

Writing out the appropriate subscripts and superscripts we dropped earlier, we obtain the expression:

$$||\mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T||_{\mathbf{L}_n} = \sum_{(i,k)\in\mathcal{E}} ||\tilde{\mathbf{E}}_n^t(i,:) - \tilde{\mathbf{E}}_n^t(k,:)||_2^2 \tag{5.4}$$

$$= \sum_{(i,k)\in\mathcal{E}} ||[\mathbf{V}_n(i,i)]^{-\frac{1}{2}}\mathbf{E}_n^t(i,:) - [\mathbf{V}_n(k,k)]^{-\frac{1}{2}}\mathbf{E}(k,:)||_2^2 \tag{5.5}$$

Notice that for terms where $(i,k) \notin \mathcal{E}$, i.e. there is no anatomical connection between nodes i and k, the corresponding error term in the summation drops out. Said another way, this construction minimizes the sum of the square difference between the rs-fMRI reconstruction profiles ($\tilde{\mathbf{E}}_n^t(i,:)$ and $\tilde{\mathbf{E}}_n^t(k,:)$) between nodes ($i$ and k) that are adjacent via the DTI graph. This effectively reweights the rs-fMRI reconstruction profiles of anatomically connected nodes according to their relative degrees ($\mathbf{V}_n(i,i)$ and $\mathbf{V}_n(k,k)$) in the DTI graph pairwise. Thus, the functional connectivity at a particular node is directly influenced by its anatomical connections with other nodes in the graph. At a high level, this construction implicitly regularizes the rs-fMRI reconstruction loss according to the underlying anatomical connectivity prior.

Finally, based on the formulation in Eq. (5.3), the final sr-DDL objective $\mathcal{D}(.)$ can be expressed as follows:

$$\mathcal{D}(\cdot) = \sum_{t} \frac{1}{T_n} ||\mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T||_{\mathbf{L}_n} \;\; s.t. \;\; \mathbf{c}_{nk}^t \geq 0, \;\; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K \tag{5.6}$$

5.3 Deep Network for Multidimensional Prediction

As seen in the gray box in Fig. 5.1, the subject-specific coefficients $\{\mathbf{c}_n^t\}$ are input to an LSTM-ANN to predict the clinical scores, as parametrized by the weights $\mathbf{\Theta}$. The M clinical scores for each individual are concatenated into a vector $\mathbf{y}_n := [\mathbf{y}_{n1} \quad ... \quad \mathbf{y}_{nM}]^T \in \mathcal{R}^{M \times 1}$. The LSTM models the temporal variations in the coefficients $\{\mathbf{c}_n^t\}$ to generate a hidden representation $\{\mathbf{h}_n^t\}_{t=1}^{T_n}$. From here, the Predictor ANN (P-ANN) generates a time varying estimates of the scores $\{\hat{\mathbf{y}}_n^t\}_{t=1}^{T_n} \in \mathcal{R}^{M \times 1}$. At the same time, the Attention ANN (A-ANN) generates T_n scalars from the hidden representation. These are then softmax across time to obtain the attention weights: $\{a_n^t\}_{t=1}^{T_n}$. The final prediction is an attention-weighted average across the time estimates, which takes the following form:

$$\hat{\mathbf{y}}_n = \sum_{t} \hat{\mathbf{y}}_n^t a_n^t \tag{5.7}$$

Effectively, the attention weights determine which time points for each subject are most relevant for behavioral prediction. Additionally, they allow us to handle rs-fMRI scans of varying durations. Mathematically, we compute the multi-score prediction error $\mathcal{L}(.)$ using the Mean Squared Error (MSE) loss function as follows:

$$\mathcal{L}(\{\mathbf{c}_n^t\}, \mathbf{y}_n; \mathbf{\Theta}) = ||\hat{\mathbf{y}}_n - \mathbf{y}_n||_F^2 = \left|\left|\sum_{t=1}^{T_n} \hat{\mathbf{y}}_n^t a_n^t - \mathbf{y}_n\right|\right|_F^2 \tag{5.8}$$

At a high level, the deep network distills the temporal information to best predict each subject's clinical profile.

We would like to highlight that our choice of the LSTM over a Recurrent Neural Network (RNN) allows us to track the temporal evolution of connectivity over longer horizons, while avoiding issues with convergence [172]. Our two branched ANN in conjunction with the LSTM directly pools together time-varying estimates of clinical severity by focusing on the portions of the rs-fMRI scan most relevant to prediction. We notice that this construction naturally allows us to handle scans of varying length, while at same time obviating the need for additional sequence padding as would be required by a competing $1D$ CNN.

In Section 5.5, we will develop a coupled optimization procedure to jointly estimate our unknowns $\{\mathbf{B}, \{\mathbf{c}_n^t\}, \mathbf{\Theta}\}$. We will show that our estimation procedure for the coefficients and neural network weights only relies on backpropagated gradients from the neural network loss and the parametric gradients from the dictionary learning. From the joint objective in Eq. (5.9), we can see that the choice of neural network architecture does not directly affect the dictionary learning gradients. So long as we can backpropagate the deep network loss to the coefficients $\mathbf{c}_n^t$, we can effectively adopt our optimization

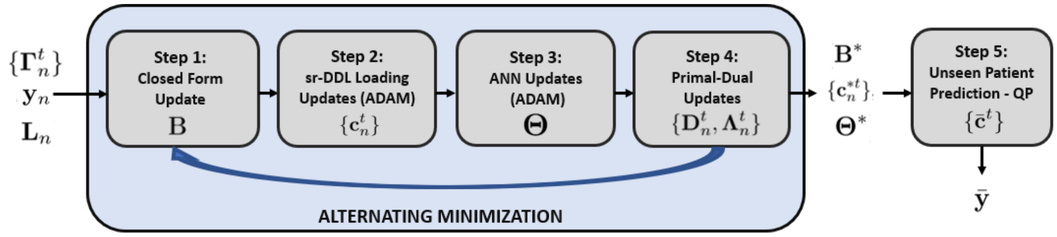


Figure 5.2: Alternating minimization strategy for joint optimization of Eq. (5.10)

strategy to handle an alternative architecture. Said another way, our coupled optimization procedure is agnostic to the specific neural network choice.

5.4 Joint Objective

We combine the complementary viewpoints in Eq. (5.6) and Eq. (5.8) into a single joint objective below:

$$\mathcal{J}(\mathbf{B}, \{\mathbf{c}_n^t\}, \boldsymbol{\Theta}; \{\boldsymbol{\Gamma}_n^t\}, \mathbf{L}_n, \{\mathbf{y}_n\}) = \underbrace{\sum_n \mathcal{D}(\mathbf{B}, \{\mathbf{c}_n^t\}; \{\boldsymbol{\Gamma}_n^t\}, \mathbf{L}_n)}_{\textbf{sr-DDL loss}} + \lambda \underbrace{\sum_n \mathcal{L}(\boldsymbol{\Theta}, \{\mathbf{c}_n^t\}; \mathbf{y}_n)}_{\textbf{deep network loss}}$$

$$= \sum_n \sum_t \frac{1}{T_n} ||\boldsymbol{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T||_{\mathbf{L}_n} + \lambda \sum_n \mathcal{L}(\boldsymbol{\Theta}, \{\mathbf{c}_n^t\}; \mathbf{y}_n) \;\; s.t. \;\; \mathbf{c}_{nk}^t \geq 0, \; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K \tag{5.9}$$

Here, λ is a hyperparameter than balances the tradeoff between the representation loss $\mathcal{D}(.)$ and the prediction loss $\mathcal{L}(.)$.

5.5 Joint Inference Strategy

We employ the alternating minimization technique in order to infer the set of hidden variables $\{\mathbf{B}, \{\mathbf{c}_n^t\}, \boldsymbol{\Theta}\}$. Namely, we optimize Eq. (5.9) for each output variable, while holding the other unknowns constant.

We utilize the fact that there is a closed-form Procrustes solution for quadratic objectives of the form $||\mathbf{M} - \mathbf{B}||_F^2$ [173]. However, Eq. (5.9) is bi-quadratic in $\mathbf{B}$, so it cannot be directly applied. Therefore, we adopt the strategy in [157, 160, 39] of introducing $\sum_n T_n$ constraints of the form $\mathbf{D}_n^t = \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)$. These constraints are enforced via the Augmented Lagrangian algorithm with corresponding constraint variables $\{\mathbf{\Lambda}_n^t\}$. Thus, our objective from Eq. (5.9) now becomes:

$$\mathcal{J}_c = \sum_{n,t} \frac{1}{T_n} ||\mathbf{\Gamma}_n^t - \mathbf{D}_n^t \mathbf{B}^T||_{\mathbf{L}_n} + \lambda \sum_n \mathcal{L}(\mathbf{\Theta}, \{\mathbf{c}_n^t\}; \mathbf{y}_n)$$

$$+ \sum_{n,t} \frac{\gamma}{T_n} \Big[\text{Tr} \left[(\mathbf{\Lambda}_n^t)^T (\mathbf{D}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)) \right] \Big] + \sum_{n.t} \frac{\gamma}{T_n} \Big[\frac{1}{2} \, ||\mathbf{D}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)||_F^2 \Big]$$

$$s.t. \;\; \mathbf{c}_{nk}^t \geq 0, \mathbf{B}^T\mathbf{B} = \mathcal{I}_K \quad (5.10)$$

The Frobenius norm terms $||\mathbf{D}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)||_F^2$ regularize the trace constraints during the optimization. Observe that Eq. (5.10) is convex in the set $\{\mathbf{D}_n^t\}$, which allows us to optimize this variable via standard procedures. The constraint parameter is fixed at $\gamma = 20$, based on the guidelines in [174].

Fig. 5.2 depicts our alternating minimization strategy. We describe each individual block in detail below:

5.5.1 Step 1: Closed form solution for B

Notice that Eq. (5.10) reduces to the following quadratic form in **B**:

$$\mathbf{B}^* = \text{argmin}_{\mathbf{B}:\ \mathbf{B}^T\mathbf{B}=\mathcal{I}_K} ||\mathbf{M} - \mathbf{B}||_F^2 \tag{5.11}$$

Given the singular value decomposition $\mathbf{M} = \mathbf{USV}^T$, we have the following closed form solution :

$$\mathbf{B}^* = \mathbf{UV}^T$$

where **M** is computed as follows:

$$\mathbf{M} = \sum_n \frac{1}{T_n} \sum_t (\mathbf{\Gamma}_n^t \mathbf{L}_n + \mathbf{L}_n \mathbf{\Gamma}_n^t) \mathbf{D}_n^t + \sum_n \frac{1}{T_n} \Big[\sum_t \frac{\gamma}{2} \mathbf{D}_n^t \mathbf{diag}(\mathbf{c}_n^t) + \gamma \mathbf{\Lambda}_n^t \mathbf{diag}(\mathbf{c}_n^t) \Big]$$

Essentially, **B** spans the anatomically weighted space of subject-specific dynamic correlation matrices.

5.5.2 Step 2: Updating the sr-DDL loadings $\{\mathbf{c}_n^t\}$

The objective $\mathcal{J}_c(\cdot)$ in Eq. (5.10) decouples across subjects. We can also incorporate the non-negativity constraint $\mathbf{c}_{nk}^t \geq 0$ by passing an intermediate vector $\hat{\mathbf{c}}_n^t$ through a ReLU. Thus:

$$\mathbf{c}_n^t = ReLU(\hat{\mathbf{c}}_n^t) \tag{5.12}$$

The ReLU pre-filtering allows us to optimize an unconstrained version of Eq. (5.10), as follows:

$$\mathcal{J}(\cdot)_{\hat{c}} = \lambda \sum_{n} \mathcal{L}(\mathbf{\Theta}, \{\mathbf{c}_n^t\}; \mathbf{y}_n) + \sum_{n,t} \frac{\gamma}{T_n} \Big[\text{Tr} \left[(\mathbf{\Lambda}_n^t)^T (\mathbf{D}_n^t - \mathbf{Bdiag}(\mathbf{c}_n^t)) \right] \Big]$$

$$+ \sum_{n.t} \frac{\gamma}{T_n} \left[\frac{1}{2} \, ||\mathbf{D}_n^t - \mathbf{Bdiag}(\mathbf{c}_n^t)||_F^2 \right] \quad (5.13)$$

This optimization can be performed via the stochastic ADAM algorithm [175] by backpropagating the gradients from the loss in Eq. (5.13) upto the input $\{\hat{\mathbf{c}}^t\}$. Experimentally, we set the initial learning rate to be 0.02, scaled by 0.9 per 10 iterations. Essentially, this optimization couples the parametric gradient from the Augmented Lagrangian formulation with the backpropagated gradient from the deep network (parametrized by fixed $\mathbf{\Theta}$). After convergence, the thresholded loadings $\mathbf{c}_n^t = ReLU(\hat{\mathbf{c}}_n^t)$ are used in the subsequent steps of the minimization.

5.5.3 Step 3: Updating the Deep Network weights-$\mathbf{\Theta}$

We use backpropagation on the loss $\mathcal{L}(\cdot)$ to solve for the unknowns $\mathbf{\Theta}$. Notice that we can handle missing clinical data by dropping the contributions of the unknown value of $\mathbf{y}_{nm}$ to the network loss during backpropagation. Again, we use the ADAM optimizer [175] with random initialization at the first main iteration of alternating minimization. We employ a learning rate of $0.2e^{-4}$, scaled by 0.95 every 5 epochs, and batch-size 1. Additionally, we train the network only for 60 epochs to avoid overfitting.

5.5.4 Step 4: Updating the Constraint Variables $\{\mathbf{D}_n^t, \mathbf{\Lambda}_n^t\}$

Each of the primal variables $\{\mathbf{D}_n^t\}$ has a closed form solution given by:

$$[\mathbf{D}_n^t]^k = \mathbf{KF} \tag{5.14}$$

where, $\mathbf{K} = (\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T + \mathbf{\Gamma}_n^t\mathbf{L}_n\mathbf{B} + \mathbf{L}_n\mathbf{\Gamma}_n^t\mathbf{B} - \gamma\mathbf{\Lambda}_n)$ and $\mathbf{F} = (\gamma\mathcal{I}_K + 2\mathbf{L}_n)^{-1}$ We update the dual variables $\{\mathbf{\Lambda}_n\}$ via gradient ascent:

$$[\mathbf{\Lambda}_n^t]^{k+1} = [\mathbf{\Lambda}_n^t]^k + \eta_k([\mathbf{D}_n^t]^k - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)) \tag{5.15}$$

We cycle through the primal-dual updates for $\{\mathbf{D}_n^t\}$ and $\{\mathbf{\Lambda}_n^t\}$ in Eq. (5.14-5.15) to ensure that the constraints $\mathbf{D}_n^t = \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)$ are satisfied with increasing certainty at each iteration.

The learning rate parameter η_k for the gradient ascent step is selected to a guarantee sufficient decrease in the objective for every iteration of alternating minimization. In practice, we initialize η_0 to 10^{-3}, and scale it by 0.75 at each iteration k.

5.5.5 Step 5: Prediction on Unseen Data

In our cross-validated setting, we must compute the sr-DDL loadings $\{\bar{\mathbf{c}}^t\}_{t=1}^{\bar{T}}$ for a new subject based on the $\mathbf{B}^*$ obtained from the training procedure and the new rs-fMRI correlation matrices $\{\bar{\mathbf{\Gamma}}^t\}$ and DTI Laplacians $\bar{\mathbf{L}}$. As we do not know the score $\bar{\mathbf{y}}$ for this individual, we need remove the contribution $\mathcal{L}(\cdot)$ from Eq. (5.10) and assume that the constraints $\bar{\mathbf{D}}^t = \mathbf{B}^*\mathbf{diag}(\bar{\mathbf{c}}^t)$ are satisfied with equality. This effectively eliminates the Lagrangian terms. Essentially, the optimization for $\{\bar{\mathbf{c}}^t\}$ now reduces to $\bar{T}_n$ decoupled quadratic programming

(QP) objectives $\mathcal{Q}_t$:

$$\bar{\mathbf{c}}^{*t} = \text{argmin}_{\bar{\mathbf{c}}^t} \frac{1}{2} \, (\bar{\mathbf{c}}^t)^T \bar{\mathbf{H}} \bar{\mathbf{c}}^t + \bar{\mathbf{f}}^T \bar{\mathbf{c}}^t \;\; s.t. \;\; \bar{\mathbf{A}} \bar{\mathbf{c}}^t \leq \bar{\mathbf{b}}$$

$$\bar{\mathbf{H}} = 2({\mathbf{B}^*}^T \bar{\mathbf{L}} \mathbf{B}^*);$$

$$\bar{\mathbf{f}} = -[\mathcal{I}_K \circ ({\mathbf{B}^*}^T (\bar{\mathbf{\Gamma}} \bar{\mathbf{L}} + \bar{\mathbf{L}} \bar{\mathbf{\Gamma}}^t) \mathbf{B}^*)]\mathbf{1};$$

$$\bar{\mathbf{A}} = -\mathcal{I}_K \; \bar{\mathbf{b}} = \mathbf{0}$$

Notice that decoupling the objective across time allows us to parallelize this computation. Additionally, since $\bar{\mathbf{H}}$ is positive semi-definite, the formulation above is convex, leading to an efficient QP solution. Finally, we estimate $\bar{\mathbf{y}}$ via a forward pass through the LSTM-ANN.

Overall, our alternating minimization training procedure explicitly couples the Dictionary Learning (sr-DDL) and Deep Network (LSTM-ANN) blocks within the optimization. In contrast, the setup at test time consists of two steps, namely the coefficient update followed by a forward pass through the LSTM-ANN. We will demonstrate via our experiments (i.e. Section 5.8) that the coupled training is key to generalization. Finally, we discuss the effect of this difference between the training and testing procedures further in Section 5.10

5.6 Model Evaluation

We evaluate our deep-generative hybrid on two separate cohorts. The first dataset is a cohort of 150 healthy individuals from the Human Connectome

Project (HCP) database [176] having both the rs-fMRI and DTI scans. We refer to this as the HCP dataset. Cognitive outcomes such as fluid intelligence are believed to be closely connected to structural (SC) and function connectivity (FC) in the human brain [177]. Thus, jointly modeling multimodal neuroimaging and cognitive data helps exploit this fundamental interweave and uncover the neural underpinnings of cognition. Finally, we chose to focus on a modest sized dataset ($N = 150$) to demonstrate that our framework is suitable for clinical rs-fMRI applications, many of which have limited sample sizes. Our second dataset is the clinical ASD dataset described in Chapter 2.

5.6.1 Secondary HCP Dataset

5.6.1.1 Neuroimaging Data

As described in [176], the HCP S1200 dataset was acquired on a Siemens 3T scanner (TR/TE= $0.72ms/0.33ms$, spatial resolution $= 2 \times 2 \times 2$mm). The rs-fMRI scans were processed according to the standard pre-processing pipeline described in [178], which includes additional processing to account for confounds due to motion and physiological noise. We opted to use a 15 minute interval (typical of clinical rs-fMRI studies of neurodevelopmental disorders) from the second scan of each subject's first visit for our analysis.

The DTI data from the HCP dataset was processed using the standard Neurodata MR Graphs package (ndmg) [179]. This consists of co-registration to anatomical space via FSL [135], followed by tensor estimation in the MNI space and probabilistic tractography to compute the fibre tracking streamlines.

5.6.1.2 Behavioral Data

For the HCP database, we examine the Cognitive Fluid Intelligence Score (CFIS) described in [180, 181], adjusted for age. This is scored based on a battery of tests measuring cognitive reasoning, considered a nonverbal estimate of fluid intelligence in subjects. The dynamic range for the score is $70 - 150$, with higher scores indicating better cognitive abilities.

5.6.2 Implementation Details

5.6.2.1 Architectural Details

Our proposed ANN architecture is highlighted in the white box to the bottom left of Fig. 5.1. Our modeling choices carefully control for representational capacity and convergence of our coupled optimization procedure. Since the input to the network, i.e. the coefficient vector $\mathbf{c}_n^t$ is essentially low dimensional, we opt for a two layered LSTM with the hidden layer width as 40. Both the P-ANN and the A-ANN are fully connected neural networks with two hidden layers of width 40. Since the A-ANN outputs a scalar, the width of its output layer is one, while that of the P-ANN is of size M, i.e. the number of behavioral scores. We use a Rectified Linear Unit (ReLU) as the activation function for each hidden layer, as we found that this choice is robust to issues with vanishing gradients and saturation that commonly confound the training of deep neural networks [182].

5.6.2.2 Parameter Settings

In order to fix the hyperparameters for our model and the baselines, we make use of a second subset of 130 individuals from the HCP database (hereby referred to as HCP-2). Note that these individuals have no overlap with those used characterize the performance in Section 5.8 to avoid biasing the results. First, we set aside 30 of these patients as a validation set to determine appropriate learning rates for our method and baselines. Recall that our deep-generative hybrid has two free parameters: namely the penalty λ, which controls the tradeoff between data representation and clinical prediction, and K, the number of networks. For our experiments, we chose $K = 15$ (See Fig. 5.3) for both datasets based on the knee point of the eigenspectrum of the correlation matrices $\{\mathbf{\Gamma}_n^t\}$. Based on the results of a 5 fold cross validation and grid search on HCP-2, we fix $\lambda = 2.5$. We will further discuss the robustness to λ in Section 5.10. Additionally, our sliding window protocol is defined by two parameters, namely the window length and stride. Although these are not hyperparameters for the sr-DDL per se, they affect the predictive performance by controlling the information overlap between successive dynamic rs-fMRI correlation matrices. Again, these are set based on the cross validation performance on HCP-2. We will further discuss the robustness to these parameters in Section 5.10.

Our experiments rely on the Automatic Anatomical Labelling (AAL) atlas [142] parcellation for the rs-fMRI and DTI data. AAL consists of 116 cortical, subcortical and cerebellar regions. We employ a sliding window protocol as shown in Fig. 2.3. Due to the different TR, we set the sliding window parameters to window length $= 156$ and stride $= 17$ for the HCP dataset, and window length $= 45$ and stride $= 5$ for the KKI dataset to extract dynamic correlation matrices from the 116 average time courses. We discuss the sensitivity to this choice in Section 5.10. Thus, for each individual, we have correlation matrices of size 116×116 based on the Pearson's Correlation Coefficient between the average regional time-series. Empirically, we observed a consistent noise component with nearly unchanging contribution from all brain regions and low predictive power for both datasets. Therefore, we subtracted out the first eigenvector contribution from each of the correlation matrices and used the residuals as the inputs $\{\mathbf{\Gamma}_n\}$ to the algorithm and the baselines.

5.6.2.3 Initialization

Our coupled optimization strategy requires us to initialize the basis $\mathbf{B}$, coefficients $\{\mathbf{c}_n^t\}$, the deep network weights $\mathbf{\Theta}$ and the constraint variable pairs

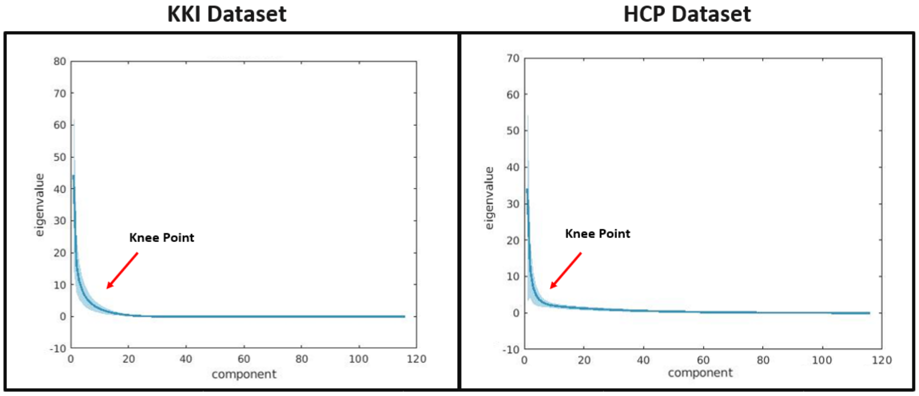


Figure 5.3: Scree Plot of the correlation matrices to corroborate the selected values for K. **(L)** KKI Dataset **(R)** HCP Dataset. The thick line denotes the mean eigenvalue, while the shaded area indicates the standard deviation across subjects and time points.

$\{\mathbf{D}_n^t, \mathbf{\Lambda}_n^t\}$. We randomly initialize the deep network weights at the first main iteration. We employ a soft-initialization for $\{\mathbf{B}, \{\mathbf{c}_n^t\}\}$ by solving the dictionary objective in Eq. (5.6) without the LSTM-ANN loss terms for 20 iterations. We then initialize $\mathbf{D}_n^t = \mathbf{Bdiag}(\mathbf{c}_n^t)$ and $\mathbf{\Lambda}_n^t = \mathbf{0}$ which lie in the feasible set for our constraints. We empirically observed that this soft initialization helps stabilize the optimization to provide improved predictive performance in fewer main iterations when compared with a completely random initialization.

5.6.3 Baseline Methods

We evaluate the performance of our framework against three different classes of baselines, each highlighting the benefit of specific modeling choices made by our method.

Our first baseline class is a two stage configuration as illustrated in Fig. 5.4 that combines feature extraction on the dynamic rs-fMRI and DTI data, with a deep learning predictor. These feature engineering techniques are drawn from a set of well established statistical (Independent Component Analysis in Subsection **??**) and graph theoretic techniques (Betweenness Centrality in Subsection 5.6.3.1), known to provide rich feature representations. The learned features are then input to the same deep LSTM-ANN network used by our method. This network is trained separately to predict the clinical outcomes. Note that these baselines incorporate multimodal and dynamic information, but do not directly operate on the network structure of the connectomes. Our second baseline class omits the two step approach in lieu of an end-to-end convolutional neural network based on the work of [44]. We train this model

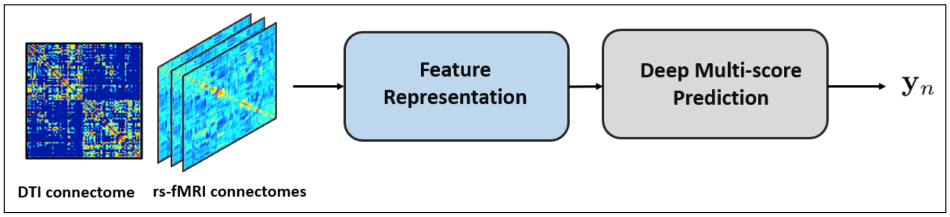


Figure 5.4: A typical two stage baseline. We input the dynamic correlation matrices and DTI connectomes to Stage 1, which performs Feature Extraction. This step could be a technique from machine learning, graph theory or a statistical measure. Stage 2 is a deep network that predicts the clinical scores

on the static rs-fMRI and DTI connectomes in tandem to predict the clinical scores. This baseline operates directly on the correlation and connectivity matrices, but ignores the dynamic evolution of functional connectivity. Next, we present the comparison of our deep sr-DDL by omitting the structural regularization. This helps us evaluate the benefit provided by the multimodal integration of DTI and rs-fMRI data. Our final baseline highlights the benefit of our joint optimization procedure. In this experiment, we decouple the optimization of the dynamic matrix factorization and deep network in Fig. 5.1 similar to the two stage pipelines.

5.6.3.1 Graph Theoretic Feature Selection

Notice that the subject-specific correlation rs-fMRI matrices $\{\mathbf{\Gamma}_n^t\}$ and the corresponding binary DTI adjacency matrices $\mathbf{A}_n$ indicate time-varying functional and anatomical connectivity between the ROIs respectively. Therefore, we multiply the two to generate the time-varying multimodal graphs whose nodes are the brain ROIs and edges are defined by the temporal connectivity between these ROIs. We denote the corresponding adjacency matrices for these graphs by $\{\mathbf{\Psi}_n^t = \mathbf{A}_n \circ \mathbf{\Gamma}_n^t \in \mathcal{R}^{P\times P}\}$, where we threshold each $\mathbf{\Psi}_n^t$ to remove negative values. Each element $[\mathbf{\Psi}_n^t]_{ij}$ gives the strength of association between two communicating sub-regions i and j in individual n at time t. We summarize the topology of these graphs via **Betweenness Centrality (C_B)** to obtain a time-varying estimate of brain connectivity for each ROI [28, 82]. $\mathbf{C}_B(v)$ for region v is calculated as:

$$\mathbf{C}_B^t(v) = \sum_{s \neq v \neq u \in V} \frac{\boldsymbol{\sigma}_{su}^t(v)}{\boldsymbol{\sigma}_{su}^t} \tag{5.16}$$

$\boldsymbol{\sigma}_{su}^t$ is the total number of shortest paths from node s to node u at time t, and $\boldsymbol{\sigma}_{su}^t(v)$ is the number of those paths that pass through v. This measure quantifies the number of times a node acts as a bridge along the shortest path between two other nodes and has found wide usage in characterizing small-worlded networks in brain connectivity [28]. We effectively reduce the dimensionality of the connectivity features. Again, the collection of features $\{\mathbf{C}_B^t\}$ are used to train an LSTM-ANN predictor from Fig. 5.1 with two hidden layers having width 200 due to the higher input feature dimensionality.

5.6.3.2 ICA Feature Selection

This baseline employs **Independent Component Analysis (ICA)** combined an the LSTM-ANN predictor. ICA is a statistical technique that extracts representative spatial patterns from the rs-fMRI time series. It has now become ubiquitous in fMRI analysis for its ability to identify group level differences as well as model individual-specific connectivity signatures. Essentially, ICA decomposes multivariate signals into 'independent' non-Gaussian components based on the data statistics.

This algorithm can be extended to the multi-subject analysis setting via Group ICA (G-ICA). Specifically, we extract independent spatial patterns common across patients, by combining the contribution of the individual time courses. For this baseline, we first perform G-ICA using the GIFT toolbox [149], and derive independent spatial maps for each subject from their raw rs-fMRI scans. We then compute the average time courses for each spatial map considering the constituent voxels. This provides us with a feature representation of reduced dimension equal to the number of specified maps ($d << L$) for each individual. For our experiments, we extract 15 ICA components. These time courses are input into the LSTM-ANN network in Fig. 5.1 with two hidden layers of width 40 to predict the clinical outcomes.

5.6.3.3 BrainNet Convolutional Neural Network

The BrainNet CNN [44] relies on specialized fully convolutional layers for feature extraction, and was originally used to predict cognitive and motor outcomes from DTI connectomes. Fig. 5.5 provides a pictorial overview of the original architecture adapted for clinical outcome prediction from multimodal data. Each branch of the network accepts as input a $P \times P$ connectome, to which it applies a cascade of two edge-edge (E-E) convolutional operations. This E-E operation combines individual convolutions acting on the row and column to which the input element belongs. It is followed by a series of edge-node (E-N) blocks that reduce the dimensionality of the intermediate outputs, followed by a node-graph (N-G) operation for pooling. Finally, the output clinical scores are predicted via a fully connected artificial neural network for regression.

We feed the rs-fMRI static connectomes ($\hat{\mathbf{\Gamma}}_n$) and DTI Laplacians $\mathbf{L}_n$ into two disjoint fully convolutional branches with the architecture described above. We integrate the learned features via concatenation and input them into the fully connected layers described in Fig. 5.5, but with the number of outputs equal to the dimensionality of the clinical severity vector $\mathbf{y}_n$. We set the learning rate, momentum and weight decay parameters according to the guidelines in [44].

5.6.3.4 Deep sr-DDL without DTI regularization

In this baseline, we examine the effect of excluding the structural regularization provided by the DTI data from the joint objective in Eq. (5.9). The

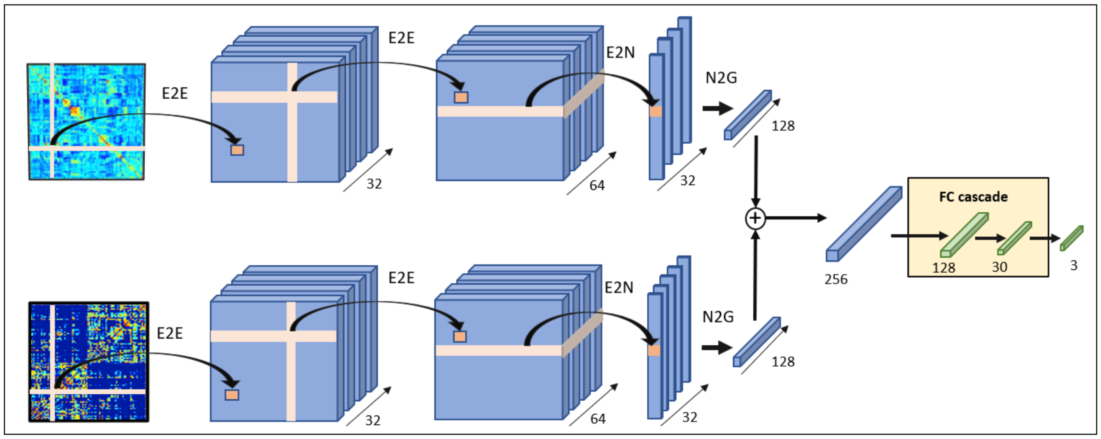


Figure 5.5: The BrainNet CNN baseline [44] for severity prediction from multimodal data

resulting objective function takes the following form:

$$\mathcal{J}_w(\mathbf{B}, \{\mathbf{c}_n^t\}, \mathbf{\Theta}; \{\mathbf{\Gamma}_n^t\}, \{\mathbf{y}_n\}) = \sum_n \sum_t \frac{1}{T_n} ||\mathbf{\Gamma}_n^t - \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T||_F^2$$

$$+ \lambda \sum_n \mathcal{L}(\mathbf{\Theta}, \{\mathbf{c}_n^t\}; \mathbf{y}_n) \;\; s.t. \;\; \mathbf{c}_{nk}^t \geq 0, \;\; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K. \quad (5.17)$$

Notice that amounts to replacing the Weighted Frobenius Norm formulation by a regular ℓ_2 penalty. This allows us to adopt the alternating minimization procedure in Section 5.4 to optimize Eq. (5.17) with a few minor modifications. Specifically, instead of T_n constraints per subject, we use a single constraint of the form $\mathbf{D} = \mathbf{B}$, enforced via a single Augmented Lagrangian $\mathbf{\Lambda}$. This effectively ensures that the new objective has a quadratic form in $\mathbf{B}$, along with a closed form update for $\mathbf{D}$. As before, we cycle through four individual steps, namely:

- Closed form Procrustes solution for the basis $\mathbf{B}$
- Updating the temporal loadings $\{\mathbf{c}_n^t\}$ (ADAM)
- Updating the Neural Network Parameters $\mathbf{\Theta}$ (ADAM)
- Augmented Lagrangian updates for the constraint variables $\{\mathbf{D}, \mathbf{\Lambda}\}$

We use $K = 15$ networks as inputs to the LSTM-ANN network with two hidden layers of width 40 to predict the clinical outcomes.

5.6.3.5 Deep sr-DDL without dynamics

This baseline examines the effectiveness of using an LSTM-ANN framework to track the temporal changes in connectivity data. We use the same matrix decomposition framework as the sr-DDL, but remove the LSTM-ANN and replace it with just a simple ANN model. For each dynamic connectivity matrix, this model provides an estimate of the clinical severity profile at the given time point, effectively treating each time point as independent. To obtain the final severity profile, we average the individual estimates. Again, we use $K = 15$ networks and an ANN with hidden layer width 40

5.6.3.6 Decoupled Deep sr-DDL

Our final baseline examines the efficacy of our coupled optimization procedure in Section 5.4 with regards to generalization onto unseen subjects. Here, we first run the feature extraction using the sr-DDL optimization to extract the basis $\mathbf{B}$ and temporal loadings $\{\mathbf{c}_n^t\}$. We then use the $\{\mathbf{c}_n^t\}$ as inputs to train the LSTM-ANN network in Fig. 5.1 to predict the scores $\mathbf{y}_n$. This is akin to the two-stage baselines delineated in Fig. 5.4. Again, we use $K = 15$ networks with an a two layered LSTM-ANN having hidden layer width 40

5.7 Experiments on Synthetic Data

This experiment allows us to assess the behavior of our algorithm under various noise scenarios. The equivalent generating process for our framework is captured by the graphical model in Fig. 5.6. The observed variables are the temporal correlation matrices $\{\mathbf{\Gamma}_n^t\}$, the DTI Laplacians $\mathbf{L}_n$, and the clinical scores $\{\mathbf{y}_n\}$, while the latent variables are the basis $\mathbf{B}$, the coefficients $\{\mathbf{c}_n^t\}$, and the neural network weights $\mathbf{\Theta}$. Note that the dynamic correlation matrices $\{\mathbf{\Gamma}_n^t\}$ are completely described by the basis $\mathbf{B}$, the coefficients $\{\mathbf{c}_n^t\}$

and the Laplacian weighting $\mathbf{L}_n$. We further observe that the rs-fMRI data decompositions for each subject couple only through the shared basis and the clinical predictions through the shared network weights $\mathbf{\Theta}$. Conditioned on these variables, $\{\{\mathbf{\Gamma}_n^t\}, \mathbf{L}_n, \{\mathbf{c}_n^t\}, \mathbf{\Theta}, \mathbf{y}_n\}$ are independent across subjects. Fig. 5.6 captures these conditional relationships.

We start by generating a basis matrix $\hat{\mathbf{B}} \in \mathcal{R}^{P \times K}$ by drawing its entries independently from a zero mean Gaussian with variance one. We then use the Gram-Schmidt procedure to compute an orthogonal basis $\mathbf{B}_o = \mathbf{orth}(\hat{\mathbf{B}})$. Finally, we simulate corruptions to this basis via additive Gaussian noise $\mathbf{B} = \mathbf{B}_o + \mathcal{N}(0, \sigma_{\mathbf{B}})$. Effectively, the value of $\sigma_{\mathbf{B}}$ quantifies the deviations of $\mathbf{B}$ from orthogonality, which is an assumption of our model. Note that the coefficient values in $\mathbf{c}_n$ are independent across networks and subjects, but not across time.

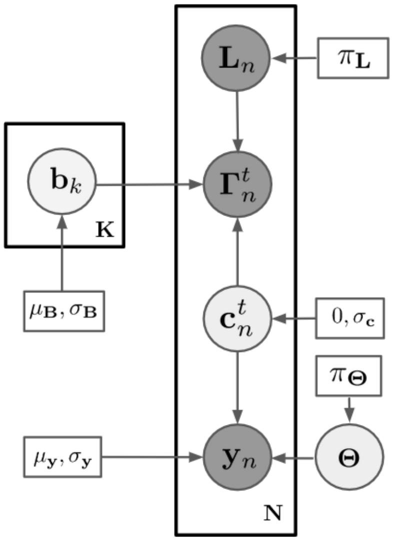


Figure 5.6: The graphical model for generating synthetic data. We fix the model parameters $\sigma_{\mathbf{c}} = 4$, number of subjects N at 60, and number networks K at 4. The dimensionality of $\mathbf{y}_n$ is $M = 3$ and the length of the scan $T_n = 30$ for each subject. The shaded circles denote observed variables, while the clear circles indicate latent variables.

Thus, for each subject, we generate the temporal coefficients using a isotropic Gaussian process with zero mean, and variance $\sigma_{\mathbf{c}}$. These values are clipped at 0 to reflect the non-negativity in the coefficients. The variance parameter $\sigma_{\mathbf{c}}$ defines the scale of the coefficients. Next, we simulate the Graph Laplacians $\mathbf{L}_n$ for each subject based on structural connectivity priors computed using real-world data. Specifically, for each region pair, we first create a histogram of connectivity using binary adjacency matrices from the HCP database. With $\pi_{\mathbf{L}}$ denoting the probability of a connection between ROI pairs, we sample a symmetric graph adjacency matrix $\mathbf{A}_n$ per subject via a Bernouilli distribution with parameter $\pi_{\mathbf{L}}$. We then compute the corresponding Laplacians $\mathbf{L}_n$ from $\mathbf{A}_n$. This choice of prior helps us generate realistic structural connectivity profiles. Now, recall that our model seeks to approximate the rs-fMRI dynamic correlation matrices by $\mathbf{\Gamma}_n^t \approx \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T$. Additionally, this decomposition is regularized by the individual Laplacians $\mathbf{L}_n$. Since we wish to evaluate the quality of this approximation, our generative model simulates $\mathbf{\Gamma}_n^t$ by adding structured noise (parametrized by $\mathbf{L}_n$) to $\mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T$. Specifically, we use the eigenbasis $\mathbf{X}$ of $\mathbf{L}_n$ to generate additive noise $\mathbf{N} = \sigma_{\mathbf{\Gamma}}\mathbf{X}\mathbf{X}^T$. We then compute the correlation matrices as $\mathbf{\Gamma}_n^t = \mathbf{B}\mathbf{diag}(\mathbf{c}_n^t)\mathbf{B}^T + \mathbf{N}$. Note that this procedure preserves the positive semi-definiteness of the decomposition. Effectively, the parameter σ_{Γ} controls the level of corruption in the observed dynamic correlation matrices. Finally, the observed variable $\{\mathbf{y}_n\}$, translates to a Gaussian with mean $\mu_{\mathbf{y}_n} = \mathcal{F}_{\mathbf{\Theta}}(\{\mathbf{c}_n^t\}) \in \mathcal{R}^{M\times 1}$, and variance $\sigma_{\mathbf{y}_n}\mathbf{I}_M$. The function mapping $\mathcal{F}_{\mathbf{\Theta}}$ refers to the LSTM-ANN network with the parameters $\mathbf{\Theta}$ - which we randomly initialize. This is again folded to reflect positive values of $\mathbf{y}_n$. Here, $\sigma_{\mathbf{y}}$ controls the noise in the clinical scores.

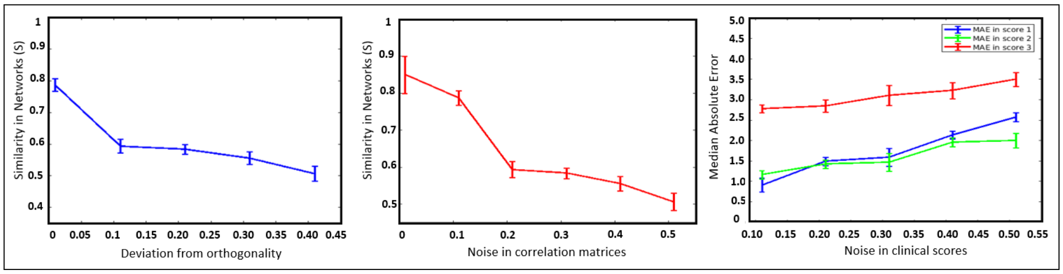


Figure 5.7: Performance on synthetic experiments. (**L**): Varying the level of deviation from orthogonality ($\sigma_{\mathbf{\Gamma}} = 0.2$, $\sigma_{\mathbf{Y}} = 0.2$), (**M**): Varying the level of noise in $\mathbf{\Gamma}$ ($\sigma_{\mathbf{B}} = 0.2$, $\sigma_{\mathbf{y}} = 0.2$) , (**R**): Varying the level of noise in $\mathbf{y}_n$ under ($\sigma_{\mathbf{B}} = 0.2$, $\sigma_{\mathbf{\Gamma}} = 0.2$) Values on the x-axis have been normalized to reflect a $[0-1]$ range by dividing by the maximum value of the variable. We report deviations from the mean for recovered similarity/MAE at each parameter setting in terms of a standard error value. The reported x-axis range reflects the regimes within which the algorithm converges to a local solution

There are two sources of noise for the observed variables. The first is error in the correlation matrices $\mathbf{\Gamma}_n^t$, controlled by changing $\sigma_{\mathbf{\Gamma}}$. The second case is error in the clinical scores $\mathbf{y}_n$, quantified by the parameter $\sigma_{\mathbf{y}}$. Additionally, we are also interested in evaluating the performance under varying levels of deviations of the basis from orthogonality, controlled by the parameter $\sigma_{\mathbf{B}}$.

We evaluate the efficacy of our algorithm using two separate metrics. The first is an average absolute cosine similarity measure S between each recovered network, $\bar{\mathbf{b}}_k$, and its corresponding best matched ground truth network, $\mathbf{b}_k$, normalizing the latter to unit norm, that is:

$$S = \frac{1}{K}\sum_k \frac{|\mathbf{b}_k^T \bar{\mathbf{b}}_k|}{||\mathbf{b}_k||_2}. \tag{5.18}$$

The second metric is the Median Absolute Error (MAE) between the output of the trained LSTM-ANN $\hat{\mathbf{y}}_n$ and the true scores $\mathbf{y}_n$.

Fig. 5.7 depicts the performance of the algorithm in these three cases. In the each subplots, the x-axis corresponds to increasing the levels of noise. In the first two subplots, the y-axis indicates the similarity metric S computed for the particular setting, while in the rightmost subplot, we plot the MAE for predicting the three scores. All numerical results have been aggregated over 50 independent trials.

In the leftmost plot, an x-axis value close to 0 indicates low levels of deviation of $\mathbf{B}$ from orthogonality, while increasing values corresponds to a more severe deviation from the modeling assumptions. During this experiment, the values of the other free parameters in Fig. 5.6 were held constant. We observed that the MAE of the three scores remains roughly constant for all noise settings (score 1—1.49 ± 0.09, score 2—1.34 ± 0.07, score 3—3.10 ± 0.11). The middle plot evaluates subnetwork recovery when the noise in the dynamic correlation matrices, i.e. $\sigma_{\mathbf{\Gamma}}$ is increased. The **x**-axis reports normalized values of $\sigma_{\mathbf{\Gamma}_n}$ while the remaining free parameters were held constant. Similar to the previous scenario, the MAE remains roughly constant for varying noise settings (score 1—1.50 ± 0.08, score 2—1.50 ± 0.06, score 3—2.96 ± 0.50). Finally, the rightmost plot in Fig. 5.7 indicates performance under varying noise in the scores $\mathbf{y}_n$. Again, normalized $\sigma_{\mathbf{y}}$ values are reported on the x-axis. For this experiment, we observed that $S = 0.87 \pm 0.05$ for varying noise levels.

As expected, increased noise in the correlation matrices and deviations from orthogonality worsens recovery performance of the algorithm. This is reflected by the decay in the similarity measure along with increasing noise parameters. Since the parameter $\sigma_{\mathbf{y}}$ is held constant, we do not observe much variation in the the MAE values upon increasing the noise. Lastly, we notice

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test
CFIS	Median	N/A	13.51 ± 9.97	N/A	0
	BC & LSTM-ANN	7.23 ± 6.24	16.50 ± 13.60	0.53	0.72
	ICA & LSTM-ANN	4.87 ± 4.84	16.45 ± 14.7	0.58	**0.77**
	BrainNet CNN	3.50 ± 2.1	16.89 ± 12.20	0.79	0.73
	Decoupled	3.72 ± 4.33	18.10 ± 14.04	0.78	0.70
	W/O DTI reg.	0.77 ± 0.66	20.02 ± 15.04	**0.88**	0.74
	W/O dynamics.	0.97 ± 0.21	15.14 ± 14.71	**0.79**	0.71
	Deep sr-DDL	**0.44** ± **0.15**	**14.76** ± **12.77**	0.86	**0.77**

Table 5.1: HCP Dataset: Performance evaluation on the HCP dataset against our prior work according to **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)**. We also report the standard deviation for the MAE Lower MAE and higher NMI indicate better performance. Best performance is highlighted in bold.

that the algorithm performs better when the level of noise in the scores is lower. This is indicated by the increasing values of MAE in the right subplot in Fig. 5.7. Since $\sigma_{\mathbf{B}}$ is held constant for this experiment, the metric S remains fairly constant even upon increasing the noise in the scores.

Taken together, our simulations indicate that the optimization procedure is robust in the noise regime $(0.01 - 0.2)$ estimated from the real-world rs-fMRI data. In addition, these experiments help us identify the stable parameter settings ($\lambda = 1 - 10$, learning rates) which govern the convergence of the algorithm which guide our real world experiments.

5.8 Population Studies

Fig. 5.8 illustrates the performance comparison of our deep sr-DDL framework against the baselines in Section 5.6.3 on the HCP dataset for predicting the CFIS. Fig. 5.9 presents the same comparison on the KKI dataset for multi-score prediction. In each figure, the scores predicted by the algorithm are plotted on the **y**-axis against the measured ground truth score on the **x**-axis. The bold

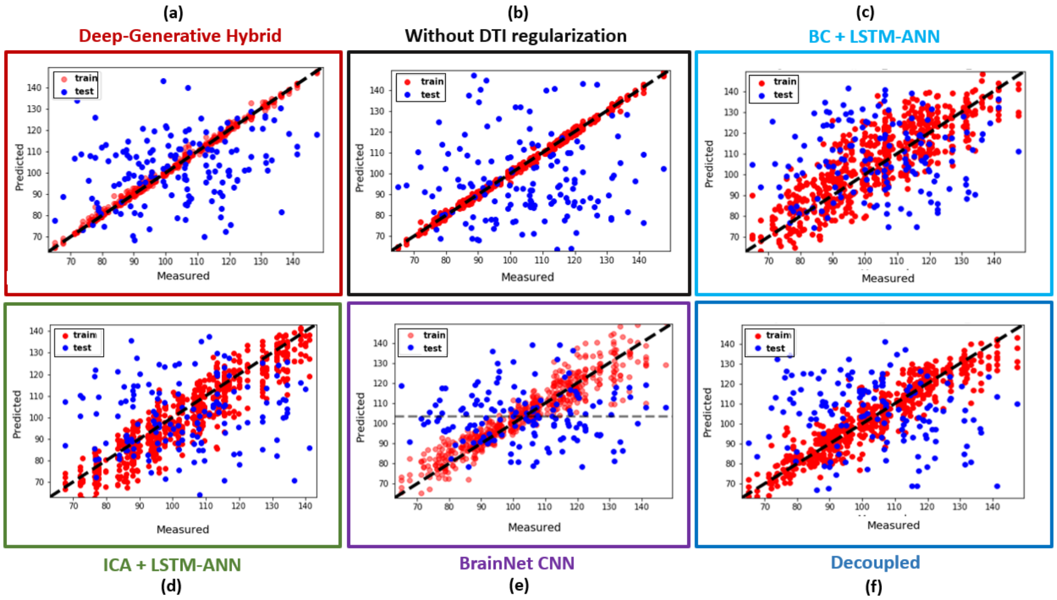


Figure 5.8: HCP dataset: Prediction performance for the Cognitive Fluid Intelligence Score by the (a) **Red Box:** Deep sr-DDL. (b) **Black Box:** Deep sr-DDL model without DTI regularization (c) **Light Purple Box:** Betweenness Centrality on DTI + dynamic rs-fMRI multimodal graphs followed by LSTM-ANN predictor (d) **Green Box:** ICA timeseries followed by LSTM-ANN predictor (e) **Purple Box**: Branched BrainNet CNN [44] on DTI and rs-fMRI static graphs (f) **Blue Box:** Decoupled DDL factorization followed by LSTM-ANN predictor

$\mathbf{x} = \mathbf{y}$ line represents ideal performance. The red points represent the training data, while the blue points indicate the held out testing data for all the cross validation folds.

We observe that the training performance of the baselines is good (i.e. the red points follow the $\mathbf{x} = \mathbf{y}$ line) in all cases for both datasets. However, in case of testing performance, our method outperforms the baselines in all cases. This performance gain is particularly pronounced in the case of multiscore prediction (KKI dataset). Empirically, we are able to tune the baseline hyperparameters to obtain good testing performance on the KKI dataset for a single score (ADOS for ICA+LSTM-ANN), but the prediction of

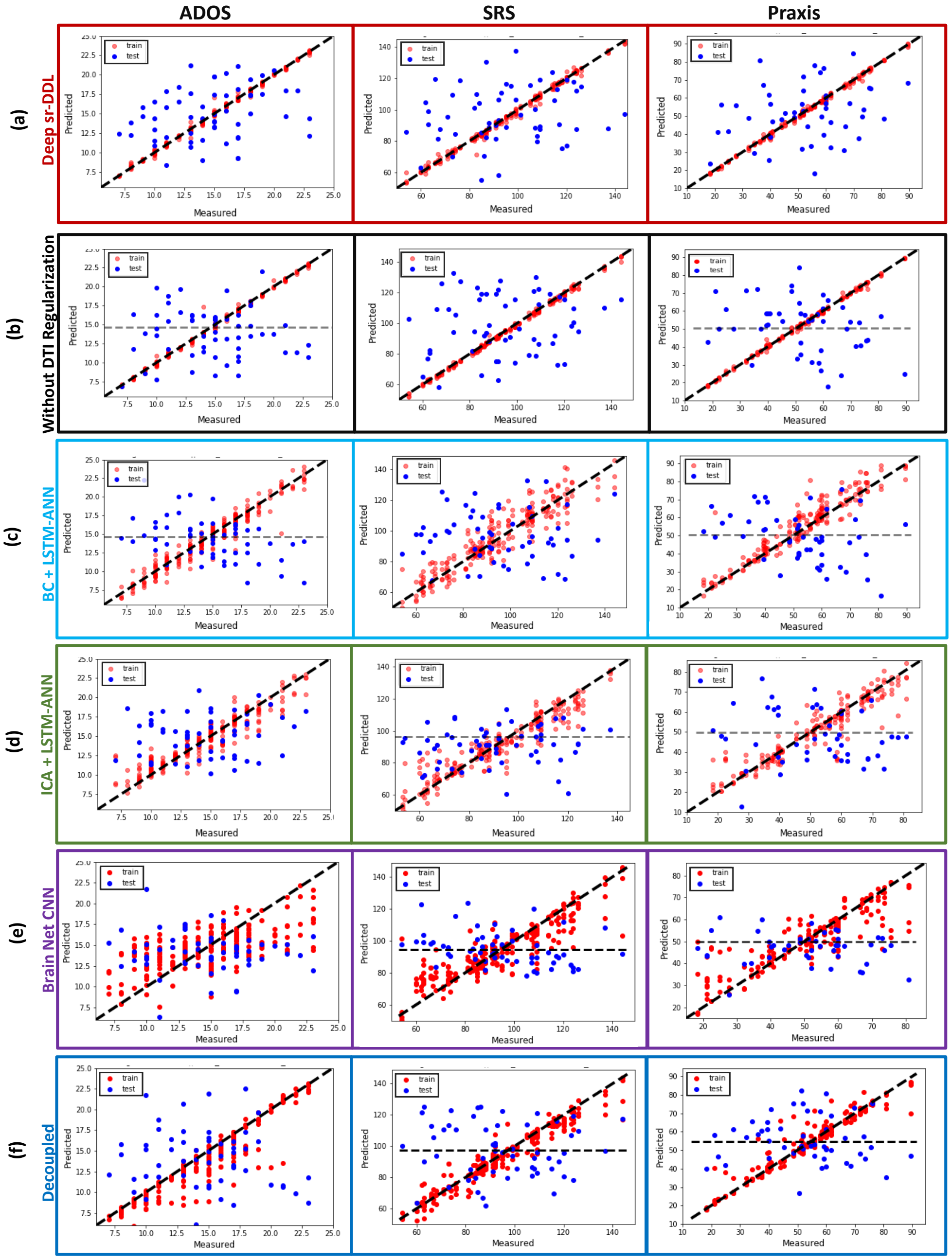


Figure 5.9: KKI dataset: Multiscore prediction performance for the **(L)** ADOS, **(M)** SRS, and **(R)** Praxis by the **(a) Red Box:** Deep sr-DDL **(b) Black Box:** Model without DTI regularization **(c) Light Purple Box:** Betweenness Centrality on DTI + dynamic rs-fMRI multimodal graphs followed by LSTM-ANN predictor **(d) Green Box:** ICA timeseries followed by the LSTM-ANN predictor **(e) Purple Box**: Branched BrainNet CNN [44] on DTI Laplacian and rs-fMRI static graphs **(f) Blue Box:** Decoupled DDL factorization followed by LSTM-ANN predictor

the remaining scores (SRS and Praxis for the KKI dataset) suffers. Notice that the prediction on one or more of scores (KKI dataset) and CFIS (HCP dataset) hovers around the population median of the score in several cases. In fact, in some of the multi-score prediction cases, it performs worse than predicting the median. This is testament to the inherent difficulty of the prediction task at hand. Finally, we notice that omitting the structural regularization from the deep sr-DDL performs worse than our method.

In contrast to the baselines, the testing predictions of our framework follow

Score	Method	MAE Train	MAE Test	NMI Train	NMI Test
ADOS	Median	N/A	2.33 ± 2.01	N/A	0
	BC & LSTM-ANN	0.68 ± 0.57	4.36 ± 3.36	0.89	0.29
	ICA & LSTM-ANN	0.9 ± 0.54	**2.47** ± **2.04**	0.91	**0.41**
	BrainNet CNN	1.90 ± 0.086	3.50 ± 2.20	0.96	0.25
	Decoupled	1.34 ± 0.51	3.93 ± 2.10	0.68	0.29
	W/O DTI reg.	0.25 ± 0.099	3.50 ± 3.09	0.99	0.17
	W/O dynamics.	1.56 ± 1.51	3.17 ± 2.54	**0.95**	0.29
	Deep sr-DDL	**0.2** ± **0.09**	2.99 ± 1.99	**0.99**	0.37
SRS	Median	N/A	16.81 ± 12.8	N/A	0
	BC & LSTM-ANN	5.10 ± 4.61	18.05 ± 14.22	0.92	0.83
	ICA & LSTM-ANN	5.27 ± 3.32	**13.64** ± **12.69**	0.76	0.59
	BrainNet CNN	5.25 ± 2.5	18.96 ± 15.65	0.83	0.75
	Decoupled	2.10 ± 2.98	21.45 ± 13.73	0.76	0.78
	W/O DTI reg.	**0.72** ± **0.61**	22.20 ± 14.78	0.95	0.65
	W/O dynamics.	3.25 ± 2.74	19.05 ± 18.19	0.93	0.73
	Deep sr-DDL	1.21 ± 0.66	18.70 ± 13.51	**0.98**	**0.85**
Praxis	Median	N/A	10.53 ± 8.81	N/A	0
	BC & LSTM-ANN	6.61 ± 3.30	17.49 ± 9.08	0.86	0.70
	ICA & LSTM-ANN	4.56 ± 1.26	15.02 ± 11.80	0.82	0.60
	BrainNet CNN	3.78 ± 0.59	15.15 ± 11.49	0.95	0.19
	Decoupled	1.57 ± 1.12	21.67 ± 12.02	0.75	0.25
	W/O DTI reg.	**0.61** ± **0.29**	18.56 ± 14.32	**0.96**	0.65
	W/O dynamics.	1.67 ± 2.31	16.22 ± 14.91	0.94	**0.82**
	Deep sr-DDL	0.62 ± 0.36	**14.99** ± **10.17**	0.95	**0.82**

Table 5.2: KKI Dataset: Performance evaluation on the KKI dataset against our prior work according to **Median Absolute Error (MAE), Normalized Mutual Information (NMI)**. We also report the standard deviation for the MAE Lower MAE and higher NMI indicate better performance. Best performance is highlighted in bold.

the $\mathbf{x} = \mathbf{y}$ more closely. The machine learning, statistical and graph theoretic techniques we selected for a comparison are well known in literature for being able to robustly provide compact characterizations for high dimensional datasets. However, we see that ICA is unable to estimate a reliable projection of the data that is particularly useful for behavioral prediction. Similarly, the betweenness centrality measure is unable to extract informative topologies for brain-behavior integration. We conjecture that the aggregate nature of this measure is useful for capturing group-level commonalities, but falls short of modeling subject-specific differences. Furthermore, even the BrainNet CNN, which directly exploits the graph structure of the connectomes falls short of generalizing to multi-score prediction. Additionally, it ignores the dynamic information in the rs-fMRI data. In case of the baseline where we omit the structural regularization, i.e. deep sr-DDL without DTI, we notice that the method learns a representation of the rs-fMRI data that generalizes beyond the training set, but still falls short of the performance when anatomical information is included. We see very similar behavior upon removing the LSTM-ANN in the no dynamic sr-DDL baseline. This clearly demonstrates the benefit of supplementing the functional data with structural priors as well as modeling the time-varying nature of connectivity. Finally, the failure of the decoupled dynamic matrix factorization and deep-network makes a strong case for jointly optimizing the neuroimaging and behavioral representations. The basis estimated independently of behavior are not indicative of clinical outcomes, due to which the regression performance suffers. We also quantify the performance indicated in these figures in Table 5.1 (HCP dataset) and Table 5.2 (KKI dataset) based on the MAE and NMI. For reference, we have added an additional row as a ‘baseline’ in our tables where for each test subject, we simply predict the median of each score.

Our deep sr-DDL framework explicitly optimizes for a viable tradeoff between multimodal and dynamic connectivity structures and behavioral data representations jointly. The dynamic matrix decomposition simultaneously models the group information through the basis, and the subject-specific differences through the time-varying coefficients. The DTI Laplacians streamline this decomposition to focus on anatomically informed functional pathways. The LSTM-ANN directly models the temporal variation in the coefficients, with its weights encoding representations closely interlinked with behavior. The limited number of basis elements help provide compact representations explaining the connectivity information well. The regularization and constraints ensure that the problem is well posed, yet extracts clinically meaningful representations.

5.9 Clinical Evaluation

5.9.1 Subnetwork Identification

In this section, we investigate the subnetworks learned in the basis **B** by the sr-DDL model when trained on both datasets. Recall that each column of the basis consists of a set of co-activated AAL subregions. In order to robustly identify these patterns, we first train the model on 10 randomly sampled subsets of each dataset. Then, we match the obtained subnetworks based on their absolute cosine similarity. Since we have 15 subnetworks, we then illustrate the mean co-activations across the brain regions for each of them individually in Fig. 5.10 (HCP) and Fig. 5.11 (KKI). Here, the colorbar in the figure indicates subnetwork contribution to the AAL regions. Regions storing negative values (cold colors) are anticorrelated with regions storing positive ones (hot colors). Alongside, we represent the corresponding standard deviations across different regions for each of the 15 subnetworks.

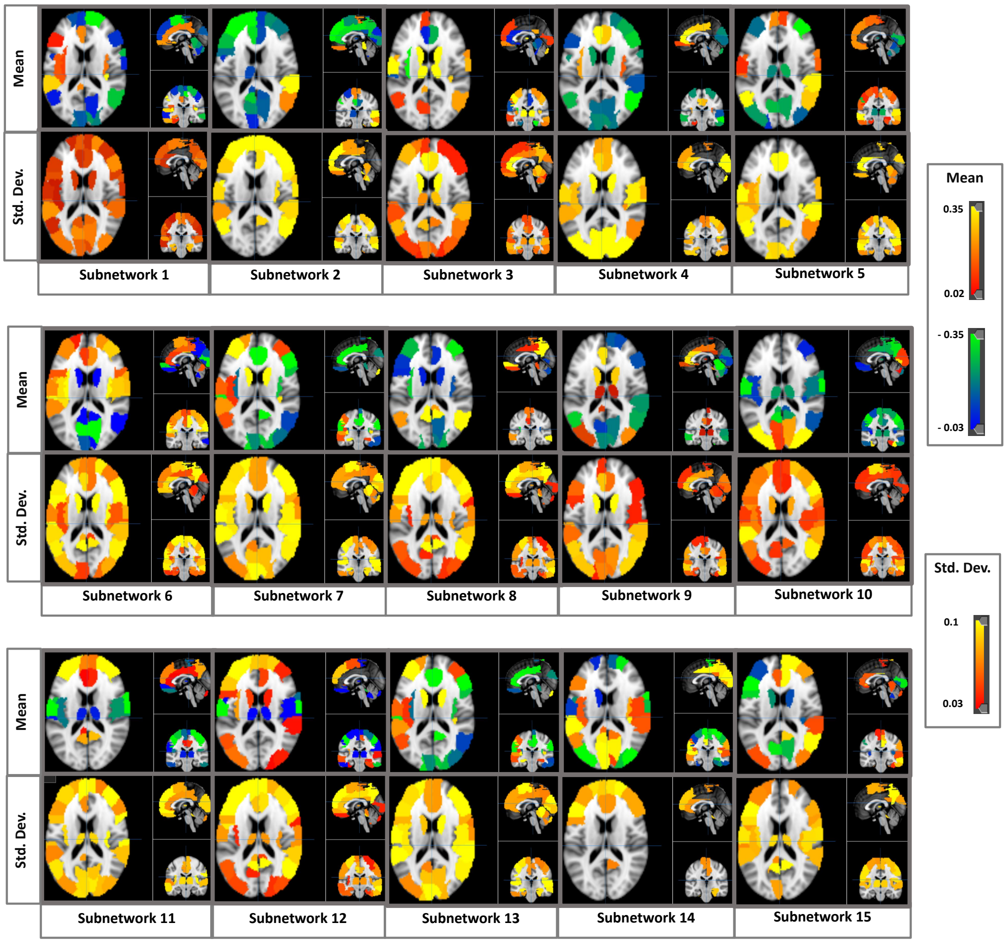


Figure 5.10: Complete set of subnetworks identified by the deep sr-DDL model for the HCP database. **Mean**: Mean regional co-activation patterns in basis **B** The red and orange regions are anti-correlated with the Purple and green regions. **Std. Dev.**: Standard deviations of regional co-activation patterns. A majority of regions exhibit small deviations from the mean. Both sets of plots have been computed across cross-validation folds

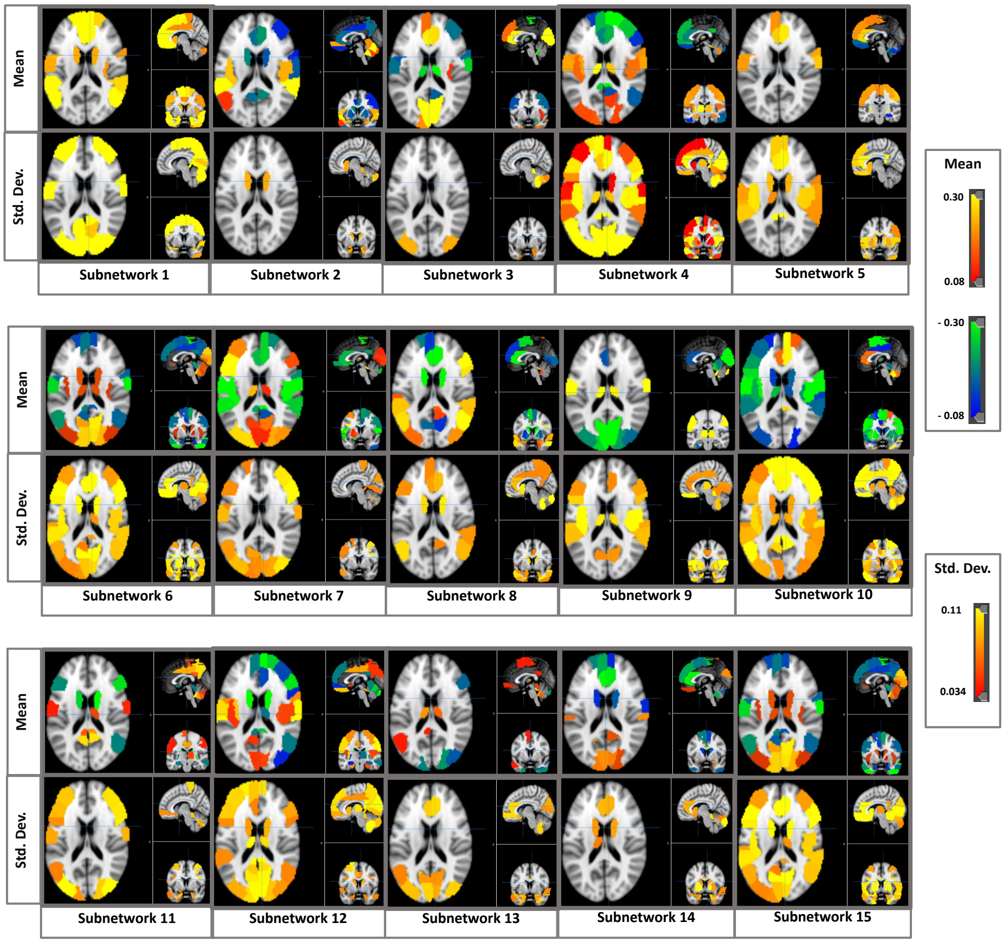


Figure 5.11: Complete set of subnetworks identified by the deep sr-DDL model for the KKI database. **Mean**: Mean regional co-activation patterns in basis **B** The red and orange regions are anti-correlated with the Purple and green regions. **Std. Dev.**: Standard deviations of regional co-activation patterns. A majority of regions exhibit small deviations from the mean. Both sets of plots have been computed across cross-validation folds

Examining the subnetworks in Fig. 5.10, we notice that Subnetworks 1 & 2, and 11 exhibits positive and competing contributions from regions of the Default Mode Network (DMN), which has been widely inferred in the resting state literature [35] and is believed to play a critical role in consolidating memory [183], as also in self-referencing and in the theory of mind [184]. At the same time, Subnetworks 2 and 11 have competing and positive contributions from regions in the Frontoparietal Network (FPN) respectively. The FPN is known to be involved in executive function and goal-oriented, cognitively demanding tasks [185]. Subnetworks 1, 6, 7, 11 and 13 are comprised of regions from the Medial Frontal Network (MFN). The MFN and FPN are known to play a key role in decision making, attention and working memory [186, 187], which are directly associated with cognitive intelligence. Subnetworks 1, 3, and 9 include contributions from the subcortical and cerebellar regions, while Subnetworks 10, 2, 14 and 11 include contributions from the Somatomotor Network (SMN). Taken together, these networks are believed to be important functional connectivity biomarkers of cognitive intelligence and consistently appear in previous literature on the HCP dataset [188, 189].

For the KKI dataset, in Fig. 5.11, Subnetwork 1 includes regions from the DMN, and the SMN. Similarly, Subnetwork 6 includes competing contributions from the SMN and DMN regions. Aberrant connectivity within the DMN and SMN regions have previously been reported in ASD [150, 98]. Subnetworks 7, 3, and 6 exhibit contributions from higher order visual processing areas in the occipital and temporal lobes along with competing sensorimotor regions. At the same time, Subnetwork 9 exhibits competing contributions from the visual network. These findings concur with behavioral reports of reduced visual-motor integration in autism [98]. Subnetworks 11 and 8 exhibit contributions from the central executive control network (CEN) and insula. Subnetwork 10 also exhibits anticorrelated CEN contributions. These regions are believed to be essential for switching between goal-directed and self-referential behavior [152]. Subnetwork 5 and Subnetwork 3 includes prefrontal and DMN regions, along with subcortical areas such as the thalamus, amygdala and hippocampus. The hippocampus is known to play a crucial role in the consolidation of long and short term memory, along with spatial memory to aid navigation. Altered memory functioning has been shown to manifest in children diagnosed with ASD [151]. The thalamus is responsible for relaying sensory and motor signals to the cerebral cortex in the brain and has been implicated in autism-associated sensory dysfunction, a core feature of ASD [190]. Along with the amygdala, which is known to be associated with emotional responses, these areas may be crucial for social-emotional regulation in ASD. [191].

Finally, we notice that the standard deviations for a majority of the regions in each of the subnetworks are small compared to the mean coactivation. Additionally, we observed an average similarity of 0.79 ± 0.13 and 0.81 ± 0.12 for these subnetworks across the runs on subsets of the HCP and KKI datasets respectively. These results suggests that our deep-generative framework is able to capture stable underlying mechanisms which robustly explain the different sets of deficits in ASD as well as robustly extract signatures of cognitive flexibility in neurotypical individuals.

5.9.2 Robustness of Biomarker Discovery

In this experiment, we study the overlap in the subnetworks in the basis $\mathbf{B}$ across different scales of subnetworks, i.e. varying the number of networks K. Recall from Section 5.6, that the knee point of the eigen-spectrum of $\{\mathbf{\Gamma}_n^t\}$

for both datasets is between $8 - 20$. Namely, we re-run the sr-DDL model on both the datasets steadily increasing the number of networks from $8 - 20$. In each case, we repeat the experiment using 10 random subsets of the data and look for subnetworks that appear most often. Fig. 5.10 and Fig. 5.11 illustrate the top ten networks that appear most frequently across different data subsets and choice of K for the HCP dataset and KKI dataset respectively. Alongside, we also report the mean and standard deviation of the absolute cosine similarity (S) for each individual subnetworks across the multiple runs. Networks which are most consistent exhibit higher similarity across runs with group 1 being the top five subnetworks (S $\geq$ 0.95), group 2 being the next five subnetworks ($S > 0.85$). Finally, a visual inspection and comparison with our results in Section 5.9.1 suggest a considerable overlap between the subnetworks in Fig. 5.10 and Fig 5.12 for the HCP dataset and between Fig. 5.11 and Fig 5.13 for the KKI dataset. These results suggest that our Deep sr-DDL robustly extracts representative neural signatures indicative of behavior in

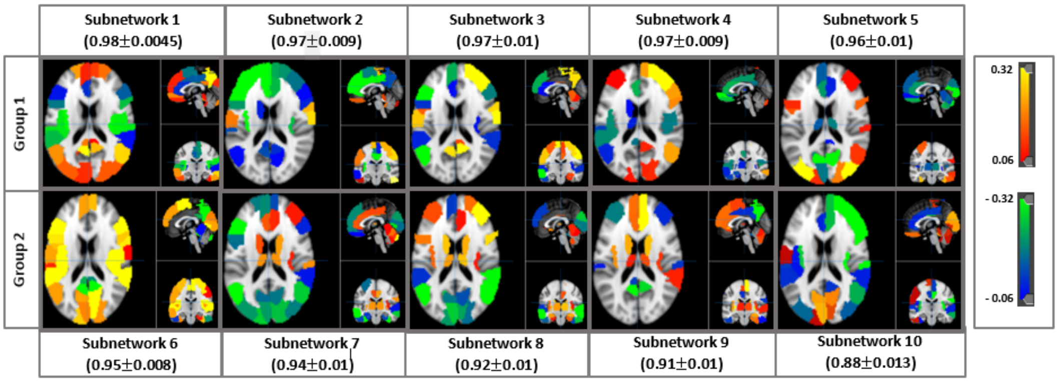


Figure 5.12: HCP dataset: Set of top 10 consistent subnetworks across different model orders. Subnetworks in group 1 exhibit above 0.95 average similarity across data subsets and model orders. Subnetworks in group 2 exhibit between $0.85 - 0.95$ average similarity across data subsets and model orders.

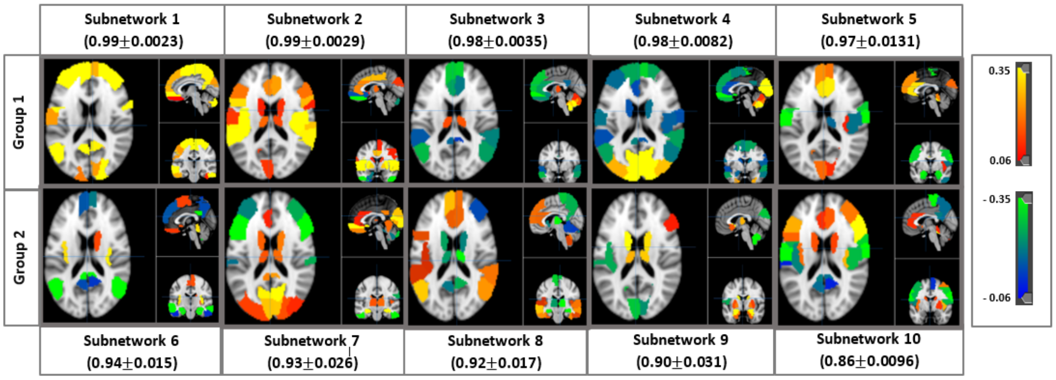


Figure 5.13: KKI dataset: Set of top 10 consistent subnetworks across different model orders. Subnetworks in group 1 exhibit above 0.95 average similarity across data subsets and model orders. Subnetworks in group 2 exhibit between $0.85 - 0.95$ average similarity across data subsets and model orders.

both healthy and autistic populations.

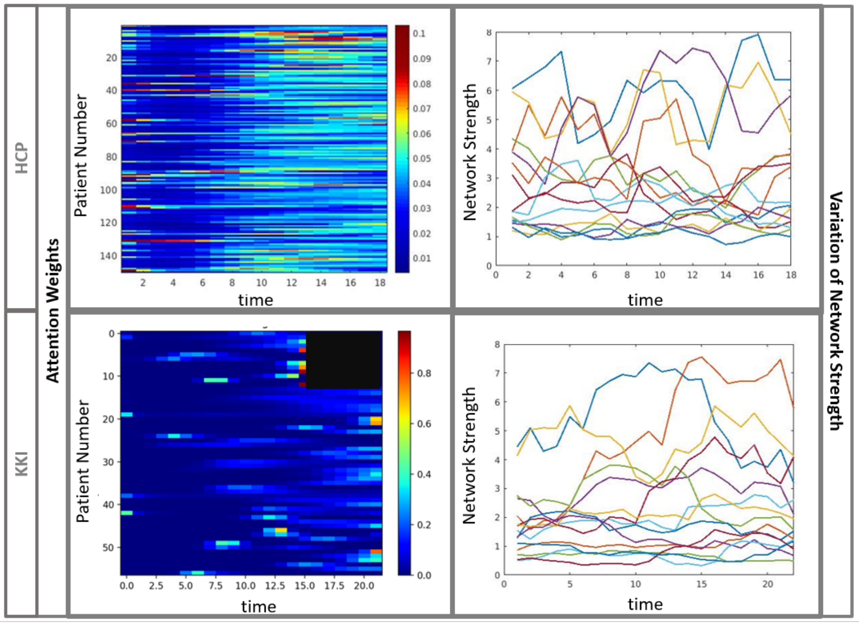


Figure 5.14: (Left) Learned attention weights **(Right)** Variation of network strength over time on the **(Top)** HCP dataset **(Bottom)** KKI dataset

5.9.3 Uncovering rs-fMRI Network Dynamics

Our deep sr-DDL allows us to map the evolution of functional networks in the brain by probing the LSTM-ANN representation. Recall that our model does not require the rs-fMRI scans to be of equal length. Fig. 5.14 (left) illustrates the learned attentions output by the A-ANN for the subjects from the HCP dataset on the top and the KKI subjects at the bottom during testing. For the KKI dataset, the patients with shorter scans have been grouped in the top of the figure. These time-points have been blackened at the beginning of the scan. The colorbar indicates the strength of the attention weights. Higher attention weights denote intervals of the scan considered especially relevant for prediction. Notice that the network highlights the start of the scan for several individuals, while it prefers focusing on the end of the scan for some others, especially pronounced in case of the KKI dataset. The patterns are comparatively more diffused for subjects in the HCP dataset, although several subjects manifest selectivity in terms of relevant attention weights. This is indicative of the underlying individual-level heterogeneity in both the cohorts.

Next, we illustrate the variation of the network strength for a representative subject from the HCP dataset and KKI dataset over the scan duration in Fig. 5.14 (right) at the top and bottom respectively. Each solid colored line corresponds to one of the 15 sub-networks in Fig. 5.11. Notice that, over the scan duration, each network cycles through phases of activity and relative inactivity. Consequently, only a few networks at each time step contribute to the patient's dynamic connectivity profile. This parallels the transient brain-states hypothesis in dynamic rs-fMRI connectivity [192], with active states as corresponding sub-networks in the basis matrix **B**.

5.10 Discussion

Our deep-generative hybrid cleverly exploits the intrinsic structure of the rs-fMRI correlation matrices through the dynamic dictionary representation to simultaneously capture group-level and subject-specific information. At the same time, the LSTM-ANN network models the temporal evolution of the rs-fMRI data to predict behavior. The compactness of our representation serves as a dimensionality reduction step that is related to the clinical score of interest, unlike the pipelined treatment commonly found in the literature. Our structural regularization helps us fold in anatomical information to guide the functional decomposition. Overall, our framework outperforms a variety of state-of-the-art graph theoretic, statistical and deep learning baselines on two separate real world datasets.

We conjecture that the baseline techniques fail to extract representative patterns from structural and functional data. These techniques are quite successful at modelling group level information, but fail to generalize to the entire spectrum of cognitive, symptomatic or connectivity level differences among subjects. Consequently, they overfit the training data.

5.10.1 Examining Generalizability

Notice that the training examples (red points) in Figs. 5.8 and 5.9 follow the $\mathbf{x} = \mathbf{y}$ line perfectly, which may suggest overfitting. This phenomenon can be explained by the difference between our training procedure, where we

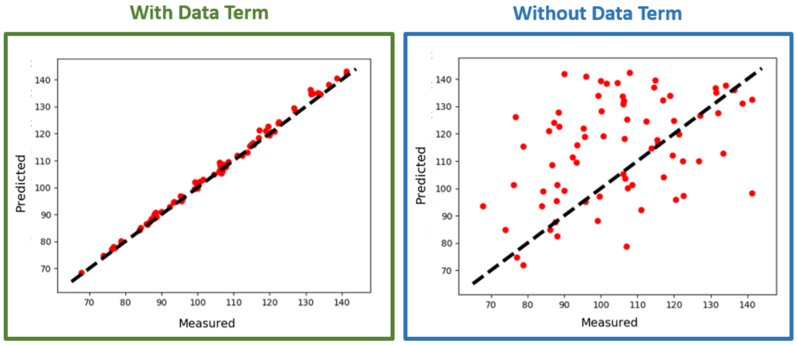


Figure 5.15: Prediction Performance of the Deep sr-DDL for the CFIS score on training data when **(L)** The data term is included in computing $\{\mathbf{c}_n^t\}$ **(R)** The data term is excluded from the computation of $\{\mathbf{c}_n^t\}$

optimize our joint objective in Eq. (5.9) assuming the scores are known, and our testing procedure. Recall that Section 5.5.5 describes the procedure for calculating the temporal sr-DDL loadings for an unseen patient i.e. $\bar{\mathbf{c}}_n^t$ from the basis $\mathbf{B}^*$ obtained during training. Since the subject is not a part of the training set, the corresponding value of $\hat{\mathbf{y}}$ is unknown. Effectively, we must set the contribution from the data term, i.e., the deep network loss $\mathcal{L}(\cdot)$ in Eq. (5.9) to 0. Here, we examine the effect of employing the same strategy to

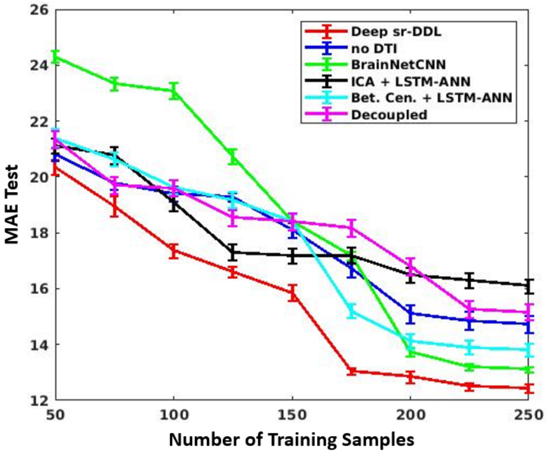


Figure 5.16: Median Absolute Error on the Test Set varying the number of samples used for training. The vertical bars indicate standard errors for each setting

calculate the coefficients for the training patients. In essence, we estimate the corresponding severity $\hat{\mathbf{Y}}$ now excluding the deep network loss. Accordingly, Fig. 5.15 highlights the differences in training fit with and without this term included in estimating $\{\mathbf{c}_n^t\}$ for the HCP dataset. Notice that in the latter, the training accuracy for the CFIS score has the same distribution as the testing points in Fig. 5.8. In contrast, inclusion of the deep network loss in our coupled optimization overparamterizes the search space of solutions for $\{\mathbf{c}_n^t\}$ to yield a near perfect fit.

To further probe the generalization capabilities of our Deep sr-DDL, we examine the effect of training the models on different sized datasets. For this experiment, we first set aside 50 individuals from the HCP database as a test set on which we evaluate the generalization performance. We then sweep the training set size from $N = 50 - 200$ in increments of 25 subjects. To avoid biasing the results, none of these subjects overlap with the HCP-2 validation set used for parameter tuning in Section 5.6. For each training set size, we randomly sample the subjects 10 times and compute the generalization performance on the held-out set.

Fig. 5.16 displays the MAE of the CFIS score prediction on the test set as a function of the training set size. As expected, we observe that with increasing training data, the performance on the test set improves at first but eventually saturates for all methods. This is evinced by a lowering of the MAE in the initial parts of the curve followed by a subsequent plateau at roughly $150 - 200$ samples. Based on these results, we conjecture that further addition of training data does not substantially improve the generalization

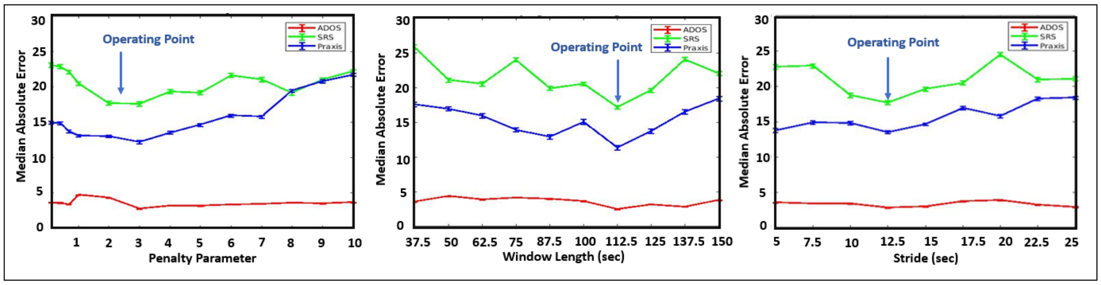


Figure 5.17: Performance of the Deep sr-DDL upon varying **(L):** the penalty parameter λ **(B):** window length **(R):** stride. Our operating point is indicated by the blue arrow

capabilities of our model or the baselines. We also note that the deep sr-DDL outperforms the baselines across the entire regime. In conjunction with our results from Section 5.8, we conclude that the deep sr-DDL model performs reasonably well for small to moderately sized datasets. This is especially important against the backdrop of potential clinical applications, many of which have datasets of modest sizes.

5.10.2 Assessing Model Robustness

Our deep sr-DDL framework has only two free hyperparameters. The first is the number of subnetworks in **B**. As described in Section 5.6, we use the eigen-spectrum of $\{\mathbf{\Gamma}_n^t\}$ to fix this at 15 for both datasets. The second is the penalty parameter λ, which controls the trade-off between representation and prediction. Recall that our data pre-processing includes a sliding window protocol in Fig. 2.3, which is defined by two parameters, i.e. the sliding window length and the stride. From a mathematical perspective, our deep sr-DDL formulation as such is agnostic to these parameters, as they are simply folded into the input data dimension. However, empirically, they balance the context size and information overlap within the rs-fMRI correlation matrices

$\{\mathbf{\Gamma}_n^t\}$ and affects the prediction performance.

In this section, we evaluate the performance of our framework under three scenarios. Specifically, we sweep λ, the window length and the stride parameter independently, keeping the other two values fixed. We use five fold cross validation with the MAE metric to quantify the multi-score prediction performance, which as shown in Section 5.8, is more challenging than single score prediction. Fig. 5.17 plots the performance for the three scores on the KKI dataset with MAE value for each score on the **y** axis and the parameter value on the **x** axis.

We observed that our method gives stable performance for fairly large ranges of each parameter settings. As expected, low values of λ $(0.01 - 1)$ result in higher MAE values, likely due to underfitting. Similarly, higher values (> 6) result in overfitting to the training dataset, degrading the generalization performance. Additionally, lower values of window lengths result in higher variance among the correlation values due to noise, and hence less reliable estimates of dynamic connectivity [106]. On the other hand, very large context windows tend to miss nuances in the dynamic evolution of the scan. Empirically, we observe that a mid-range of window length $100 - 125$s yields a good tradeoff between representation and prediction. The training of LSTM networks with very long sequence lengths is known to be particularly challenging owing to vanishing/exploding gradient issues during backpropagation. However, having too short a sequence confounds a reliable estimation of the LSTM weights from limited data. The stride parameter helps mitigate these issue by compactly summarizing the information in the sequence while simultaneously controlling the overlap across subsequent samples. Our experiments found a stride length between $10 - 20$s to be suitable for our application.

In summary, the guidelines we identified for each of the parameters are- $\lambda \in (2 - 5)$, window length $\in (100 - 125)$s, and stride $\in (10 - 20)$s. Additionally, our experiments on the HCP dataset using the same settings indicate that the results of our method are reproducible across different populations. It is also interesting to note that previous experiments on the HCP dataset in literature have found similar window lengths to be stable in classification [193] and various test-retest settings [194].

5.10.3 Clinical Relevance

Our experiments on the KKI dataset evaluate the ability of our Deep sr-DDL framework to simultaneously explain multiple clinical impairments of ASD. This multi-target prediction is a challenging task, and in fact, the baseline methods fail to generalize all three scores. At the same time, one could evaluate the performance of predicting each score independently via three single-target regression tasks. Accordingly, Table 5.3 compares the performance of our Deep sr-DDL framework in the single-target and multi-target settings. Empirically, we observe that the single-target prediction is slightly better than the multi-target prediction. Indeed, a possible counter perspective would be to optimize for prediction accuracy of individual measures explained by potentially different brain bases, for example, as in the work in [146, 39, 157]. This comparison poses a more philosophical question about the benefits of a

Score	Method	MAE	NMI
ADOS	Single-target	2.91 ± 2.71	0.44
	Multi-target	2.99 ± 1.99	0.37
SRS	Single-target	14.78 ± 14.24	0.87
	Multi-target	18.70 ± 13.51	0.85
Praxis	Single-target	12.40 ± 11.60	0.85
	Multi-target	14.99 ± 10.17	0.82

Table 5.3: Testing performance (5-fold CV) of the sr-DDL framework for single-target and multi-target prediction on the KKI dataset according to **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)**. We also report the standard deviation for the MAE. Lower MAE and higher NMI scores indicate better performance.

multi-target setup given a possible decline in predictive performance and the difficultly of the task itself.

To weigh in on this trade off, we note the growing consensus in clinical psychiatry that complex disorders, such as autism and schizophrenia, are inherently multidimensional [59]. Furthermore, there is considerable patient heterogeneity within a single diagnostic umbrella that reflect subtle differences in the underlying etiology [60]. In fact, the National Institute of Mental Health (NIHM) in the United States has released the RDoc research framework [61], which advocates for a multidimensional characterization to understand the full spectrum of mental health and illness. In this context, our Deep sr-DDL approach provides a flexible tool to map multiple measures via a consistent and stable brain basis (as shown by the results in Section 5.9.1). Thus, we view it as an important foundation to parse complex spectrum disorders that may even spur new analytical directions in brain connectomics.

Finally, our Deep sr-DDL framework is carefully designed to extract subject-level dynamic information. Namely, the attention mechanism automatically highlights portions of the rs-fMRI scan that are important for clinical prediction (Fig 5.14). In fact, a comparison of the attention weights in Fig. 5.14 suggests considerable inter-patient variability of the intervals used for multi-target prediction in the KKI dataset, as opposed to the relatively consistent attention weights in the HCP dataset. This pattern may be linked to the heterogeneity of ASD described above. In conjunction, we observe the subnetwork contributions phasing in and out prominence over the course of the scan, which is consistent with the transient brain state hypothesis [192]

In summary, the blend of classical generative modeling and deep learning prediction in our Deep sr-DDL framework allows for a finer-grained characterization of connectivity and behavior. Overall, we believe that the robustness, stability, clinical interpretability, and flexibility of our Deep sr-DDL render it a novel and valuable tool for the research community.

5.10.4 Applications, Limitations and Future Scope

As seen in our experiments in Section 5.8, our method is able to extract key predictive resting state biomarkers from healthy and autistic populations. Additionally, our deep sr-DDL makes minimal assumptions. Provided we have access to a set of consistently defined structural and functional connectivity measures and clinical scores, this analysis can be easily adapted to other neurological disorders and even predictive network models outside the medical realm. Overall, these findings broaden the scope of our method for future applications.

Although we outperform several baselines on two separate datasets, our prediction performance in Section 5.8 is far from perfect. This underscores that multi-score prediction is a challenging clinical problem. One of the key reasons can be attributed to inherent noise in the clinical measures themselves. For example, SRS is based on a parent-teacher questionnaire, which tends to be more subjective than a clinical exam. This renders the behavioral prediction task especially challenging, which partially accounts for the poor performance of several baselines we compared against. Keeping this in mind, a natural clinical direction of exploration is to adopt our method to predicting measures more directly related to functional connectivity, as opposed to those relying on clinical reports. Another avenue of exploration includes examining more coarse indicators of behavior, such as ordered levels of impairment instead of continuous measures (an ordinal regression problem), or the prevalence of ASD sub-types.

Another limitation to our method lies in the fact that our estimate of dynamic functional connectivity relies on the availability of a reliable sliding-window protocol. As illustrated in Section 5.10.2, an inappropriate window-length and stride choice has a direct bearing on the predictive performance. Moreover, this tradeoff is difficult to quantify and correct for analytically. Keeping this in mind, we are motivated to explore alternatives to the sliding window for better estimating dynamic functional connectivity, which can at the same time be robustly integrated into multimodal data-analysis frameworks such as ours.

From the methodological standpoint, we recognize that our model is simplistic in its assumptions, particularly in the sr-DDL formulation. The DTI priors guide a data-driven classical rs-fMRI matrix decomposition in a regularization framework. This modeling choice was deliberately employed to preserve interpretability in the basis and simplify the inference procedure. A key limitation of this approach is that it does not directly consider multi-stage pathways, which may be an important mediator of functional relationships between communicating sub-regions.

To this end, graph neural networks have shown great promise in brain connectivity research due to their ability to capture subtle and multi-stage interactions between communicating brain regions while exploiting the underlying hierarchy of brain organization. Consequently, these methods are emerging as important tools to probe complex pathologies in brain functioning and diagnose neurodevelopmental disorders [195, 196]. In this vein, the Appendix (Chapter 8) presents an exploratory end-to-end graph convolutional network that jointly models rs-fMRI and DTI data.

Chapter 6

Beyond Dictionary Decompositions and Phenotypic Prediction: Geometric Frameworks to Characterize Complementary Connectivity Spaces

In this chapter, we examine a slightly different problem that goes beyond just multidimensional phenotypic prediction. Instead, we are interested in studying the interplay between function and structure more carefully. A natural direction to pursue is that of multi-view representation learning of the two connectivity spaces. In this light, we examine the prediction of structural connectivity from functional connectivity. This helps us better qualify the complementarity between the two data spaces.

In the clinical neuroscience realm, several studies have found both direct and indirect correspondences between structural and functional connectivity [11, 12]. Going a step further, structural and functional connectivity have been shown to be predictive of each other at varying scales [15, 16, 17]. As a result, multimodal integration of these viewpoints has become a key area of focus for characterizing neuropsychiatric disorders such as autism and ADHD [197, 198]

As described in Chapter 2, techniques for integrating structural and functional connectivity focus heavily on group-wise discrimination. These works include statistical tests on edge-based features to identify significant differences in Alzheimer's disease [23], parallel ICA using structure and function to identify discriminative biomarkers of schizophrenia [128], and classical machine learning techniques to predict diagnosis [199]. While highly informative at a group level, these methods do not directly address inter-individual variability, for example by predicting finer grained patient characteristics. This divide has been partially bridged by end-to-end deep learning models. Examples include MLPs [200] for age prediction from functional connectomes and convolutional neural networks [44] for predicting cognitive and motor measures from structural connectomes. Even so, these models focus exclusively on a single neuroimaging modality and do not exploit the interplay between function and structure.

Geometric learning frameworks have recently shown great promise in multimodal connectomics studies, both for conventional manifold learning [201] and in the context of Graph Convolutional Networks (GCN) [202, 198]. Their primary advantage is the ability to directly incorporate and exploit the underlying data geometry. Beyond associative analyses, the work of [17, 203] employ multi-GCNs combined with a Generative Adversarial Network (GAN) for the alignment problem. Particularly, [17] examines the problem of recovering structural connectomes from patient functional connectomes While this paper marks a seminal contribution to multimodal integration, the representations learned by end-to-end GCNs can be hard to interpret. It can also be difficult to train GANs on modest-sized datasets [204].

Here, we develop an end-to-end matrix autoencoder that maps rs-fMRI correlation matrices to structural connectomes obtained from DTI tractography. Inspired by recent work in Riemannian deep learning [205, 206], our matrix autoencoder, estimates a low dimensional embedding from rs-fMRI correlation matrices while taking into account the geometry of the functional connectivity (FC) manifold. Our second matrix decoder uses this embedding to reconstruct patient structural connectivity (SC) matrices akin to a manifold alignment [207] between the FC and SC data spaces. For regularization, the FC embedding is also used to predict behavioral phenotypes. We demonstrate that our framework reliably traverses from function to structure and extracts meaningful brain biomarkers.

Outline: This section is organized as follows: Section 6.1 describes our Matrix Autoencoder framework to encode functional connectivity matrices, while Section 6.1.2 explains the alignment to structural connectivity and Section 6.1.3 describes our secondary phenotypic prediction task. Section 6.2 describes our experimental evaluation. This work appeared recently in MICCAI 2021 as a conference paper [208]. Finally, we use Section 6.4 to probe the theoretical aspects of the representation learned by our proposed autoencoder.

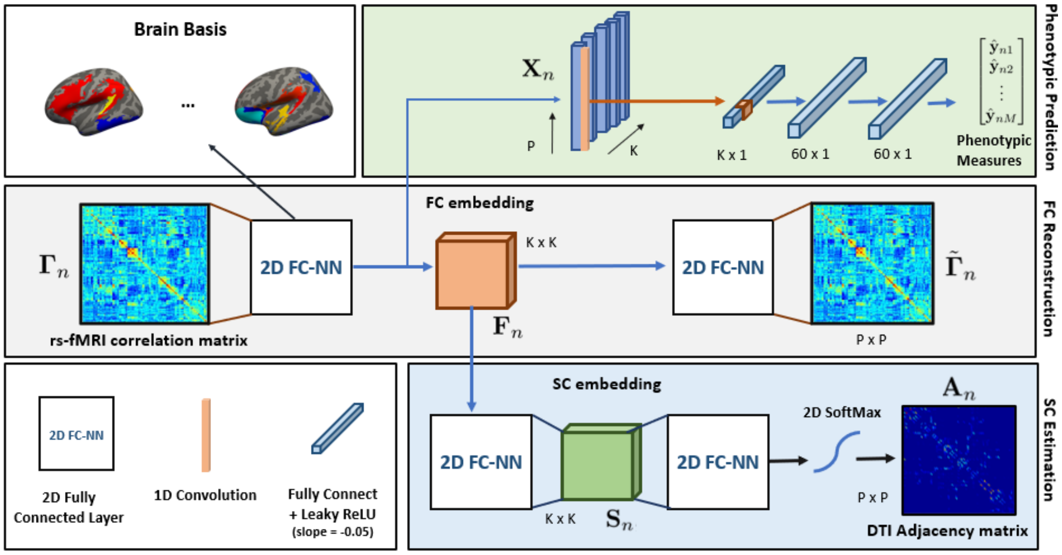


Figure 6.1: A Matrix Autoencoder for aligning the FC and SC manifolds **Gray Box:** Matrix encoder-decoder for functional connectomes. **Blue Box:** Alignment Decoder for estimating DTI connectomes **Green Box:** ANN for predicting behavioral phenotypes

6.1 A Matrix Autoencoder to Model the Functional Connectivity Space

Fig. 6.1 illustrates our matrix autoencoder framework consisting of an encoder-decoder for functional connectivity (gray box), manifold alignment for estimating structural connectivity (blue box), and ANN for prediction of behavioral phenotypes (green box). Let N be the number of patients and P be the number of ROIs in our brain parcellation. We denote the rs-fMRI correlation matrix for patient n by $\mathbf{\Gamma}_n \in \mathcal{R}^{P \times P}$. $\mathbf{A}_n \in \mathcal{R}^{P \times P}$ is the corresponding structural connectivity profile, and $\mathbf{y}_n \in \mathcal{R}^{M \times 1}$ is a vector of M concatenated phenotypic measures.

6.1.1 Functional Connectivity Reconstruction:

By construction, the correlation matrices $\mathbf{\Gamma}_n$ belong to the manifold of symmetric positive semi-definite matrices $\mathcal{P}_P^+$. Our matrix autoencoder estimates a latent functional embedding $\mathbf{F}_n \in \mathcal{P}_K^+$ using a 2D fully connected (2D FC-NN) layer [206, 205]. Formally, this mapping $\mathbf{\Phi}_{\text{ec}}(\cdot) : \mathcal{P}_P^+ \to \mathcal{P}_K^+$ is parametrized by weights $\mathbf{W} \in \mathcal{R}^{P \times K}$ and is computed as a cascade of two linear layers with tied weights: $\mathbf{F}_n = \mathbf{\Phi}_{\text{ec}}(\mathbf{\Gamma}_n) = \mathbf{W}^T \mathbf{\Gamma}_n \mathbf{W}$. Our decoder is another 2D FC-NN that estimates $\tilde{\mathbf{\Gamma}}_n$ from $\mathbf{F}_n$ via a similar transformation $\mathbf{\Phi}_{\text{dc}}(\cdot) : \mathcal{P}_K^+ \to \mathcal{P}_P^+$ that shares weights with the encoder. Mathematically, our FC reconstruction loss is represented as follows:

$$\mathcal{L}_{\text{FC}} = \frac{1}{N} \sum_n ||\mathbf{\Phi}_{\text{dc}}(\mathbf{\Phi}_{\text{ec}}(\mathbf{\Gamma}_n)) - \mathbf{\Gamma}_n||_F^2 = \frac{1}{N} \sum_n ||\mathbf{W}\mathbf{W}^T \mathbf{\Gamma}_n \mathbf{W}\mathbf{W}^T - \mathbf{\Gamma}_n||_F^2 \quad (6.1)$$

The second term of Eq. (6.1) encourages the columns of the brain basis $\mathbf{W}$ to be orthonormal. Conceptually, this specialized loss helps us learn uncorrelated patterns that explain the rs-fMRI data well while acting as an implicit regularizer.

6.1.2 Aligning to Structural Connectivity Prediction

The structural connectivity matrices $\mathbf{A}_n$ are derived from DTI tractography and belong to the manifold of symmetric (non PSD) matrices $\mathcal{S}_P$. Our alignment decoder first generates an SC embedding $\mathbf{S}_n \in \mathcal{R}^{K \times K}$ from $\mathbf{F}_n$ via a 2D FC-NN layer $\mathbf{\Phi}_{\text{align}}(\cdot) : \mathcal{P}_K^+ \to \mathcal{P}_K^+$, followed by a second 2D FC-NN layer $\mathbf{\Phi}_{\text{est}}(\cdot) : \mathcal{P}_K^+ \to \mathcal{P}_P^+$ which maps to the structural connectivity matrices. For stability our SC matrices do not have self-connections and are normalized to

$\|\mathbf{A}_n\|_1 = 1$. Accordingly, at the output layer, we suppress the diagonal elements and apply a 2D softmax $\mathcal{SF}(\cdot)$ to generate the final output $\tilde{\mathbf{A}}_n \in \mathcal{R}^{P \times P}$. Our SC estimation objective is represented as follows:

$$\mathcal{L}_{\text{SC}} = \frac{1}{N} \sum_n \left|\left|\mathcal{SF}\left[\mathbf{\Phi}_{\text{est}}(\mathbf{\Phi}_{\text{align}}(\mathbf{F}_n)) \circ [\mathbf{1}\mathbf{1}^T - \mathcal{I}_P]\right] - \mathbf{A}_n\right|\right|_F^2 \tag{6.2}$$

where $\circ$ is the element-wise Hadamard product. $\mathbf{1} \in \mathcal{R}^{P \times 1}$ is the vector of all ones, and $\mathcal{I}_P$ is the identity matrix of dimension P. Conceptually, the loss in Eq. (6.2) is akin to manifold alignment [207] between the functional and structural embeddings based on a two sided Procrustes-like objective.

6.1.3 Mapping to Phenotypes

We map the intermediate representation $\mathbf{X}_n = \mathbf{\Gamma}_n\mathbf{W} \in \mathcal{R}^{P \times K}$ learned by the FC encoder to the phenotypes $\mathbf{y}_n$ via a cascade of a 1D convolutional layer and an ANN. The convolutional layer $\mathcal{F}_{\text{conv}}(\cdot)$ collapses $\mathbf{X}_n$ along its rows via a weighted sum to generate a K dimensional feature vector. This feature vector is input to a simple two layered ANN $\mathcal{G}(\cdot)$ to jointly estimate the elements in $\hat{\mathbf{y}}_n$. We use a Mean Squared Error (MSE) loss function:

$$\mathcal{L}_{\text{phen}} = \frac{1}{NM} \sum_n ||\hat{\mathbf{y}}_n - \mathbf{y}_n||_F^2 = \frac{1}{NM} \sum_n ||\mathcal{G}(\mathcal{F}_{\text{conv}}(\mathbf{X}_n)) - \mathbf{y}_n||_2^2 \tag{6.3}$$

This prediction task is a secondary regularizer that encourages our matrix autoencoder to learn representations predictive of inter-subject variability.

6.2 Evaluation on Real Data

We validate our framework on a dataset of 275 healthy individuals from the Human Connectome Project database and on the second clinical ASD dataset described in Chapter 2. We wish to evaluate the model on the robustness of recovery of structural connectivity patterns across individuals, reliable phenotypic prediction against a variety of baselines, and on the extraction of predictive and interpretable brain biomarkers. We adopt a five fold cross-validated setting.

6.2.1 Implementation Details

We train our framework on a joint objective that combines Eqs. (6.1), (6.2) and (6.3) as follows:

$$\mathcal{L} = \mathcal{L}_{\text{FC}} + \gamma_1 \mathcal{L}_{\text{SC}} + \gamma_2 \mathcal{L}_{\text{phen}} \tag{6.4}$$

where γ_1 and γ_2 balance the tradeoff for the SC estimation and phenotypic prediction relative to the FC reconstruction objective. We employ a two layered ANN with the hidden layer size $V = 60$ with Leaky ReLU ($\phi(x) = \max(0, x) + 0.05 * \min(0, x)$) as the activation function. We employ an ADAM optimizer [161] with learning rate 0.005 and weight decay regularization [209] ($\delta = 0.0005$) run for a maximum of 400 epochs. Optimization parameters were fixed based on a validation set consisting of 30 additional patients from the HCP database. We use this strategy to set the the dimensionality of our autoencoder embedding at $K = 15$ and loss penalities to $\{\gamma_1, \gamma_2\} = 10^3, 3$. Finally, we utilize a spectral initialization for the encoder-decoder weights **W** in Eq. (6.1) based on the top K eigenvectors of the average patient correlation matrix $\bar{\mathbf{\Gamma}}_n$ for the training set. We use a similar initialization based on $\bar{\mathbf{A}}_n$ for $\mathbf{\Phi}_{\text{est}}(\cdot)$, and default initialization [210] for the remaining layers. Our model has a runtime of 10-12 minutes on an 8 core machine with 32GB RAM implemented in PyTorch (v1.5.1).

6.2.2 Evaluating Phenotypic Prediction

6.2.2.1 Baselines

We compare against the following baselines on the phenotypic prediction task:

Matrix AE without rs-fMRI decoder: We start with the architecture in Fig. 6.1 but omit the rs-fMRI decoder loss ($\mathcal{L}_{\text{FC}}$) in Eq. (6.4). This helps us evaluate the benefit of a tied encoder-decoder model for the rs-fMRI matrices.

Matrix AE without DTI decoder: We start with the architecture in Fig. 6.1 but remove the DTI decoder loss ($\mathcal{L}_{\text{SC}}$) in Eq. (6.4). This helps us evaluate the benefit of manifold alignment to constrain the functional embedding.

Decoupled Matrix AE + ANN: We start with the architecture in Fig. 6.1 but decouple the representation learning on the connectomics data from the prediction of phenotypic measures by training the models separately. This baseline provides a comparison against allowing the two competing objectives to guide each other directly during training.

BrainNetCNN: This baseline integrates multimodal connectivity data via the BrainNetCNN [44]. We modify the original architecture, which is designed for a single modality, to have two branches, one for the rs-fMRI correlation

Table 6.1: CFIS prediction on the HCP dataset against the baselines using Median Absolute Error (MAE), Normalized Mutual Information (NMI) for training and testing for the test set. Best performance is highlighted in bold.

Method	MAE Train	MAE Test	NMI Train	NMI Test
No rs-fMRI dec.	6.31 ± 5.61	16.42 ± 12.41	0.85	0.61
No DTI dec.	6.30 ± 5.80	15.44 ± 13.00	0.86	0.61
Decoupled.	2.53 ± 2.41	14.90 ± 13.60	0.87	0.59
BrainNetCNN	6.80 ± 6.25	14.95 ± 12.74	0.88	0.59
Dict. Learn. + ANN	**3.19** ± **2.19**	15.26 ± 13.99	**0.89**	0.66
Our Framework	3.19 ± 2.47	**14.08** ± **11.85**	0.86	**0.69**

matrices $\mathbf{\Gamma}_n$, and another for the DTI connectomes $\mathbf{A}_n$. The ANN is modified to output M measures of clinical severity. We set the hyperparameters according to [44]

rs-fMRI Dictionary Learning + ANN: The framework in [160] uses rs-fMRI correlation matrices for the prediction of multiple clinical measures. The model combines a dictionary learning with a neural network predictor, with these two blocks optimized in an end-to-end fashion via a coupled optimization objective.

6.2.2.2 Predicting Behavioral Phenotypes:

Table 6.1 (and Fig. 6.2) compares the model against the baselines when predicting CFIS in a five-fold cross validated setting. Lower Median Absolute Error (MAE) and higher Normalized Mutual Information (NMI) signify improved performance. Our framework outperforms the baselines during testing, though the model of [160] comes in a close second. This suggests that the Matrix Autoencoder faithfully models subject-specific variation even in unseen patients.

Table 6.2: Multi-score performance on the ASD dataset using Median Absolute Error (MAE), Normalized Mutual Information (NMI), and Correlation Coefficient (R) for testing. Best performance is highlighted in bold. Near misses are underlined

Measure	Method	MAE Test	NMI Test
ADOS	No rs-fMRI dec.	3.11 ± 2.74	0.46
	No DTI dec.	**2.61** ± **2.59**	0.41
	Decoupled	**2.64** ± **2.30**	**0.49**
	BrainNetCNN	3.89 ± 2.80	0.35
	Dict. Learn.+ANN	2.71 ± 2.40	0.43
	Our Framework	2.71 ± 1.84	**0.49**
SRS	No rs-fMRI dec.	16.84 ± 16.01	0.77
	No DTI dec.	**15.65** ± **12.69**	0.81
	Decoupled	17.40 ± 14.16	0.74
	BrainNetCNN	17.50 ± 15.18	0.73
	Dict. Learn.+ANN	16.79 ± 13.83	**0.89**
	Our Framework	16.04 ± 13.40	0.83
Praxis	No rs-fMRI dec.	14.03 ± 10.80	0.74
	No DTI dec.	19.65 ± 13.18	0.81
	Decoupled	17.08 ± 12.23	0.76
	BrainNetCNN	19.35 ± 12.56	0.74
	Dict. Learn.+ANN	**13.19** ± **10.75**	0.82
	Our Framework	**13.14** ± **10.78**	**0.86**

For the evaluation on the Autism dataset, we carry forward the same model parameters as used for the HCP dataset. Table 6.2 (and Fig. 6.3) compares the *multi-score* prediction testing performance of ADOS, SRS, and Praxis in a five fold cross validation setting. We observe that only our framework and the model of [160] can *simultaneously predict all three measures*. In contrast, the other baselines achieve good testing performance on one or two of the measures (for example, No DTI decoder baseline for ADOS and SRS) but cannot generalize all three. Overall, our experiments on both healthy (HCP) and clinical (ASD) populations suggest that our model is robust across cohorts and generalizes effectively even with modest dataset sizes.

6.2.3 Functional to Structural Association

We evaluate three aspects of our functional to structural manifold alignment. First is our ability to recover structural connectivity matrices during testing. Here, we compare two distance metrics: (1) F_{self} is the Frobenius norm between a test example $\mathbf{A}_n$ and the model prediction for the same example $\hat{\mathbf{A}}_n$, and (2) F_{other} is $\hat{\mathbf{A}}_n$ and other SC matrices $\mathbf{A}_m, (m \neq n)$. As shown in the left of Fig. 6.4(a) (HCP) and Fig. 6.6 (a) (KKI) , F_{self} is consistently smaller than F_{other}, with statistical significance determined using the Wilcoxon rank sum test. This indicates that individual differences in SC are preserved by our framework. In the same plot, we also benchmark the recovery performance of our framework against a baseline Matrix encoder-decoder (gray box in Fig. 6.1) with the SC matrices as *input and output*. We also compare against a linear

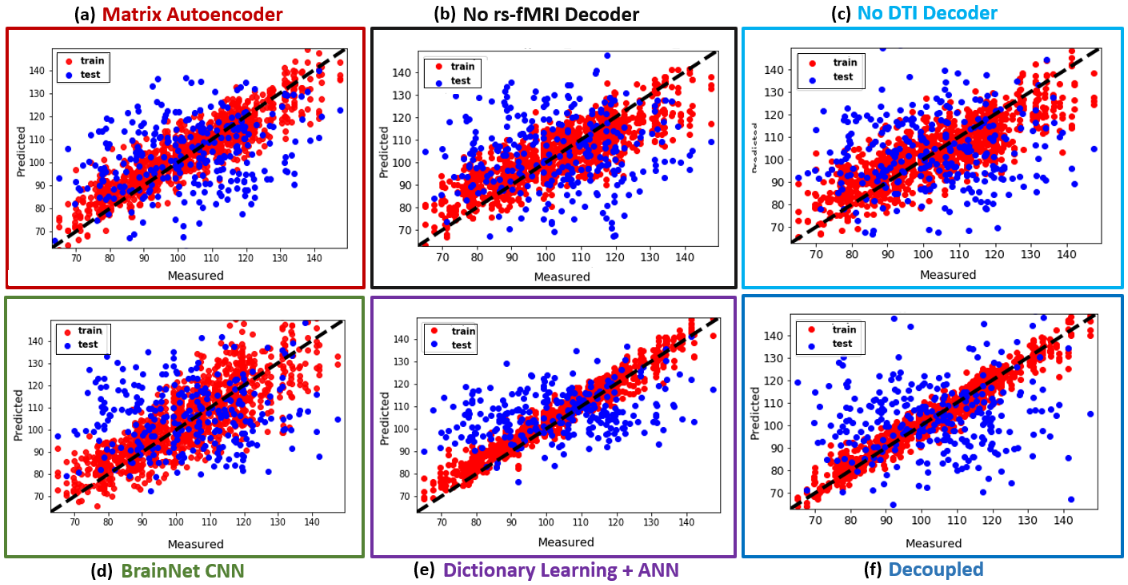


Figure 6.2: HCP Dataset Prediction of CFIS by **(a)** Our Framework **(b)** Matrix AE without rs-fMRI Decoder **(c)** Matrix AE without DTI Decoder **(d)** BrainNet CNN **(e)** Dictionary Learning + ANN **(f)** Decoupled Matrix AE and ANN

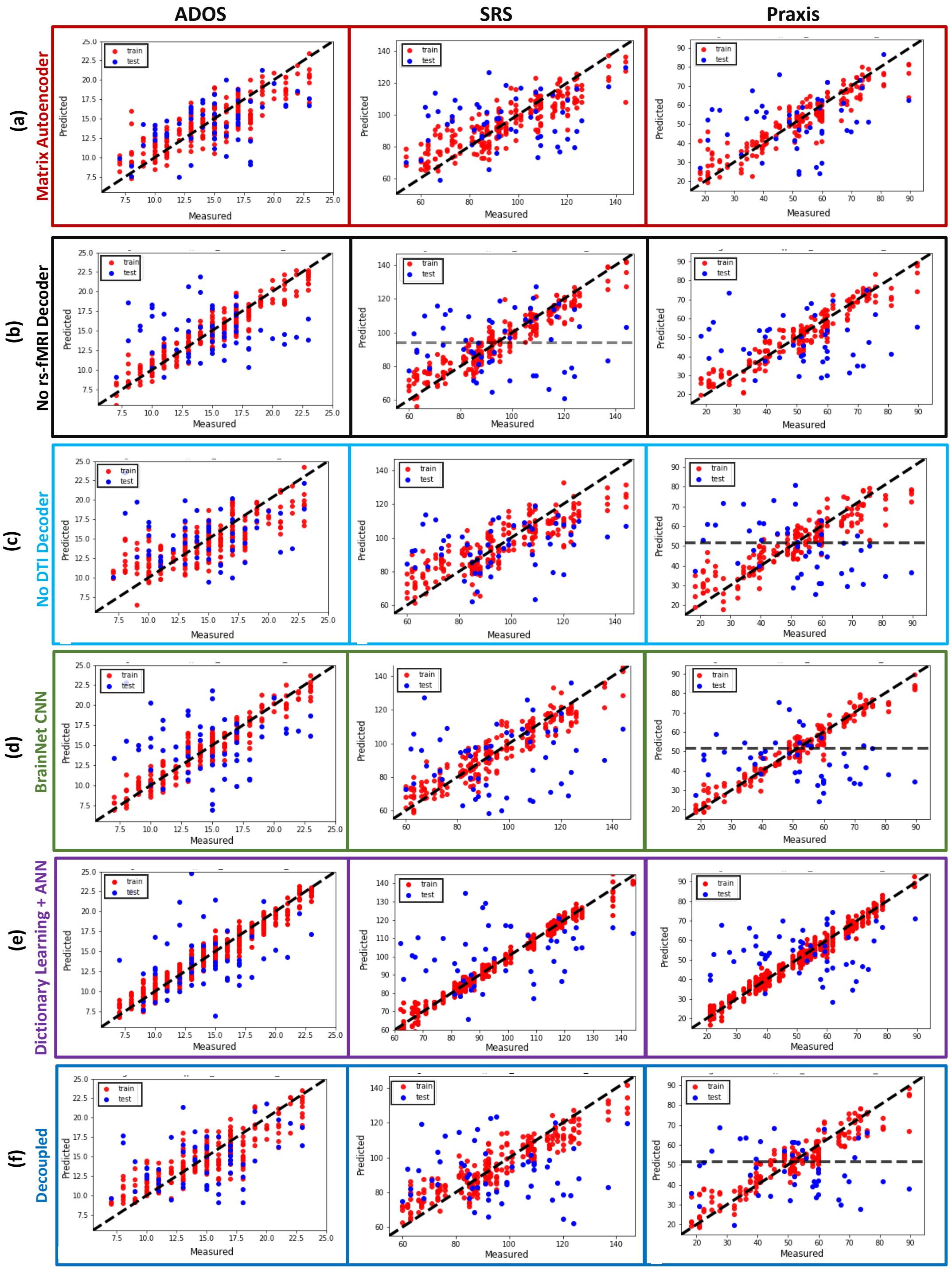


Figure 6.3: ASD Dataset: Multi-output prediction performance of **(L):** ADOS **(M):** SRS **(R):** Praxis by **(a)** Our Framework **(b)** Matrix Autoencoder without rs-fMRI Decoder **(c)** Matrix Autoencoder without DTI Decoder **(d)** BrainNet CNN **(e)** Dictionary Learning + ANN **(f)** Decoupled Matrix Autoencoder and ANN

regression between the vectorized upper diagonal FC features (input) and SC features (output) to help evaluate the benefit of our matrix decomposition. As seen, our function → structure decoding achieves similar performance as directly encoding/decoding the structural connectivity. At the same time, the linear regression baseline performs worse than both of these techniques. This suggests that the ability to directly leverage the low rank matrix structure is key to preserving individual differences during reconstruction.

Second, we use t-SNE to visualize the symmetric FC and SC embeddings, $\mathbf{F}_n$ and $\mathbf{S}_n$, respectively. Fig. 6.4 (b) and Fig. 6.6 (b) (KKI) displays the 2D t-SNE representation computed from the upper-triangle entries of the embedding. As seen, the FC and SC are clustered in two different locations within this space. Interestingly, the learned representations are non-overlapping without explicit enforcement. This suggests that the alignment decoder $\mathbf{\Phi}_{\text{align}}(\cdot)$ is learning a conversion between manifolds.

Third, we examine the stability of the transformation learned by the alignment decoder, i.e. the weights $\mathbf{W}_{\text{align}} \in \mathcal{R}^{K\times K}$ of $\mathbf{\Phi}_{\text{align}}(\cdot)$. We first match the

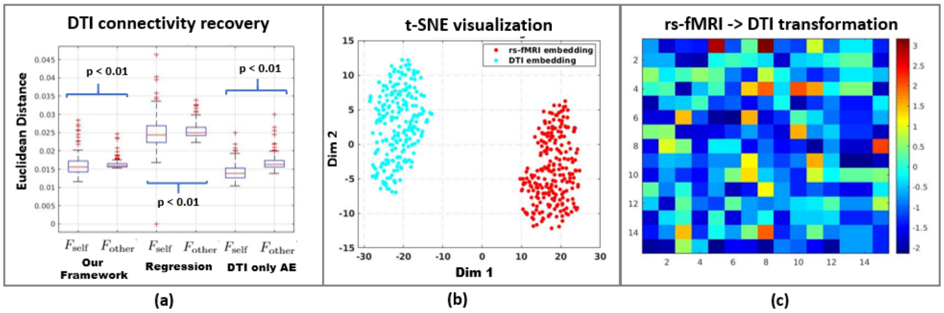


Figure 6.4: **(a)** Recovery of SC for **(L):** Our Framework **(M):** Linear Regression **(R):** DTI only Autoencoder **(b)** t-SNE visualization for FC and SC embeddings **(c)** Coefficient of Variation (C_v) (log scale) for the weights of $\mathbf{\Phi}_{\text{align}}(\cdot)$. Cold colors imply small deviations, i.e. better stability

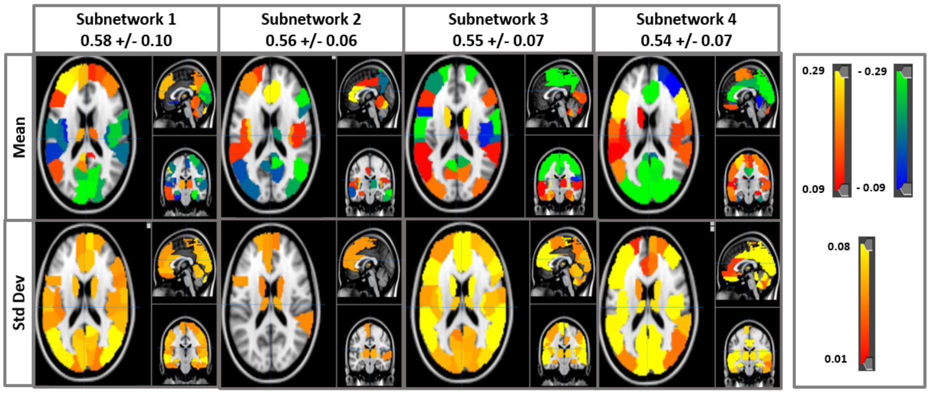


Figure 6.5: Top four bases learned by the Matrix Autoencoder measured by the absolute correlation coefficient across cross validation folds and initializations.

columns of $\mathbf{W}_{\text{align}}$ across cross validation folds according to correspondences between the functional brain basis. For each entry of $\mathbf{W}_{\text{align}}$, we compute the coefficient of variation (C_v), i.e. the ratio of the standard deviation to the mean (in absolute value). Lower values of C_v indicate smaller deviations from the mean values, i.e. better stability. Fig. 6.4(c) (HCP) and Fig. 6.6(c) (KKI) displays the log coefficient of variation $\log(C_v)$, where the cool colors indicate smaller C_v. As seen, a majority of the entries of $\mathbf{W}_{\text{align}}$ have low variation over the mean pattern value. Overall, our results suggest that our framework learns a stable mapping across the manifolds that explains individual patterns of structural connectivity faithfully.

6.2.4 Evaluating Functional Biomarkers

We explore the functional connectivity patterns learned by our framework by first matching the brain bases (i.e., columns of $\mathbf{W}$) across the cross validation folds based on the absolute correlation coefficient. We run this experiment five times with different initializations for the ANN branch to check for consistency

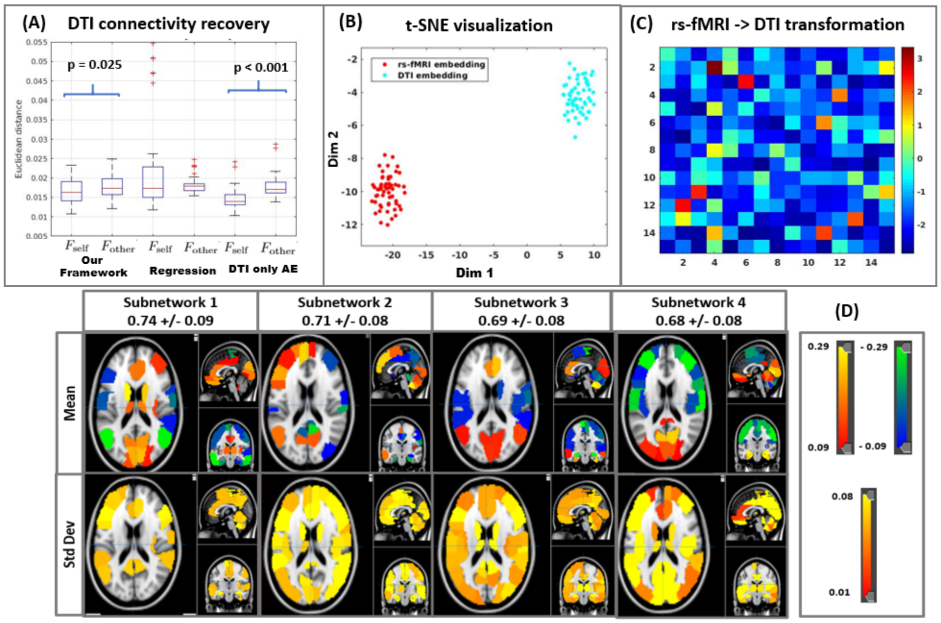


Figure 6.6: KKI Dataset: (A) SC recovery by **(L):** Our Framework **(M):** Linear Regression **(R):** DTI only AE **(B)** t-SNE visualization of embeddings **(C)** Coeff. of Var. (C_v) (log scale) for $\mathbf{\Phi}_{\text{align}}(\cdot)$. Cold colors imply better stability **(D)** Top four FC bases

in the learned representation. Fig 6.5 (HCP) displays the four most consistent bases, as projected onto the brain using the region definitions of the AAL atlas. In each case, we report the mean and standard deviation of the basis across folds. We notice that while there is spatial overlap between the bases, the standard deviations are small, which indicates that our framework is learning stable patters in the data. Subnetwork 1 highlights regions from the default mode network, which is widely inferred within the resting state literature, and known to play a critical role in consolidating working memory [183]. Subnetworks 1, 3 and 4 highlight regions from the somatomotor network and visual cortex, together believed to be important functional biomarkers of cognitive intelligence [189]. Finally, Subnetwork 2 and 4 displays contributions from the frontoparietal network and the medial prefrontal network. These areas are believed to play a role in working memory, attention, and decision making, all of which are associated with cognitive intelligence [186]. Fig. 6.6 (D) displays the bases learned when we train the Matrix Autoencoder on the KKI Dataset.

6.3 Summary

We have introduced a novel matrix autoencoder to map the manifold of rs-fMRI functional connectivity to the manifold of DTI structural connectivity. Our framework is strategically designed to leverage the underlying geometry of the data spaces and robustly recover brain biomarkers that are simultaneously explanative of behavioral phenotypes. We demonstrate that our framework offers both interpretability and generalizability, even for multi-score prediction on modest sized datasets. Finally, our framework makes minimal assumptions, and can potentially find application both within and outside the medical realm.

6.4 Probing the Encoder-Decoder Representation

Beyond clinical applicability, we gear the rest of this discussion towards aspects of representation learning. This is primarily motivated by an interesting empirical observation that arose in our experiments on the Matrix Autoencoder. In Fig. 6.7, we plot the deviation of the basis (column normalized $\mathbf{W}$ in Eq. (6.1)) from orthogonality, i.e. $||\mathbf{W}^T\mathbf{W} - \mathcal{I}_K||_F^2$. On the left, we plot this quantity for the basis of the encoder-decoder (gray box in Fig. 6.1). On the

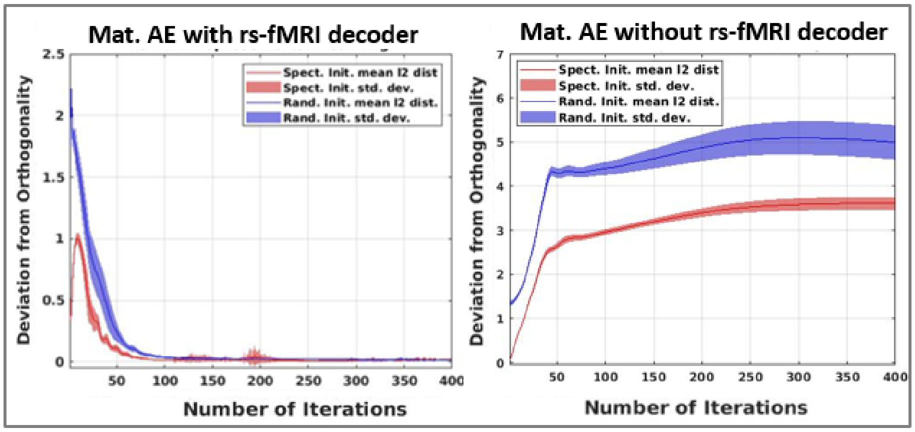


Figure 6.7: HCP Dataset: Deviation of basis from orthogonality **(L)** Matrix Autoencoder with rs-fMRI decoder **(R) Matrix Autoencoder without rs-fMRI decoder**

right, we plot this quantity for the basis of $\mathbf{\Phi}_{\text{enc}}(\cdot)$ when the rs-fMRI decoder $\mathbf{\Phi}_{\text{dec}}(\cdot)$ is removed. In each case, the red plot corresponds to a spectral initalization for the basis, which uses the top eigenvectors of the average correlation matrix ($\bar{\mathbf{\Gamma}}$) for the training examples. The blue plots correspond to a default random initialization. We repeat this experiment on the HCP dataset 10 times using a subset of the 275 patients for training. Interestingly, we observe that for the plots on the left, as the training proceeds, the recovered basis move closer to being nearly orthogonal regardless of the initialization. On the other hand, when we remove the rs-fMRI decoder, we no longer observe this behavior. We also observed the same trend in experiments with the KKI dataset. Our subsequent analysis aimed at uncovering the theoretical underpinnings behind this behavior.

6.4.1 Parallels with Common Principal Components

Recall that in previous chapters, our matrix decomposition strategy on rs-fMRI correlation matrices was heavily inspired by the Common Principal Components (CPC) [211] formulation.

From a modeling standpoint, CPC is the generalization of principal components to several populations. CPC was designed to model covariance matrices (henceforth referred to as data matrices $\{\mathbf{\Gamma}_n\}$) among groups of multidimensional datasets. Their inner products are constrained to share the same eigenvectors and are therefore simultaneously diagonalizable by a common decorrelator matrix.

In CPC, the generating common basis $\mathbf{B} \in \mathcal{R}^{P\times K}$ is a concatenation of K elemental bases vectors $\mathbf{b}_k \in \mathcal{R}^{P\times 1}$, i.e. $\mathbf{B} := [\mathbf{b}_1 \quad \mathbf{b}_2 \quad ... \quad \mathbf{b}_K]$, where $K \ll P$. The basis vectors are also constrained to be orthogonal to each other to ensure that the learned bases are uncorrelated, yet explain the data well.

While the bases (eigenvectors) are shared, the strength of their combination differs across individual data matrices. These loadings (eigenvalues) are denoted by the set $\{\mathbf{c}_n\}$ and combine the bases uniquely to constitute $\mathbf{\Gamma}_n$. The non-negativity constraint $\mathbf{c}_n$ ensures the positive semi-definiteness of $\mathbf{\Gamma}_n$. In a noiseless setting, the complete data representation is as follows:

$$\mathbf{\Gamma}_n \approx \sum_k \mathbf{c}_{nk}\mathbf{b}_k\mathbf{b}_k^T \quad s.t. \quad \mathbf{c}_{nk} \geq 0, \;\; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K, \tag{6.5}$$

where $\mathcal{I}_K$ is the $K \times K$ identity matrix. As seen in Eq. (6.5), the loading vectors, $\mathbf{c}_n := [\mathbf{c}_{n1} \quad ... \quad \mathbf{c}_{nK}]^T \in \mathcal{R}^{K\times 1}$ seek to model the variation in the dataset. Denoting $\mathbf{diag}(\mathbf{c}_n)$ as a diagonal matrix with the K coefficients on the diagonal and off-diagonal terms set to zero, Eq. (6.6) can be re-written in the following matrix form:

$$\mathcal{L}_{\text{CPC}}(\{\boldsymbol{\Gamma}_n\}; \mathbf{B}, \{\mathbf{c}_n\}) = ||\boldsymbol{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 \quad s.t. \quad \mathbf{c}_{nk} \geq 0, \;\; \mathbf{B}^T\mathbf{B} = \mathcal{I}_K \tag{6.6}$$

This objective serves to minimize the discrepancy between the observed data and the assumed CPC generating process.

6.4.1.1 CPC Inference Procedure

We can adopt the alternating minimization procedure in Section 5.4 to optimize Eq. (6.6) with a few minor modifications. Specifically, instead of T_n constraints per subject, we can use a single constraint of the form $\mathbf{D} = \mathbf{B}$, enforced via a single Augmented Lagrangian $\boldsymbol{\Lambda}$. This effectively ensures that the new objective has a quadratic form in $\mathbf{B}$, along with a closed form update for $\mathbf{D}$. As before, we cycle through three individual steps, namely:

- Closed form Procrustes solution for the basis $\mathbf{B}$
- Updating the loadings $\{\mathbf{c}_n\}$ (having a quadratic form)
- Augmented Lagrangian updates for the constraint variables $\{\mathbf{D}, \boldsymbol{\Lambda}\}$. $\mathbf{D}$ has a closed form Procrustes solution as well.

6.4.1.2 Comparing Representational Aspects:

Reverting back our attention to our end-to-end geometric framework, we observe several parallels with CPC. For the purposes of this discussion, we refer to the representation in the gray box in Fig. 6.1 and Eq. (6.1) as the Matrix

Autoencoder representation, which we examine in isolation. For simplicity, we optimize for the parameters of this network using the Stochastic Gradient Descent algorihthm with a learning rate of 0.0005. The matrix factorization imposed by this framework is represented as follows:

$$\mathcal{L}_{\text{MatAE}}(\{\mathbf{\Gamma}_n\};\mathbf{W}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{W}\mathbf{W}^T\mathbf{\Gamma}_n\mathbf{W}\mathbf{W}^T||_F^2 \tag{6.7}$$

$$= \sum_n ||\mathbf{\Gamma}_n - \mathbf{W}\mathbf{F}_n\mathbf{W}^T||_F^2 \;\text{ where }\; \mathbf{F}_n = \mathbf{W}\mathbf{\Gamma}_n\mathbf{W}^T \tag{6.8}$$

Both CPC in Eq. (6.6) and the Matrix Autoencoder decomposition in Eq. (6.8) optimize for a rank K canonical outerproduct decomposition. By construction, the factor $\mathbf{c}_n$ is constrained to be a non-negative diagonal loading matrix. As a consequence of the orthonormality constraint in Eq. (6.6), we can show [212] that for a fixed estimate of the basis $\mathbf{B}$, the optimal solution for the loading vector is $\mathbf{diag}(\mathbf{c}_n) = \mathbf{diag}(\mathbf{B}^T\mathbf{\Gamma}_n\mathbf{B}) = (\mathbf{B}^T\mathbf{\Gamma}_n\mathbf{B}) \circ \mathcal{I}_K$ at each iterative estimate. On the other hand, the factor $\mathbf{F}_n = \mathbf{W}^T\mathbf{\Gamma}_n\mathbf{W}$ in the Matrix Autoencoder has a similar projection form, but is allowed to have off diagonal terms along with the non-negative diagonal terms.

6.4.2 Experiments on Synthetic Data

Through our experiments on synthetic data, we wish to qualitatively and quantitatively compare aspects of representation learning across the CPC and Matrix Autoencoder formulations. To this end, we sample random positive semi-definite matrices $\{\mathbf{\Gamma}_n\}$ according to a common principal components [211] generating process.

We start by generating a common low rank basis matrix $\hat{\mathbf{B}}_{no} \in \mathcal{R}^{P \times K}$ by drawing its entries independently from a zero mean Gaussian with variance one. We choose $K = 30$, $P = 116$ and number of examples $N = 100$. We then use the Gram-Schmidt procedure to compute an orthogonal basis $\mathbf{B} = \mathbf{orth}(\hat{\mathbf{B}}_{no})$. Note that the coefficient values in $\mathbf{c}_n$ are independent across networks and data examples. Thus, for each example, we generate the coefficients using a Gaussian distribution with zero mean, and variance 1. The true data matrices are $\mathbf{\Gamma}_n = \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T$.

We introduce additive corruptions to the data generating process of the form $\sigma^2 \mathbf{X}_n \mathbf{X}_n^T$. This also ensures that the resulting data matrices are positive semi-definite. We experiment with two noise scenarios. In the first, we draw $\mathbf{X}_n$ randomly from the null space of $\mathbf{B}$, constraining the scale at $||\mathbf{X}_n||_2 = 1$. For low noise regimes, these corruptions are expected to not generate large perturbations in the column space of the ground truth basis. In the second scenario, the entries of $\mathbf{X}_n$ are drawn from a random normal distribution, which is a case where noise corruptions have contributions in the column space of the basis even for low noise regimes. We sweep σ from $0.1 - 0.45$. We generate box plots using 10 repeated trials using different random initializations. This helps us visualize and quantify variations in the recovery of the generating process across such trials.

6.4.2.1 Perturbations in the Null Space

We first run a sanity check by adding noise strictly to the null space of the generating low dimensional basis. Fig. 6.8 plots the recovery performance

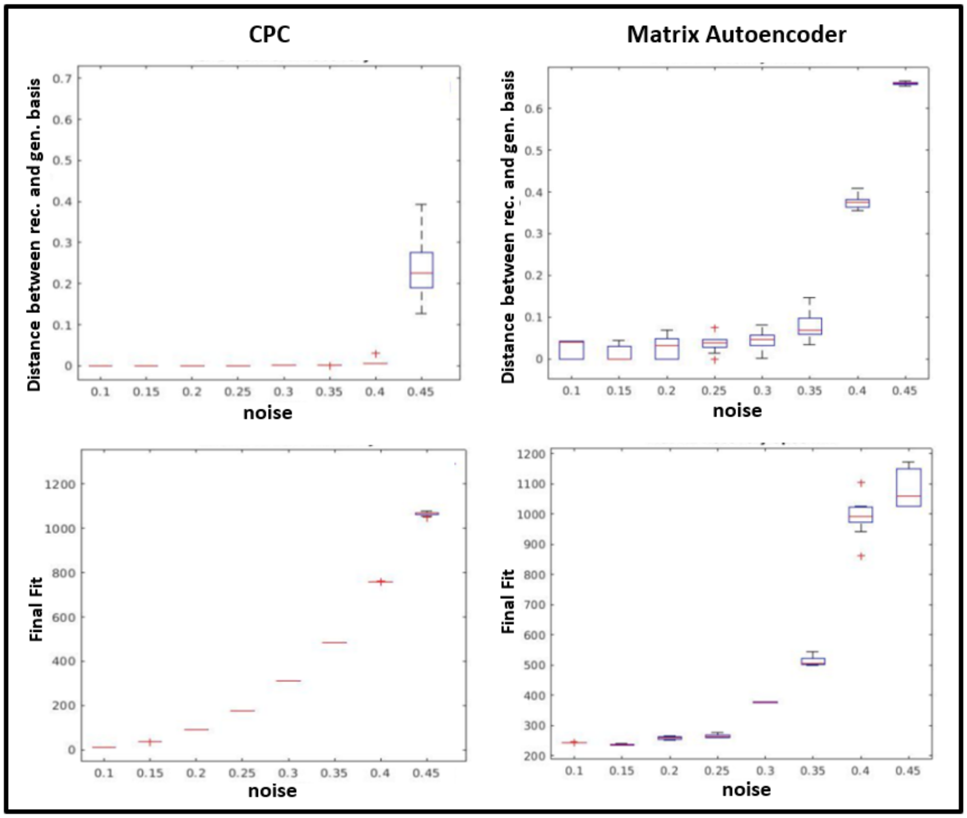


Figure 6.8: Recovery Performance using **(L)** CPC algorithm and **(R)** Matrix Autoencoder according to **(Top)** Distance between generating and recovered bases and **(Bottom)** Data fit. Perturbations are restricted to the null space of the generating basis

of the CPC algorithm to the left and the Matrix Autoencoder to the right. The x-axis represents the noise, while the y axis plots the distance between the generated and recovered basis in the top set of plots, and the fit on the training data. Since solutions may have rotational (and scaling for the Matrix Autoencoder) equivalence, we normalize the columns to unit norm and perform a Procrustes alignment. The residual fit is thus obtained. As expected, the recovery performance of the CPC algorithm is better than that the Matrix Autoencoder (as indicated by the smaller fit error bars) since it is reflective of the generating process for the data. For large noise settings, the CPC is

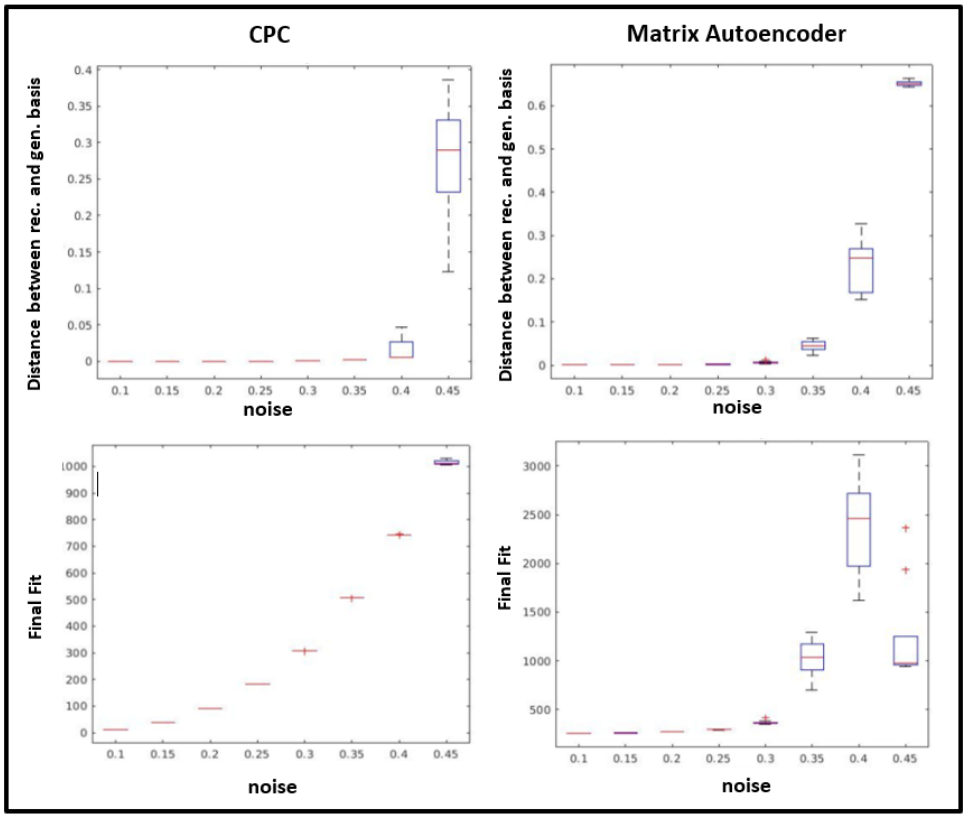


Figure 6.9: Recovery Performance using **(L)** CPC algorithm and **(R)** Matrix Autoencoder according to **(Top)** Distance between generating and recovered bases and **(Bottom)** Data fit. Perturbations are random and can generate corruption onto the ground truth basis

more robust in terms of the recovery of the generating bases, while the matrix autoencoder canonical bases are corrupted more easily, resulting in a worse data fit. For low-moderate noise settings ($\sigma^2 \leq 0.35$), there is also a large overlap between the bases recovered by the two algorithms after accounting for rotation and scaling.

6.4.2.2 Random Perturbations in the Data Space

In this scenario, we draw the entries of $\mathbf{X}_n$ from a random normal. This inherently creates corruptions in the column space of the generating basis. We observe that the fidelity of the Matrix Autoencoder representation under these perturbations in comparison with the CPC is lower even for relatively moderate scales of noise.

Interestingly, we observe that the recovered bases for both algorithms have a considerable overlap even when the bases across examples do not share a common generating process. Taken together, all these observations suggest a fundamental link between the two representation learning schemes.

6.4.2.3 Examining Modeling Constraints

CPC Relaxation: Since the Matrix Autoencoder in its formulation does not impose constraints on the basis, a natural direction would be to examine the effect of relaxing the orthonormality and non-negativity constraints in the CPC algorithm (following in Section 6.4.1.1). We refer to this setting as the Unconstrained CPC via the objective below:

$$\mathcal{L}_{\text{CPC unc}}(\{\mathbf{\Gamma}_n\}; \mathbf{B}, \{\mathbf{c}_n\}) = ||\mathbf{\Gamma}_n - \mathbf{B}\mathbf{diag}(\mathbf{c}_n)\mathbf{B}^T||_F^2 \tag{6.9}$$

An interesting observation we made is that the recovered bases for the unconstrained case are still orthogonal. At the same time, there arises an expected difference in scaling magnitude factor for recovered solutions $\{\mathbf{B}_k\alpha, \frac{1}{\alpha^2}\mathbf{c}_{nk}\}$. We note that this behavior continues to persists even when we break the generating process by not using a common generating basis across examples. In

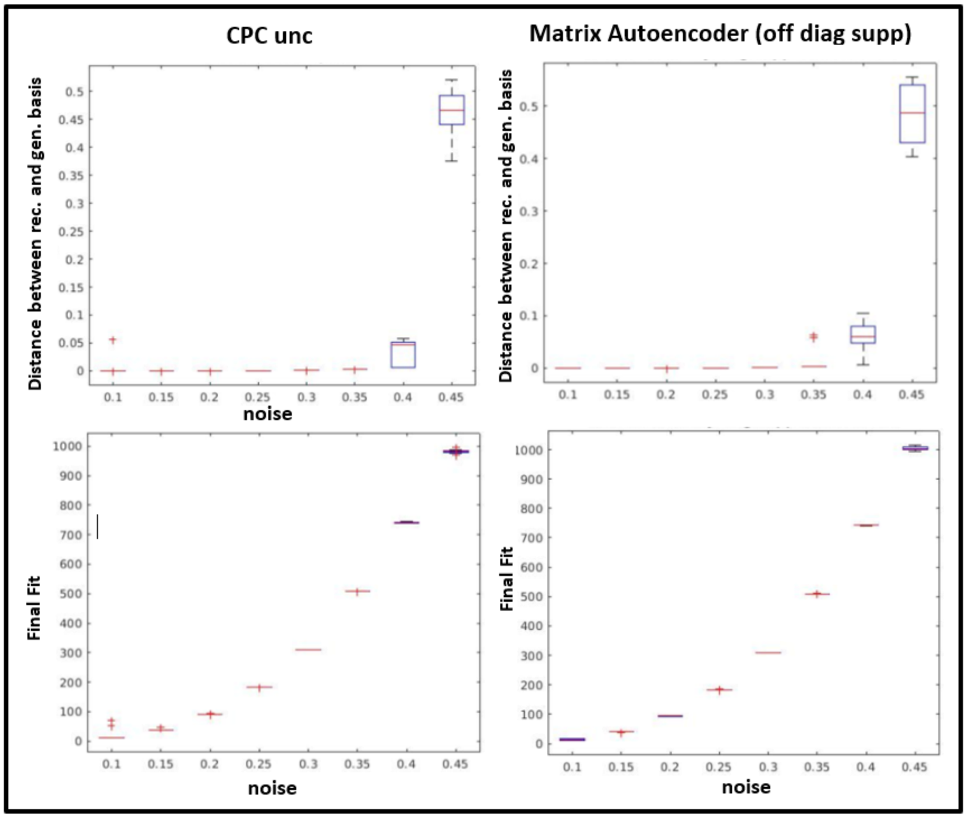


Figure 6.10: Recovery Performance using **(L)** Unconstrained CPC algorithm and **(R)** Matrix Autoencoder (off diagonal suppressed) according to **(Top)** Distance between generating and recovered bases and **(Bottom)** Data fit. Perturbations are random and can generate corruption onto the ground truth basis

fact, it can be analytically shown that when the estimate of **B** is held constant in the objective in Eq. (6.9), the optimal estimates for the coefficients can be computed as:

$$\mathbf{c}_n = \mathbf{A}[\mathbf{diag}(\mathbf{B}^T\boldsymbol{\Gamma}_n\mathbf{B})] = \mathbf{A}[(\mathbf{B}^T\boldsymbol{\Gamma}_n\mathbf{B}) \circ \mathcal{I}_k]\mathbf{1}_K \tag{6.10}$$

$$\text{where } \mathbf{A} = [(\mathbf{B}^T\mathbf{B}) \circ (\mathbf{B}^T\mathbf{B})]^{-1} \tag{6.11}$$

with **A** arising as a result of the relaxation of the orthonormality constraint.

Inference wise, we continue to utilize an alternating minimization procedure to optimize for Eq. (6.9). However, the closed form solutions for $\mathbf{B}, \{\mathbf{c}_n\}$ (and $\mathbf{D}$) are recomputed to reflect the removal of the constraints. At the same time, the observation is nevertheless surprising given that no explicit orthogonality constraints are enforced on the basis representation by any of the updates.

Constraining the Matrix Autoencoder: Recall that the matrix autoencoder learns a low dimesnional projection matrix of the form $\mathbf{F}_n = \mathbf{W}^T\mathbf{\Gamma}_n\mathbf{W}$. This symmetric product is allowed to have both off-diagonal and (non-negative) diagonal contributions. We now examine aspects of the recovery when we artificially discard the off diagonal contributions in the forward pass, i.e. set $\mathbf{F}_n = \mathbf{diag}(\mathbf{W}^T\mathbf{\Gamma}_n\mathbf{W}) = (\mathbf{W}^T\mathbf{\Gamma}_n\mathbf{W}) \circ \mathcal{I}_K$. This is similar in spirit to the diagonal CPC loading in Eq. (6.6), albeit without additional constraints on the weights $\mathbf{W}$. Mathematically, we can write this as:

$$\mathcal{L}_{\text{MatAE ODS}}(\{\mathbf{\Gamma}_n\};\mathbf{W}) = \sum_n ||\mathbf{\Gamma}_n - \mathbf{W}\mathbf{diag}(\mathbf{W}^T\mathbf{\Gamma}_n\mathbf{W})\mathbf{W}^T||_F^2 \tag{6.12}$$

$$= \sum_n ||\mathbf{\Gamma}_n - \mathbf{W}\mathbf{F}_n\mathbf{W}^T||_F^2 \text{ where } \mathbf{F}_n = (\mathbf{W}\mathbf{\Gamma}_n\mathbf{W}^T) \circ \mathcal{I}_K \tag{6.13}$$

We continue to encounter the puzzling phenomenon of recovering nearly orthonormal bases weights despite the lack of explicit modeling constraints.

The plots on the left of Fig. 6.10 illustrate the performance of the unconstrained CPC. for the same random perturbations in the previous subsection, we observe that the recovery performance here is similar to that of the CPC (with the orthonormality constraints) across the noise regime. The plots on the right of Fig. 6.10 correspond to the Matrix Autoencoder (with off diagonal terms suppressed) indicate that the recovery performance shows marked improvement over the original Matrix Autoencoder formulation. Additionally, we obtain recovery performance comparable to our experiments on the CPC (both constrained and unconstrained), which we can think of as ground truth for this experiment. We also observed a high overlap between the recovered bases (modulo rotation and scaling equivalence) across the CPC (constrained, unconstrained) and this formulation.

6.4.3 Scope and Ongoing Work

Our work in this thesis has heavily relied on the representational power of carefully crafted mathematical models that incorporate data geometry. Specifically in the context of this Chapter, our preliminary experiments suggest a fundamental link between the classical and end-to-end geometric representations studied in Eq. (6.6-6.13). Our current understanding of these frameworks does not sufficiently explain the observed representational phenomenon. We suspect this would require machinery from the geometry of Riemannian (specifically PSD) manifolds and from the optimization literature. As an ongoing effort, we are working on developing such an analytic approach to refine our theoretical understanding of these frameworks.

Chapter 7

Conclusion

This chapter serves as a summary of the main ideas and frameworks developed in this thesis. In a nutshell, we introduced powerful mathematical modeling strategies to jointly analyze brain connectivity and behavior for clinical applications.

This concluding chapter is organized as follows. We will first summarize the main ideas and the scope of this work and our findings. Next, we consolidate the comparisons across models we developed in each chapter to carefully weight the advantages and tradeoffs of each framework. To this end, we will provide qualitative and quantitative discussions to solidify these comparisons. Finally, we provide a brief discussion on potential technical and clinical extensions to the ideas presented.

7.1 Discussion

7.1.1 Overview

In Chapter 3, we built a generative-discriminative framework (i.e. the **Joint Network Optimization (JNO)** model) that can predict behavior from rs-fMRI connectomes. This work extended the field of representation learning in functional connectivity for clinical characterizations. We learn a mapping between the two spaces which is both generalizable and interpretable. Chapter 4 explored extensions of the predictive framework beyond linear prediction (i.e. the **Coupled Manifold Optimization (CMO)** and the **Dictionary Learning + ANN** frameworks) in the form of non-parametric regression and deep models respectively. This helps us improve the representational power of the discriminative framework. We observe that combining classical models with deep learning allows us to expand our behavioral characterizations to multidimensional phenotypes. We view this characterization as a first step to improving our understanding of the pathogenesis of complex disorders.

Chapter 5 focused on two key technical modeling innovations. We introduced the use of structural tractography (DTI) as anatomical priors that provide complementary connectivity information. To this end, we developed a deep generative hybrid model (i.e. the **Deep structurally regularized Dynamic Dictionary Learning (Deep sr-DDL)**. Secondly, our deep-generative framework also incorporated a time-varying picture of functional connectivity. Our comparisons in Tables 5.1 and 5.2 help us evaluate the benefit of each component (dynamic functional connectivity and multimodal integration) individually via ablation studies. Moreover, our strategically designed deep networks allow us to mimic the evolution of constituent brain states by modeling them as canonical subnetworks. This framework tracks the temporal phasing of these states and parallels the transient brain states hypothesis in dynamic connectivity. Our use of temporal attention models adds to model interpretability by automatically underscoring clinically predictive scan time points. Together, such dynamic modeling of brain states could pave way for a more nuanced comparison across patient sub-populations.

In all of these works, our optimization procedure was shown to be key to obtaining *generalizable* representations. This property was verified by detailed performance comparisons against pipelined versions of each framework in Tables 3.1,4.1,5.1,5.2 . Overall, this inference procedure allowed us to explicitly couple the neuroimaging and clinical spaces, ultimately leading to the extraction of clinically relevant bases. This joint inference is a departure from prior work in this area, where the feature selection and prediction modules are typically decoupled.

In Chapter 6, we examined a slightly different yet important representation learning paradigm, i.e. that of examining the complementarity between function and structure. Deriving inspiration from classical models, we developed an end-to-end geometric framework (the **Matrix Autoencoder** model) to learn an explicit mapping from functional to structural connectivity matrices. At the same time, we used multidimensional phenotypes as a secondary guide in a predictive setting. From a technical standpoint, this framework marries the

Meas.	Method	MAE Test	NMI Test
CFIS	JNO	16.36 ± 14.28	0.63
	CMO	15.91 ± 14.78	0.64
	Dict. Learn. + ANN	15.26 ± 13.99	0.66
	Deep sr-DDL.	16.31 ± 15.43	0.67
	Matrix Autoencoder	14.08 ± 11.85	0.69
	M-GCN	**12.87** ± **9.65**	**0.73**

Table 7.1: HCP Dataset: Evaluation of single target regression using the **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)** for the test set. Best performance is highlighted in bold. Second best is underlined

best of two worlds, i.e. the interpretability in classical models with the simplicity of end-to-end deep stochastic optimization. From a clinical standpoint, this helps us better understand the interplay between function, structure and behavior.

Lastly, Chapter 8 takes an alternate end-to-end deep learning approach that deviates from classical decomposition based models. Instead, we treat the brain as a network graph entity, with the communication patterns of the brain dictated by functional and structural connectivity data. Our **Multimodal Graph Convolutional (M-GCN)** framework is capable of exploiting topological properties of the brain graph via carefully designed graph filtering operations. Overall, this provides improved phenotypic prediction performance in comparison with our previous models.

7.1.2 Comparing Representational Frameworks

This section provides cross comparisons across the suite of mathematical models introduced in this thesis, i.e. the Joint Network Optimization (JNO), the Coupled Manifold Optimization (CMO), the Dictionary Learning + ANN, the

Meas.	Method	MAE Test	NMI Test
ADOS	JNO	2.79 ± 2.35	0.32
	CMO	3.17 ± 2.00	0.35
	Dict. Learn. + ANN	2.71 ± 2.40	0.43
	Deep sr-DDL	2.84 ± 2.79	0.34
	Matrix Autoencoder	**2.71** ± **1.84**	**0.49**
	M-GCN	2.71 ± 2.15	0.45
SRS	JNO	43.27 ± 30.14	0.61
	CMO	33.11 ± 28.07	0.51
	Dict. Learn. + ANN	16.79 ± 14.83	**0.89**
	Deep sr-DDL	17.81 ± 16.09	0.88
	Matrix Autoencoder	16.04 ± 13.40	0.83
	M-GCN	**16.50** ± **9.44**	0.85
Praxis	JNO	27.12 ± 29.66	0.53
	CMO	30.11 ± 26.47	0.61
	Dict. Learn. + ANN	13.19 ± 10.75	0.82
	Deep sr-DDL	13.50 ± 11.55	0.85
	Matrix Autoencoder	13.14 ± 10.78	**0.86**
	M-GCN	**12.82** ± **12.04**	**0.86**

Table 7.2: KKI Dataset: Evaluation of multiscore prediction using the **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)**. Best performance is highlighted in bold. Near misses are underlined

Deep structurally regularized dynamic dictionary learning (Deep sr-DDL), the Matrix Autoencoder framework, and the Multimodal Graph Convolutional framework (M-GCN).

To streamline the discussion, we focus our quantitative comparisons on the multi-score phenotypic prediction task on our Autism dataset and the prediction of cognitive fluid intelligence score on the healthy controls in the HCP dataset. Tables 7.1 and 7.2 provide a head to head comparison of these models on the HCP and KKI dataset respectively.

In Chapters 3 and 4, we observed that the JNO and the CMO frameworks are one of the first frameworks that generalize onto clinical prediction, outperforming several machine learning pipelines. By design, these frameworks allow us to probe the learned representation by extracting subnetwork patterns that are most clinically predictive. While they can be tuned to predict a single measure (CFIS in Table 7.1 or the Autism severity measures in Tables 3.1 and 4.1) faithfully, they do not have the flexibility to predict multiple measures at the same time. We can see from the predictive performance in Table 7.2, they can be tuned to predict one of the three measures well (ADOS), but do not generalize to predicting Praxis and SRS. In fact, we found that adjusting the hyperparameters for these models allows us to predict one of the scores well, but at the expense of generalization to the other two measures.

On the other hand, the Dictionary learning + ANN and Deep sr-DDL models provide us with this ability. The added representational flexibility offered by deep learning allows us to extract a consistent set of brain bases that can explain a spectrum of behavioral deficits. In addition, the Deep sr-DDL provides allow us to track the temporal evolution of brain states (See Fig. 5.14) and incorporate anatomical priors into our functional connectivity representation. Potentially, this could allow us to perform more nuanced comparisons across sub-types within the same population and better understand complex disorders and their behavioral manifestation.

The Matrix Autoencoder framework can be thought of as an end-to-end version of the previous models, where the low-rank matrix structure is key to representation learning. In turn, we notice that it provides comparable generalization to multi- and single score phenotypic prediction on both datasets

(Tables 7.1 and 7.2) against the Dictionary Learning+ ANN and Deep sr-DDL model and second best performace overall. However, there are two notable benefits to the autoencoder representation. Firstly, the end-to-end training of the autoencoder allows us to avail the computational speed-up and simplicity of end-to-end stochastic optimization as opposed to alternating inference strategies. Secondly, it allows us to explicitly learn a mapping between the functional and structural connectivity spaces, which our other frameworks do not explore. In turn, this setup incorporates and implicitly leverages the geometry of two spaces, offering elegant strategies to model connectivity data.

In terms of predictive performance, the M-GCN model clearly provides the best generalization on the HCP dataset. It also demonstrates improved performance on the multiscore prediction task (best performance on SRS and Praxis and close to best performance on ADOS). As mentioned previously, the heterogeneity in the measures renders this task particularly challenging. As opposed to the previous models, the M-GCN is a geometric model of the multimodal brain graph. The filtering operations are strategically designed to leverage topological information within the architecture. This also departs from a one-nearest neighbour flavour of regularization (as with the sr-DDL) as it allows us to incorporate multi-hop pathways of structural connectivity. The structural regularization allows the framework to efficiently extract generalizable representations from limited data. Despite its successes, this model is less straightforward to interpret. In this vein, recent advances in the field of geometric deep learning suggest tools to formalize explainability and interpretability [213] of such graph models and could be beneficial for future applications of the M-GCN.

7.2 Scope

As seen in this thesis, our methods allow us to hone in on key predictive resting state biomarkers from healthy and autistic populations. Additionally, our frameworks makes minimal assumptions. Provided we have access to a set of consistently defined structural and functional connectivity measures and clinical scores, these tools can be easily adapted to other neurological disorders . Long term, such studies may spurn discoveries and advances in challenging translational and clinical paradigms such as biomarker development, behavioral therapeutics, neuro-surgical pre-planning etc. In fact, such principles may even benefit predictive network models outside the medical realm. Overall, these findings broaden the scope of our method for future applications.

7.2.1 Limitations

Despite the benefits of these frameworks, they do suffer from certain limitations.

Although our models have been shown to outperform several baselines on two separate datasets, our prediction performance is far from perfect, both for single and multi-target prediction. One of the key reasons can be attributed to inherent noise in the clinical measures themselves. For example, SRS is based on a parent-teacher questionnaire, which tends to be more subjective than a clinical exam [214]. This renders the behavioral prediction task especially challenging, and can partially account for the poor performance of predictive models. Even diagnostic examinations such as ADOS, in which trained clinicians score patients tend to suffer from some variability in reporting across individual, often at floor and ceiling values of these scales [215, 216]. Current findings in literature have found to cap out at a predictive performance of about ten percent of total dynamic range [217, 201, 218]. Going one step further, the performance trends within our frameworks and the baselines suggest that multi-score prediction is a notoriously challenging clinical problem. Given the exploratory nature of this work, careful investigation is warranted in clinically interpreting and working towards improving performance results in practice.

As another example, the assumptions made by these models may be restrictive for capturing the full complexity of the brain. Our frameworks are relatively simple to lay the groundwork for such analyses. Thus they do not directly explore and explicitly incorporate notions of heirarchical and modular organization of the brain, as well as sophisticated machinery for temporal connectivity tracking such as conditional correlations.

Another limitation lies in the lack of robustness to distribution shifts. Such shifts are quite common when data is aggregated across multiple sites. As a result, this may hinder the identification and characterization of clinically relevant brain biomarkers across heterogeneous populations. These are key points of consideration when studying dysfunction associated with clinical disorders.

7.3 Future Work

Keeping these points in mind, we identify three avenues of proposed research that may improve our understanding of brain connectivity and benefit clinical studies in the long term:

- **Multi-Site Representation Learning:** While open repositories such as ABIDE (Autism Brain Imaging Data Exchange) [154, 18] are becoming extremely popular, such studies pose their own set of unique challenges. For example, the subject demographics and scanning protocols is known to vary tremendously across sites. This variation introduces site-specific biases, which are difficult to account for and often confound neurobiological discovery. As of today, there are very few studies that go beyond case/control prediction on ABIDE. It is also unclear how to explicitly account for patient heterogeneity arising from site differences [219]. Such research questions are also fundamentally related to studies which seek to quantify, understand, and account for the effect of covariates and confounders [220] when evaluating prediction within heterogenous cohorts.

  As such, our models have been developed on focused clinical datasets and are not designed to handle the distribution shifts thus arising. A potential direction of exploration would be to include modelling constraints that incorporate site related information (as a second level of patient heterogeneity) directly within the framework and inference procedure. Such models may benefit from concepts within the stochastic optimization literature such as continual learning [221] or distribution alignment schemes [222] that are designed for parallel scenarios in other AI application domains.

- **Uncovering Nuanced Clinical Characterizations:** This thesis has largely focused on multidimensional phenotypic prediction, which is a challenging yet important clinical paradigm. Nevertheless, our prediction performance is far from perfect. One of the key reasons can be attributed to inherent noise in the severity measures derived from clinical reports themselves. For example, reporting across parent-teacher evaluations could have more variability when compared with measures scored by trained clinicians. At the same time, such phenotypic scores alone may not paint a complete clinical picture of such complex and heterogenous disorders.

  In an effort to build up a more holistic picture, one may benefit from examining more coarse indicators of behavior. Examples include studying ordered levels of impairment or dysfunction [223], or the prevalence of behavioral sub-types in disorders such as ASD or ADHD [224, 225], or co-morbidities among developmental disorders [226, 227], or nosological relationships between psychiatric disorders [228]. From a technical standpoint, these modeling extensions may borrow from clustering or ordinal regression to supplement our discriminative models. Other interesting avenues of exploration include models which uncover causal relationships between connectivity and disease phenotypes [229, 230].

- **Sophisticated Modeling of Brain Connectivity:** The mathematical models presented in this thesis restrict the analysis of brain connectivity onto a set of clinically predictive matrix factors. While this treatment ensures computational tractability, the actual complexity of interactions within the brain may not be sufficiently explored by such frameworks. Simple examples where such matrix/tensor factorization frameworks have shown promise is in characterizing the hierarchical [231] and modular [232] organization of the brain, which our models do not directly incorporate into the representation.

  Another scenario of great clinical interest is in the longitudinal modeling of brain connectivity [220, 233], which is a currently unexplored within this thesis. We envision that the ability to effectively track and forecast changes (disruptions) in brain connectivity patterns over disease progression could ultimately further our understanding of the associated pathogenesis.

In summary, our models make very minimal assumptions and can potentially be adapted to a wide variety of clinical and neuroscientific applications. The proposed modeling formulations could serve as powerful tools for brain connectivity analysis. We are confident that such explorations will go a long way to advance the field by providing novel insights into the organization of the human brain.

Chapter 8

Appendix: Graph Convolutional Frameworks for Multidimensional Phenotypic Prediction from Multimodal Connectivity Data

As seen in the chapters leading into this thesis, the rise of machine learning has prompted a shift in connectomics towards subject-level predictions. This shift has been accelerated by deep learning, which provides unparalleled representational power. In this section, we take an alternate strategy designed to model the complex topology of brain organization. We revert to the network-centric view of the brain and propose a geometric deep learning framework based on graph convolutions to map from the brain connectivity to the behavioral space.

As mentioned in previous chapters, the bulk of deep learning methods focus on diagnostic classification. These approaches range from Multi-Layered

Perceptrons [234], Deep Belief Networks [129], to Convolutional Neural Networks [43]. Methods to predict finer-grained characteristics (e.g, demographics or behavior) are sparser and largely focus on a single modality. For example, the authors of [44] introduced a convolutional neural network that mapped DTI connectivity matrices to cognitive and motor measures. The work of [200] proposes an artificial neural network for age prediction from structural connectomes. While these methods achieve good empirical performance, they ignore the interplay between structure and function in the brain.

Our work in [160] takes the alternative approach of combining a generative dictionary learning framework with a predictive artificial neural network to simultaneously map multiple clinical measures. However, it still focuses on functional connectivity alone. To address this gap, we extended [160] to combine dynamic rs-fMRI correlations with DTI tractography using a structurally-regularized matrix decomposition in [165, 164]. While promising, even this method relies on only immediate structural neighbourhood relationships to guide the representation learning (See Eq. (5.5)). Said another way, this method does not provide explicit control over the extent to which multi-hop (indirect) structural connections mediate functional connectivity.

8.1 Graph Neural Networks in Neuroimaging

Graph neural networks are designed to build representations of nodes and edges within graph structured data, and have found applications in a variety of domains where data naturally assumes a network-like organization

[235]. These architectures have shown great promise for modeling multi-stage interactions between brain regions that also reflect the hierarchy of brain organization. Hence, these techniques have become important tools in brain connectivity research. Examples include: modeling dynamic functional connectivity for groupwise discrimination [193], diagnosis of neurodevelopmental disorders [195, 196] from rs-fMRI correlation inputs, or structural connectivity modeling for disease classification [236]. However, most current approaches do not leverage the complementarity between the structural and functional graphs and examine dimensional measures of behavior beyond diagnostic classification.

Outline: This work describes a multimodal graph convolutional network (M-GCN) to integrate functional and structural connectivity from rs-fMRI and DTI data respectively, and map this information to phenotypic measures in Section 8.2. This employ specialized graph convolutional filters based on [237, 44] that operate on functional connectivity inputs, as guided by the subject-level structural graph topology. In Section 8.4, we demonstrate that our framework generalizes to prediction of phenotypic measures on two separate real world datasets and learns to extract predictive brain biomarkers from limited data. The work described here first appeared as a conference paper in [238].

8.2 Multimodal Graph Convolutional Framework

Fig. 8.1 illustrates our graph convolutional framework, which consists of a representation learning module on the connectomics data (Green Box) cascaded with a fully connected ANN for regression (Blue Box). Let N be the number of patients and P be the number of regions in our brain parcellation. Our framework first extracts the structural connectivity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_n)$ from DTI tractography. The nodes in $\mathcal{V}$ are brain ROIs defined by the parcellation, while the edges in $\{\mathcal{E}_n\}$ indicate the presence of at least one fiber tract between these regions. Let $\mathbf{A}_n \in \mathcal{R}^{P \times P}$ be the adjacency matrix for $\mathcal{G}$. Correspondingly, we assume that the functional connectivity profile is a signal that rides on the fixed graph montage and is given by rs-fMRI correlation matrices $\mathbf{\Gamma}_n \in \mathcal{R}^{P \times P}$.

Traditional convolutional layers assume a spatial contiguity of the input features, as in the case of 2-D images. This assumption breaks down in general graphs, as node orderings may be arbitrary. Thus, graph convolutional networks define a layer-wise propagation rule designed to aggregate information efficiently at each node based on the underlying graph topology [239, 237].

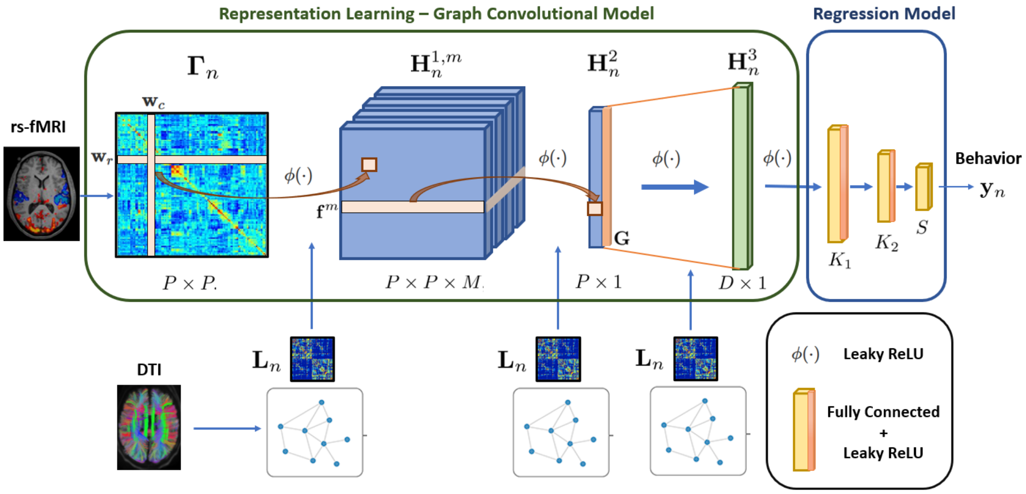


Figure 8.1: Our M-GCN framework for predicting phenotypic measures **Green Box:** Graph Convolutional Model for Representation Learning from Multimodal Connectomics Data. **Blue Box:** Fully Connected Artificial Neural Network to map to phenotypic measures.

For a generic input signal $\mathbf{X}^{l-1} \in \mathcal{R}^{P \times C_{l-1}}$, a graph filtering operation can be formulated as follows:

$$\mathbf{X}^{l} = \phi(\mathbf{L}\mathbf{X}^{l-1}\mathbf{W}) = \phi(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}^{l-1}\mathbf{W}) \tag{8.1}$$

$$\text{where} \quad \tilde{\mathbf{A}} = \mathcal{I}_P + \mathbf{A}; \; \tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij} \tag{8.2}$$

where $\mathbf{W} \in \mathcal{R}^{C_{l-1} \times C_l}$ denotes the filter weights, $\mathcal{I}_P$ is an identity matrix of dimension P, and $\mathbf{L} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}$ is the graph Laplacian of the reparameterized adjacency matrix $\tilde{\mathbf{A}}$ and degree matrix $\tilde{\mathbf{D}}$. The authors of [237] demonstrate that Eq. (8.2) is a first order approximation to spectral filtering in the graph Fourier domain.

Inspired by Eq. (8.2), we define a graph filtering operation that acts on the input functional connectivity matrix $\mathbf{\Gamma}_n$ to generate a connectivity embedding $\mathbf{H}_n^{1,m} \in \mathcal{R}^{P \times P}$ as follows:

$$\mathbf{H}_n^{1,m}(i,j) = \phi\Big((\mathbf{w}_r^m)^T \mathbf{L}_n \mathbf{\Gamma}_n(:,j) + \mathbf{\Gamma}_n(i,:)\mathbf{L}_n \mathbf{w}_c^m + \mathbf{b}^1\Big) \quad m \in \{1, \ldots M\} \tag{8.3}$$

Here, M is the number of channels, each parametrized by a row and column filter $\mathbf{w}_r^m, \mathbf{w}_c^m \in \mathcal{R}^{P \times 1}$ and a bias term $\mathbf{b}^1 \in \mathcal{R}^{P \times 1}$, resulting in a total of $(2P+1)$ learnable parameters per channel. Effective, $\mathbf{H}_n^{1,m}(i,j)$ computes a weighted sum of the functional connectivity profile of nodes i and j, further regularized by the DTI graph Laplacian $\mathbf{L}_n$. Conceptually, Eq. (8.3) is similar to the cross shaped E2E filters in [44]. We also note that, despite the symmetry of the correlation matrices $\mathbf{\Gamma}_n$, the embedding $\mathbf{H}_n^{1,m}$ can be assymmetric. This allows us to account for any laterality in functional subsystems.

Following the connectome embedding in Eq. (8.3), we use two more graph convolutional layers with pooling to first compute a node-wise representation $\mathbf{H}_n^2 \in \mathcal{R}^{P\times 1}$ and a whole-graph embedding $\mathbf{H}_n^3 \in \mathcal{R}^{D\times 1}$. Mathematically, these operations can be represented as:

$$\mathbf{H}_n^2 = \phi\Big(\sum_m \mathbf{L}_n\mathbf{H}_n^{1,m}\mathbf{f}^m + \mathbf{b}^2\Big) \qquad\qquad \mathbf{H}_n^3 = \phi\Big(\mathbf{G}\mathbf{L}_n\mathbf{H}_n^2 + \mathbf{b}^3\Big) \tag{8.4}$$

The filter weights are parameterized by the vectors $\mathbf{f}^m \in \mathcal{R}^{P\times 1}$ per M channel, the graph embedding matrix $\mathbf{G} \in \mathcal{R}^{D\times P}$, and the bias terms $\mathbf{b}^2$ and $\mathbf{b}^3$ respectively. In total, these layers add another $(M+D)P+2$ learnable parameters. Eq. (8.4) parallels the computation of centrality measures in graph theoretic literature by summarizing node-wise information based on functional similarity, as guided by structure. Finally, our graph embedding $\mathbf{H}_n^3$ is input to an ANN to map to the phenotypic measures $\mathbf{y}_n \in \mathcal{R}^{S\times 1}$ for patient n. The ANN is a simple three layered fully connected network of sizes $D\times K_1$, $K_1\times K_2$ and $K_2\times S$.

8.3 Model Evaluation

8.3.1 Implementation Details:

We train our M-GCN on a combination of ℓ_2 loss and ℓ_1 loss between the predicted $\hat{\mathbf{y}}_n$ and true measures $\mathbf{y}_n$:

$$\mathcal{L} = \frac{1}{NS}\sum_{n=1}^{N}\Big[||\mathbf{y}_n - \hat{\mathbf{y}}_n||_2 + ||\mathbf{y}_n - \hat{\mathbf{y}}_n||_1\Big] \tag{8.5}$$

The ℓ_1 loss function has been shown to be more robust to outliers as compared to the ℓ_2 loss [240], but less stable during training due to the lack of smoothness near the optimal solution [241]. We found that this combined loss empirically provided a good tradeoff between stability and generalization. Layer sizes for the M-GCN were set to $M = 32$ channels for the connectome embedding, $D = 256$ for the graph embedding and $\{K_1, K_2\} = 128, 30$, as we found these choices to be sufficient to map the connectomics data to the phenotypic measures during training. We chose a LeakyReLU ($\phi(x) = \max(0, x) + 0.1 * \min(0, x)$) as the activation function with our network layers, which we found empirically robust to saturation and exploding gradients during training. We train our M-GCN via stochastic gradient descent (SGD) algorithm with momentum ($\delta = 0.9$), batch size = 16, with an initial learning rate of 0.001 decayed by 0.9 every 10 epochs. Additionally, we utilize a weight decay of 0.001 as regularization and train our network for 40 epochs to avoid overfitting. All parameters were determined based on a validation set of 30 additional patients from the HCP dataset. We carried forward the same settings to the KKI dataset.

8.3.2 Baselines

We compare the predictive performance of our network against the following baselines:

Multimodal ANN: We use a four layer ANN that maintains the same number of parameters, activation, and loss function as the M-GCN. It operates on the vectorized $P \times (P-1)/2$ rs-fMRI correlations, each multiplied by the corresponding entry of the DTI Laplacian $\mathbf{L}_n$. This baseline evaluates the benefit of maintaining the graph structure of the data.

rs-fMRI only GCN: We use the same architecture as our M-GCN but omit the graph Laplacian in Eqs. (8.3-8.4). This baseline evaluates the benefit of DTI regularization.

BrainNetCNN: We integrate multimodal connectivity data via the BrainNetCNN [44], originally designed to predict cognitive outcomes from DTI data. We modify this architecture to have two branches, one for the rs-fMRI correlation matrices $\mathbf{\Gamma}_n$, and another for the DTI Laplacians $\mathbf{L}_n$. The ANN is modified to output S measures of clinical severity. We set the hyperparameters according to [44]

Dictionary Learning + ANN: The integrated framework in [160] uses static rs-fMRI correlation matrices ($\mathbf{\Gamma}_n$) to simultaneously predict multiple clinical or behavioral measures. The model combines a dictionary learning generative term with a neural network predictor. The two blocks are optimized jointly in an end-to-end fashion.

Dynamic Deep-Generative Hybrid: The framework in [164, 165] uses a similar joint optimization strategy but operates on dynamic rs-fMRI correlation matrices $\{\mathbf{\Gamma}_n^t\}$ and incorporates DTI regularizer in the dictionary learning term. Overall, these last two baselines evaluate the benefit of GCNs for implicit representational learning over a classical decomposition strategy. We have followed the guidelines provided by the authors to set the hyperparameters and train both of these baselines.

Meas.	Method	MAE Test	NMI Test
CFIS	Mult. ANN	14.06 ± 10.16	0.61
	rs-fMRI only GCN	14.16 ± 8.96	0.54
	BrainNetCNN	17.90 ± 17.55	0.58
	Dict. Learn. + ANN	15.26 ± 13.99	0.66
	Dyn. Deep-Gen. Hyb.	16.31 ± 15.43	0.67
	Our Framework	**12.87** ± **9.65**	**0.73**

Table 8.1: HCP Dataset: Evaluation using the **Median Absolute Error (MAE)**, **Normalized Mutual Information (NMI)** for the test set. Best performance is highlighted in bold.

8.4 Experimental Results

We validate this framework on 275 healthy individuals from the Human Connectome Project and our in-house ASD dataset to predict cognitive measures and behavioral deficits respectively. For both datasets, we again use the Automatic Anatomical Labeling (AAL) atlas [142] to define 116 cortical, sub-cortical and cerebellar brain ROIs for both the functional and structural connectivity matrices. We also subtract the first eigenvector from the rs-fMRI correlation matrices, which is a roughly constant bias, and use the residual matrices as the inputs to all models.

8.4.1 Population Studies

8.4.2 Predicting CFIS:

Table 8.1 (and Fig. 8.2) illustrates our method and baselines for predicting CFIS for the HCP dataset in a five-fold cross validated setting. We quantify the performance via the Median Absolute Error (MAE), the Normalized Mutual Information (NMI) between the actual and predicted measures. Lower MAE and higher NMI. indicate better performance. The training performance is good for all methods. However, the M-GCN clearly outperforms the baselines when generalizing to unseen testing data. As a benchmark, our validation performance (Test MAE: 13.41 ± 8.17, NMI Test: 0.71) also provides similar generalization.

8.4.3 Multidimensional Clinical Severity Prediction:

Table 8.2 (and Fig. 8.3) compares the multi-output prediction performance of ADOS, SRS, and Praxis on the KKI dataset for a five fold cross validation. Again, we observe that the M-GCN outperforms the baselines for the prediction of all three severity measures in almost every case. Note that, from a

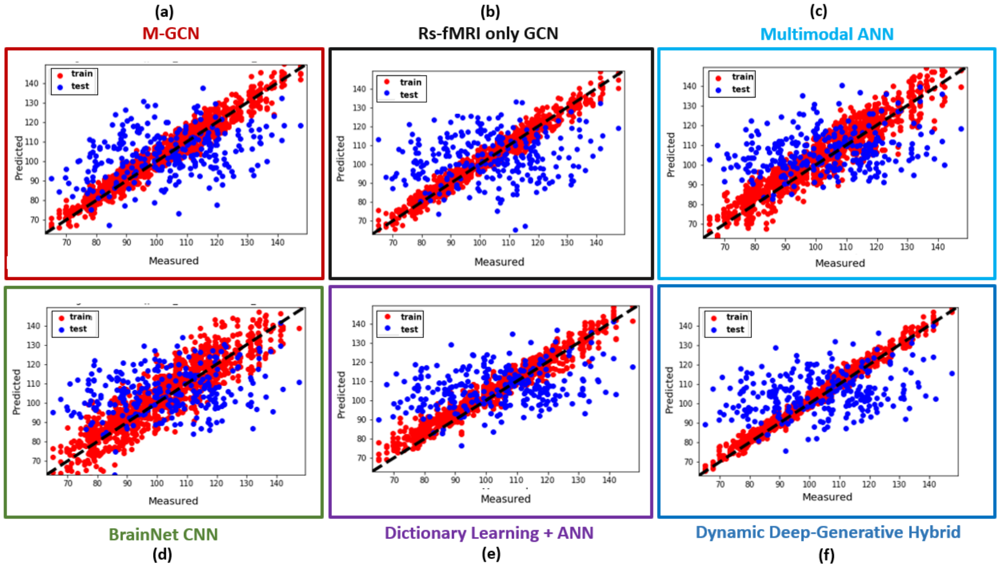


Figure 8.2: HCP Dataset: Prediction of Cognitive Fluid Intelligence Score by **(a) Red Box:** M-GCN **(b) Black Box:** rs-fMRI only GCN **(c) Light Blue Box:** Multimodal ANN **(d) Green Box:** BrainNet CNN **(e) Purple Box:** Dictionary Learning + ANN **(f) Dark Blue Box:** Dynamic Deep-Generative Hybrid

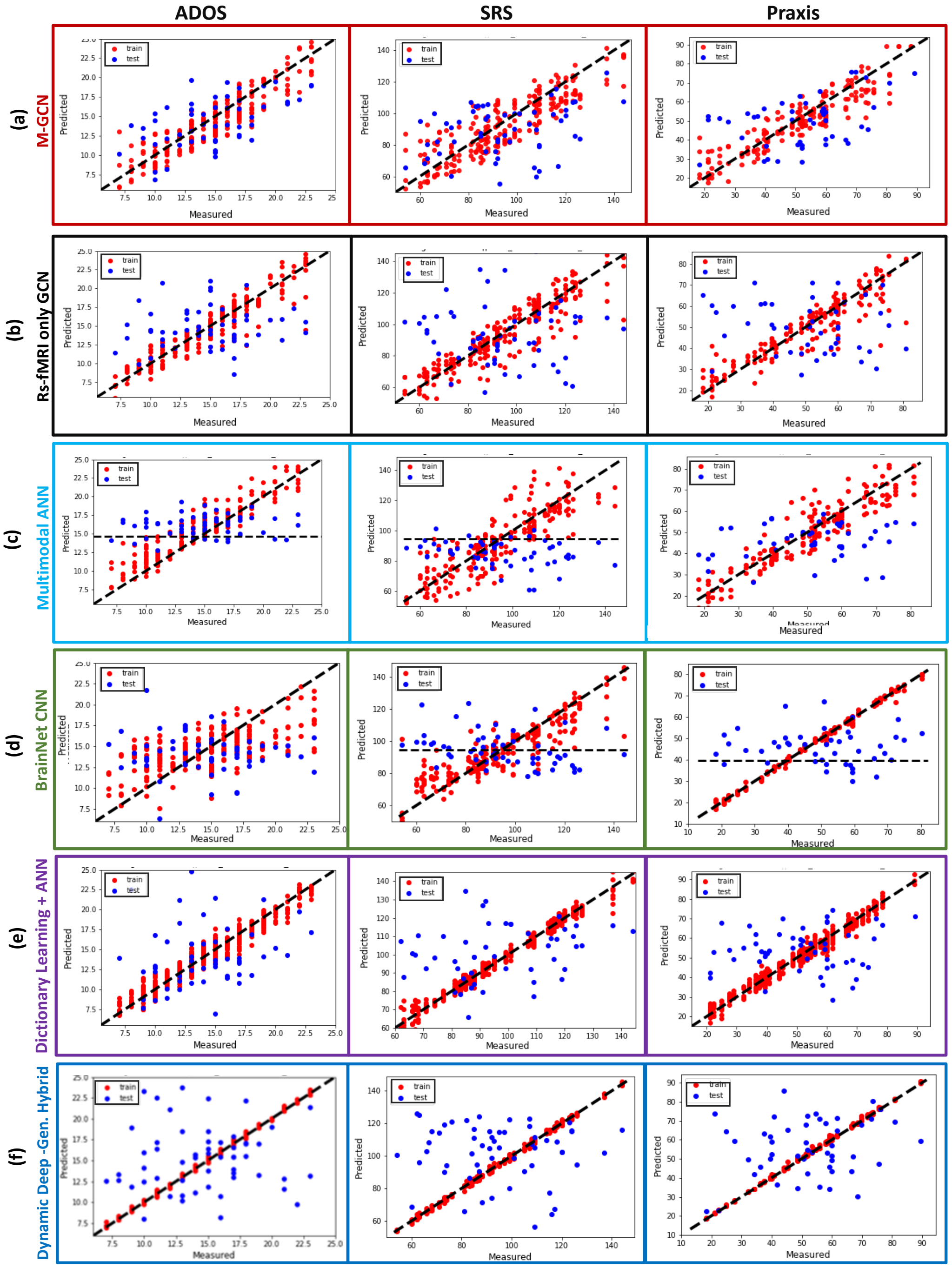


Figure 8.3: KKI Dataset: Multi-output prediction by **Left:** ADOS **Middle:** SRS **Right:** Praxis by **(a) Red Box:** M-GCN **(b) Black Box:** rs-fMRI only GCN **(c) Light Blue Box:** Multimodal ANN **(d) Green Box:** BrainNet CNN **(e) Purple Box:** Dictionary Learning + ANN **(f) Dark Blue Box:** Dynamic Deep-Generative Hybrid

Meas.	Method	MAE Test	NMI Test
ADOS	Mutl. ANN	2.96 ± 2.30	0.30
	rs-fMRI only GCN	3.14 ± 2.25	0.41
	BrainNetCNN	3.50 ± 2.20	0.25
	Dict. Learn. + ANN	**2.71** ± **2.40**	0.43
	Dyn. Deep-Gen. Hyb.	2.84 ± 2.79	0.34
	Our Framework	**2.71** ± **2.15**	**0.45**
SRS	Mult. ANN	18.47 ± 11.04	0.60
	rs-fMRI only GCN	21.34 ± 8.58	0.62
	BrainNetCNN	18.96 ± 15.65	0.75
	Dict. Learn. + ANN	16.79 ± 13.83	**0.89**
	Dyn. Deep-Gen. Hyb.	17.81 ± 16.09	<u>0.88</u>
	Our Framework	**16.50** ± **9.44**	<u>0.85</u>
Praxis	Mult. ANN	17.12 ± 16.66	0.65
	rs-fMRI only GCN	16.71 ± 16.66	0.74
	BrainNetCNN	15.15 ± 11.49	0.19
	Dict. Learn. + ANN	13.19 ± 10.75	0.82
	Dyn. Deep-Gen. Hyb.	13.50 ± 11.55	<u>0.85</u>
	Our Framework	**12.82** ± **12.04**	**0.86**

Table 8.2: KKI Dataset: Evaluation using the **Median Absolute Error (MAE), Normalized Mutual Information (NMI)**. Best performance is highlighted in bold. Near misses are underlined.

clinical standpoint, generalization to prediction of multiple deficits is inherently more challenging than predicting a single phenotypic measure. This also partially accounts for the poor performance of some of the baselines, where they perform reasonably well for the prediction of one of the measures (for example, the rs-fMRI only GCN for ADOS), but at the expense of generalization onto the other two measures. Overall, our experiments on two different real world datasets allude to reproducibility and suggest that the M-GCN generalizes effectively even with modest training sample sizes. Moreover, the performance gains against the M-GCN baseline without the DTI indicate the benefit provided by the multimodal integration via our graph convolutional framework.

8.4.4 Extracting Clinical Biomarkers:

The representations learned by the row and column filter pairs $\mathbf{w}_r$ and $\mathbf{w}_c$ at the input layer of the M-GCN (i.e. Eq. (8.3)) may illuminate key biomarkers for each population. We first match the filter pairs across the cross validation folds based on the average correlation coefficient between the row and column filter weights. Fig. 8.4 illustrates four filter pairs out of 32 that appear most frequently across subsets of the HCP and KKI dataset. In each case, we plot the average row filter (RF) and column filter (CF) weights projected onto the corresponding regions of the AAL atlas. Compared with the filters learned by the rs-fMRI only GCN (Fig. 8.5), the DTI regularization in the M-GCN offers sparsity and better spatial selectivity in the patterns captured. For the HCP dataset (Fig. 8.4 (a)), we observe that RF1, RF2, CF1 and CF2 display contributions from regions of the Default Mode Network (DMN), known to play a critical role in consolidating working memory [183] and is widely inferred within the resting state literature. RF3 and CF3 highlight regions of the Frontoparietal Network (FPN) and the Medial Prefrontal Network (MPN), believed to play a role in working memory, attention and decision making, which are associated with cognitive intelligence [186]. CF4 highlights regions from the Somatomotor Network (SMN) while RF4 includes subcortical and cerebellar regions. Together, these are believed to be important functional biomarkers of cognitive intelligence in literature [189]. For the KKI dataset (Fig. 8.4 (b)), we observe that RF1, CF1, CF2 and CF4 highlight areas from the

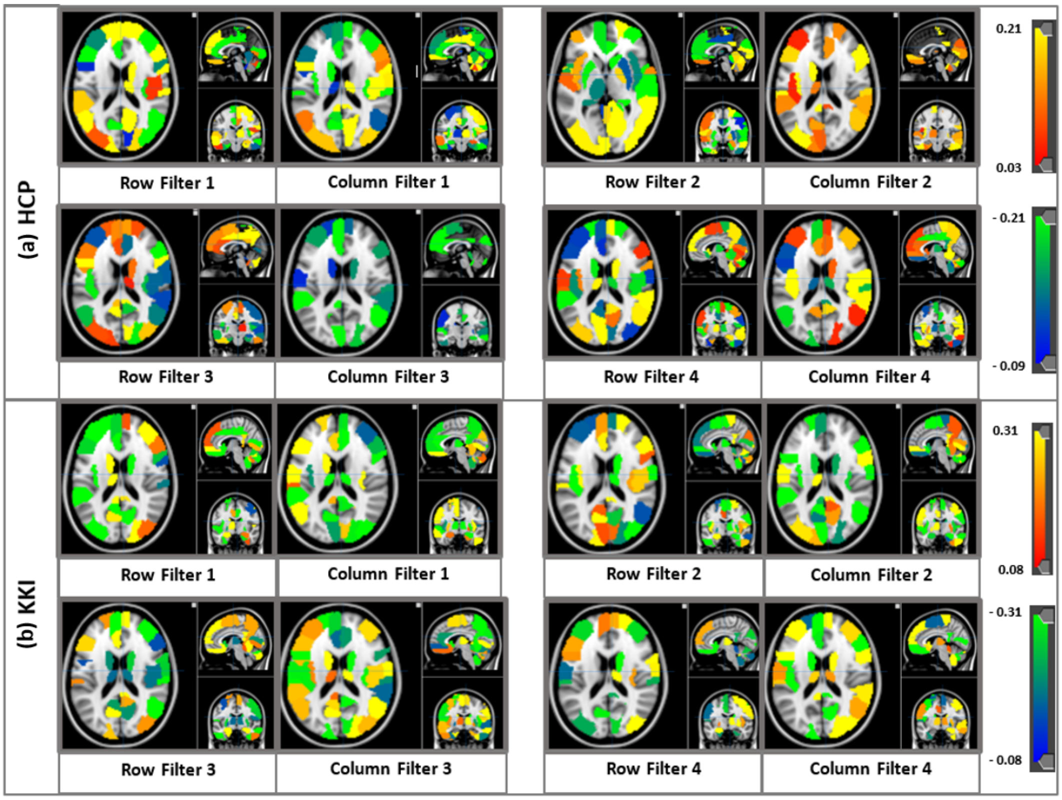


Figure 8.4: Four pairs of row & column filter weights learned by the M-GCN on the (a) HCP dataset and (b) KKI dataset. The colorbar quantifies the filter weight for each AAL ROI.

DMN and SMN. Altered connectivity within these regions is widely reported in ASD literature [98]. RF3, RF4 and CF4 also highlight contributions from the higher order visual processing areas and sensorimotor regions, which are in line with findings of reduced visual motor integration in Autism [98]. RF3, RF4 and CF4 also display contributions from subcortical regions along with the prefrontal cortex and DMN, which is believed to be relevant to social-emotional regulation in ASD [191].

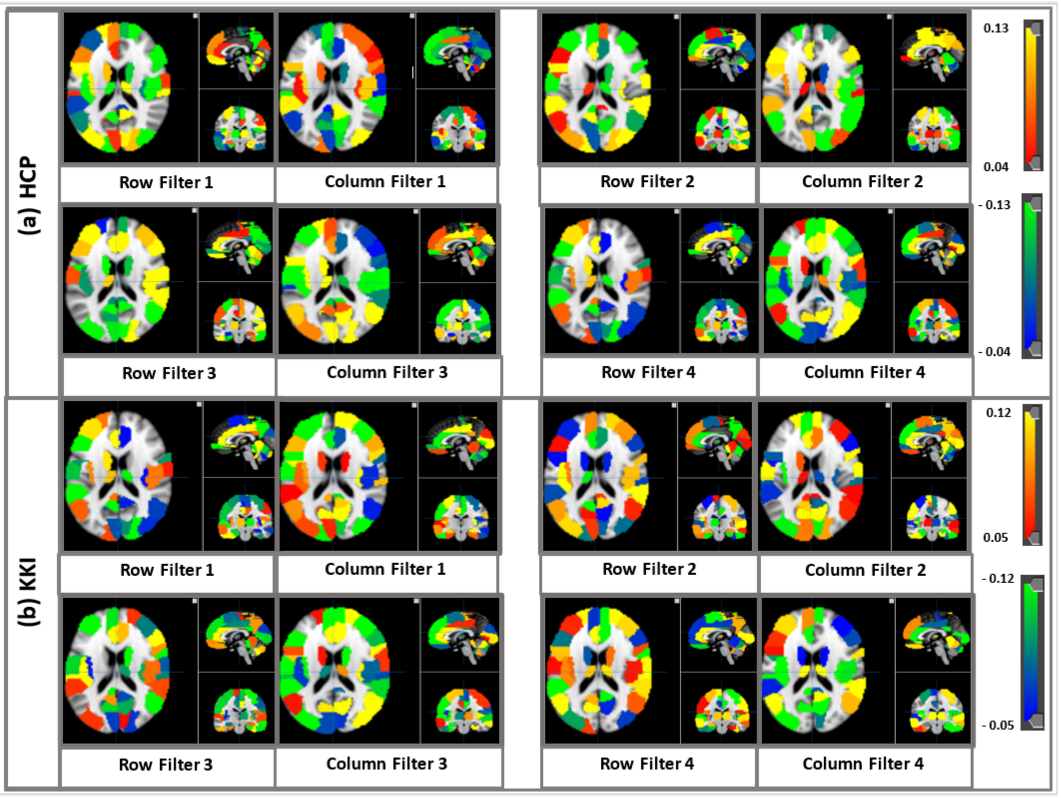


Figure 8.5: Four pairs of row & column filter weights learned by the rs-fMRI only GCN on the (a) HCP dataset (b) KKI dataset. The colorbar quantifies the filter weight for each AAL ROI. In contrast, the patterns in Fig. 3 are sparser and display lesser overlap across filters

8.5 Summary

This work introduces a novel multimodal graph convolutional framework to leverage complementary information from functional and structural connectivity. Our M-GCN is designed to effectively utilize the underlying anatomical pathways to learn rich representations from functional connectivity data that are simultaneously informative of multidimensional phenotypic characterizations. We demonstrate that this framework is able to learn effectively from limited training data and generalize well to unseen patients. Finally, our framework makes minimal assumptions, and can potentially be applied to study other neuro-psychiatric disorders (eg. ADHD, Schizophrenia) as a diagnostic tool.

References

[1] Beomsue Kim, Hongmin Kim, Songhui Kim, and Young-ran Hwang. "A brief review of non-invasive brain imaging technologies and the near-infrared optical bioimaging". In: *Applied Microscopy* 51.1 (2021), pp. 1–10.

[2] James C Bezdek, LO Hall, and L_P Clarke. "Review of MR image segmentation techniques using pattern recognition". In: *Medical physics* 20.4 (1993), pp. 1033–1048.

[3] Andrea Mechelli, Cathy J Price, Karl J Friston, and John Ashburner. "Voxel-based morphometry of the human brain: methods and applications". In: *Current Medical Imaging* 1.2 (2005), pp. 105–113.

[4] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. "Statistical parametric maps in functional imaging: a general linear approach". In: *Human brain mapping* 2.4 (1994), pp. 189–210.

[5] Yaniv Assaf and Ofer Pasternak. "Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review". In: *Journal of molecular neuroscience* 34.1 (2008), pp. 51–61.

[6] Saad Jbabdi and Heidi Johansen-Berg. "Tractography: where do we go from here?" In: *Brain connectivity* 1.3 (2011), pp. 169–183.

[7] Michael D Fox and Marcus E Raichle. "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging". In: *Nat. rev. neuro.* 8.9 (2007), pp. 700–711.

[8] Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI". In: *Magnetic resonance in medicine* 34.4 (1995), pp. 537–541.

[9] Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. "The brain's default network: anatomy, function, and relevance to disease." In: (2008).

[10] Pawel Skudlarski, Kanchana Jagannathan, Vince D Calhoun, Michelle Hampson, Beata A Skudlarska, and Godfrey Pearlson. "Measuring brain connectivity: diffusion tensor imaging validates resting state temporal correlations". In: *Neuroimage* 43.3 (2008), pp. 554–561.

[11] Christopher J Honey et al. "Predicting human resting-state functional connectivity from structural connectivity". In: *Proc. of the Nat. Acad. of Sci.* 106.6 (2009), pp. 2035–2040.

[12] Makoto Fukushima et al. "Structure–function relationships during segregated and integrated network states of human brain functional connectivity". In: *Brain Structure and Function* 223.3 (2018), pp. 1091–1106.

[13] F DuBois Bowman, Lijun Zhang, Gordana Derado, and Shuo Chen. "Determining functional connectivity using fMRI data with diffusion-based anatomical weighting". In: *NeuroImage* 62.3 (2012), pp. 1769–1779.

[14] Selen Atasoy, Isaac Donnelly, and Joel Pearson. "Human brain networks function in connectome-specific harmonic waves". In: *Nature communications* 7.1 (2016), pp. 1–10.

[15] Arnaud Messé et al. "Predicting functional connectivity from structural connectivity via computational models using MRI: an extensive comparison study". In: *NeuroImage* 111 (2015), pp. 65–75.

[16] Shu-Hsien Chu et al. "Function-specific and enhanced brain structural connectivity mapping via joint modeling of diffusion and functional MRI". In: *Sci. Rep.* 8.1 (2018).

[17] Lu Zhang, Li Wang, and Dajiang Zhu. "Recovering Brain Structural Connectivity from Functional Connectivity via Multi-GCN Based Generative Adversarial Network". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 53–61.

[18] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism". In: *Molecular psychiatry* 19.6 (2014), pp. 659–667.

[19] Meng Liang, Yuan Zhou, Tianzi Jiang, Zhening Liu, Lixia Tian, Haihong Liu, and Yihui Hao. "Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging". In: *Neuroreport* 17.2 (2006), pp. 209–213.

[20] Steven M Stufflebeam, Hesheng Liu, Jorge Sepulcre, Naoaki Tanaka, Randy L Buckner, and Joseph R Madsen. "Localization of focal epileptic discharges using functional connectivity magnetic resonance imaging". In: *Journal of neurosurgery* 114.6 (2011), pp. 1693–1697.

[21] Pierre Besson, Vera Dinkelacker, Romain Valabregue, Lionel Thivard, Xavier Leclerc, Michel Baulac, Daniela Sammler, Olivier Colliot, Stéphane Lehéricy, Séverine Samson, et al. "Structural connectivity differences in left and right temporal lobe epilepsy". In: *Neuroimage* 100 (2014), pp. 135–144.

[22] Charreau S Bell. *Seed-based correlation analysis and instantaneous global correlation analysis for resting state fMRI*. Vanderbilt University, 2018.

[23] Klaus Hahn et al. "Selectively and progressively disrupted structural connectivity of functional brain networks in Alzheimer's disease—revealed by a novel framework to analyze edge distributions of networks detecting disruptions with strong statistical evidence". In: *Neuroimage* 81 (2013), pp. 96–109.

[24] Jennifer L Whitwell, Ramesh Avula, Ankit Master, Prashanthi Vemuri, Matthew L Senjem, David T Jones, Clifford R Jack Jr, and Keith A Josephs. "Disrupted thalamocortical connectivity in PSP: a resting-state fMRI, DTI, and VBM study". In: *Parkinsonism & related disorders* 17.8 (2011), pp. 599–605.

[25] Daniel J Goble, James P Coxon, Annouchka Van Impe, Monique Geurts, Wim Van Hecke, Stefan Sunaert, Nicole Wenderoth, and Stephan P Swinnen. "The neural basis of central proprioceptive processing in older versus younger adults: an important sensory role for right putamen". In: *Human brain mapping* 33.4 (2012), pp. 895–908.

[26] Jessica R Andrews-Hanna, Abraham Z Snyder, Justin L Vincent, Cindy Lustig, Denise Head, Marcus E Raichle, and Randy L Buckner. “Disruption of large-scale brain systems in advanced aging”. In: *Neuron* 56.5 (2007), pp. 924–935.

[27] Ruth E Propper, Lauren J O’Donnell, Stephen Whalen, Yanmei Tie, Isaiah H Norton, Ralph O Suarez, Lilla Zollei, Alireza Radmanesh, and Alexandra J Golby. “A combined fMRI and DTI examination of functional language lateralization and arcuate fasciculus structure: effects of degree versus direction of hand preference”. In: *Brain and cognition* 73.2 (2010), pp. 85–92.

[28] Olaf Sporns, Dante R Chialvo, Marcus Kaiser, and Claus C Hilgetag. “Organization, development and function of complex brain networks”. In: *Trends in cognitive sciences* 8.9 (2004), pp. 418–425.

[29] Mikail Rubinov and Olaf Sporns. “Complex network measures of brain connectivity: uses and interpretations”. In: *Neuroimage* 52.3 (2010), pp. 1059–1069.

[30] Ed Bullmore and Olaf Sporns. “Complex brain networks: graph theoretical analysis of structural and functional systems”. In: *Nature Reviews Neuroscience* 10.3 (2009), p. 186.

[31] Alex Fornito, Edward T Bullmore, and Andrew Zalesky. “Opportunities and challenges for psychiatry in the connectomic era”. In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 2.1 (2017), pp. 9–19.

[32] David Rolnick and Eva L Dyer. “Generative models and abstractions for large-scale neuroanatomy datasets”. In: *Current opinion in neurobiology* 55 (2019), pp. 112–120.

[33] CY Tang, EL Eaves, JC Ng, DM Carpenter, X Mai, DH Schroeder, CA Condon, R Colom, and RJ Haier. “Brain networks for working memory and factors of intelligence assessed in males and females with fMRI and DTI”. In: *Intelligence* 38.3 (2010), pp. 293–303.

[34] Dajiang Zhu, Tuo Zhang, Xi Jiang, Xintao Hu, Hanbo Chen, Ning Yang, Jinglei Lv, Junwei Han, Lei Guo, and Tianming Liu. “Fusing DTI and fMRI data: a survey of methods and applications”. In: *NeuroImage* 102 (2014), pp. 184–191.

[35] Marcus E Raichle. “The brain’s default mode network”. In: *Annual review of neuroscience* 38 (2015), pp. 433–447.

[36] Archana Venkataraman, Nicholas Wymbs, Mary Beth Nebel, and Stewart Mostofsky. “A Unified Bayesian Approach to Extract Network-Based Functional Differences from a Heterogeneous Patient Cohort”. In: *International Workshop on Connectomics in Neuroimaging*. Springer. 2017, pp. 60–69.

[37] Meenakshi Khosla, Keith Jamison, Gia H Ngo, Amy Kuceyeski, and Mert R Sabuncu. “Machine learning in resting-state fMRI analysis”. In: *Magnetic resonance imaging* 64 (2019), pp. 101–121.

[38] Simon Wein, Gustavo Deco, Ana Maria Tomé, Markus Goldhacker, Wilhelm M Malloni, Mark W Greenlee, and Elmar W Lang. “Brain Connectivity Studies on Structure-Function Relationships: A Short Survey with an Emphasis on Machine Learning”. In: *Computational Intelligence and Neuroscience* 2021 (2021).

[39] NS D’Souza, MB Nebel, N Wymbs, SH Mostofsky, and A Venkataraman. “A joint network optimization framework to predict clinical severity from resting state functional MRI data”. In: *NeuroImage* 206 (2020), p. 116314.

[40] Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry J Bockholt, Jeffrey D Long, Hans J Johnson, Jane S Paulsen, Jessica A Turner, and Vince D Calhoun. “Deep learning for neuroimaging: a validation study”. In: *Frontiers in neuroscience* 8 (2014), p. 229.

[41] Junghoe Kim, Vince D Calhoun, Eunsoo Shim, and Jong-Hwan Lee. “Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia”. In: *Neuroimage* 124 (2016), pp. 127–146.

[42] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. “Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction”. In: *NeuroImage* 199 (2019), pp. 651–662.

[43] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert R Sabuncu. “3D convolutional neural networks for classification of functional connectomes”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 137–145.

[44] Jeremy Kawahara et al. "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment". In: *NeuroImage* 146 (2017), pp. 1038–1049.

[45] Asela Gunawardana, William Byrne, and Michael I Jordan. "Convergence Theorems for Generalized Alternating Minimization Procedures." In: *Journal of machine learning research* 6.12 (2005).

[46] Vince D Calhoun, Tülay Adali, Michael C Stevens, Kent A Kiehl, and James J Pekar. "Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ICA analysis". In: *Neuroimage* 25.2 (2005), pp. 527–538.

[47] Roland Norbert Boubela, Klaudius Kalcher, Wolfgang Huf, Claudia Kronnerwetter, Peter Filzmoser, and Ewald Moser. "Beyond noise: using temporal ICA to extract meaningful information from high-frequency fMRI signal fluctuations during rest". In: *Frontiers in human neuroscience* 7 (2013), p. 168.

[48] Shaojie Chen, Lei Huang, Huitong Qiu, Mary Beth Nebel, Stewart H Mostofsky, James J Pekar, Martin A Lindquist, Ani Eloyan, and Brian S Caffo. "Parallel group independent component analysis for massive fMRI data sets". In: *PloS one* 12.3 (2017), e0173496.

[49] Christian F Beckmann, Clare E Mackay, Nicola Filippini, Stephen M Smith, et al. "Group comparison of resting-state FMRI data using multi-subject ICA and dual regression". In: *Neuroimage* 47.Suppl 1 (2009), S148.

[50] Harini Eavani, Theodore D Satterthwaite, Raquel E Gur, Ruben C Gur, and Christos Davatzikos. "Unsupervised learning of functional network dynamics in resting state fMRI". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2013, pp. 426–437.

[51] Guifeng Xu, Lane Strathearn, Buyun Liu, and Wei Bao. "Prevalence of autism spectrum disorder among US children and adolescents, 2014-2016". In: *Jama* 319.1 (2018), pp. 81–82.

[52] Roger J Jou, Sarah J Paterson, Xenophon Papademetris, Lawrence H Staib, and Robert T Schultz. "Abnormalities in white matter structure in autism spectrum disorders detected by diffusion tensor imaging". In: *Neuroscience Research* 58 (2007), S62.

[53] Brittany G Travers, Nagesh Adluru, Chad Ennis, Do PM Tromp, Dan Destiche, Sam Doran, Erin D Bigler, Nicholas Lange, Janet E Lainhart, and Andrew L Alexander. "Diffusion tensor imaging in autism spectrum disorder: a review". In: *Autism Research* 5.5 (2012), pp. 289–313.

[54] Sarah K Noonan, Frank Haist, and Ralph-Axel Müller. "Aberrant functional connectivity in autism: evidence from low-frequency BOLD signal fluctuations". In: *Brain research* 1262 (2009), pp. 48–63.

[55] Hideya Koshino, Patricia A Carpenter, Nancy J Minshew, Vladimir L Cherkassky, Timothy A Keller, and Marcel Adam Just. "Functional connectivity in an fMRI working memory task in high-functioning autism". In: *Neuroimage* 24.3 (2005), pp. 810–821.

[56] Marcel Adam Just, Vladimir L Cherkassky, Timothy A Keller, and Nancy J Minshew. "Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity". In: *Brain* 127.8 (2004), pp. 1811–1821.

[57] Daniel H Geschwind and Pat Levitt. "Autism spectrum disorders: developmental disconnection syndromes". In: *Current opinion in neurobiology* 17.1 (2007), pp. 103–111.

[58] Chris Oliver, Katy Berg, Jo Moss, Kate Arron, and Cheryl Burbidge. "Delineation of behavioral phenotypes in genetic syndromes: characteristics of autism spectrum disorder, affect and hyperactivity". In: *Journal of autism and developmental disorders* 41.8 (2011), pp. 1019–1032.

[59] Karoline Alexandra Havdahl, Vanessa Hus Bal, Marisela Huerta, Andrew Pickles, Anne-Siri Øyen, Camilla Stoltenberg, Catherine Lord, and Somer L Bishop. "Multidimensional influences on autism symptom measures: implications for use in etiological research". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 55.12 (2016), pp. 1054–1063.

[60] Seok-Jun Hong, Sofie L Valk, Adriana Di Martino, Michael P Milham, and Boris C Bernhardt. "Multidimensional neuroanatomical subtyping of autism spectrum disorder". In: *Cerebral Cortex* 28.10 (2018), pp. 3578–3588.

[61] Thomas R Insel. "The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry". In: *American Journal of Psychiatry* 171.4 (2014), pp. 395–397.

[62] Seiji Ogawa, Tso-Ming Lee, Alan R Kay, and David W Tank. “Brain magnetic resonance imaging with contrast dependent on blood oxygenation”. In: *proceedings of the National Academy of Sciences* 87.24 (1990), pp. 9868–9872.

[63] David J Heeger and David Ress. “What does fMRI tell us about neuronal activity?” In: *Nature Reviews Neuroscience* 3.2 (2002), pp. 142–151.

[64] Martin M Monti. “Statistical analysis of fMRI time-series: a critical review of the GLM approach”. In: *Frontiers in human neuroscience* 5 (2011), p. 28.

[65] Nancy J Minshew and Timothy A Keller. “The nature of brain dysfunction in autism: functional brain imaging studies”. In: *Current opinion in neurology* 23.2 (2010), p. 124.

[66] George Bush, Eve M Valera, and Larry J Seidman. “Functional neuroimaging of attention-deficit/hyperactivity disorder: a review and suggested future directions”. In: *Biological psychiatry* 57.11 (2005), pp. 1273–1284.

[67] Margaret A Niznikiewicz, Marek Kubicki, and Martha E Shenton. “Recent structural and functional imaging findings in schizophrenia”. In: *Current Opinion in Psychiatry* 16.2 (2003), pp. 123–147.

[68] A Yoshino, Y Okamoto, G Okada, M Takamura, N Ichikawa, C Shibasaki, S Yokoyama, M Doi, R Jinnin, H Yamashita, et al. “Changes in resting-state brain networks after cognitive–behavioral therapy for chronic pain”. In: *Psychological Medicine* 48.7 (2018), pp. 1148–1156.

[69] Cristina Rosazza, Domenico Zacà, and Maria G Bruzzone. “Pre-surgical brain mapping: to rest or not to rest?” In: *Frontiers in neurology* 9 (2018), p. 520.

[70] Alex Fornito, Ben J Harrison, Emmeline Goodby, Anna Dean, Cinly Ooi, Pradeep J Nathan, Belinda R Lennox, Peter B Jones, John Suckling, and Edward T Bullmore. “Functional dysconnectivity of corticostriatal circuitry as a risk phenotype for psychosis”. In: *JAMA psychiatry* 70.11 (2013), pp. 1143–1151.

[71] Yao Yu, Dong-Yi Lan, Li-Ying Tang, Ting Su, Biao Li, Nan Jiang, Rong-Bin Liang, Qian-Min Ge, Qiu-Yu Li, and Yi Shao. “Intrinsic functional connectivity alterations of the primary visual cortex in patients with proliferative diabetic retinopathy: A seed-based resting-state fMRI

study". In: *Therapeutic Advances in Endocrinology and Metabolism* 11 (2020), pp. 204 –2018820960296.

[72] Hussam Metwali and Amir Samii. "Seed-based connectivity analysis of resting-state fMRI in patients with brain Tumors: a feasibility study". In: *World neurosurgery* 128 (2019), e165–e176.

[73] Stephanie Noble, Dustin Scheinost, Emily S Finn, Xilin Shen, Xenophon Papademetris, Sarah C McEwen, Carrie E Bearden, Jean Addington, Bradley Goodyear, Kristin S Cadenhead, et al. "Multisite reliability of MR-based functional connectivity". In: *Neuroimage* 146 (2017), pp. 959–970.

[74] Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. "Building better biomarkers: brain models in translational neuroimaging". In: *Nature neuroscience* 20.3 (2017), p. 365.

[75] AP Holmes and KJ Friston. "Generalisability, random E ects & population inference". In: *Neuroimage* 7 (1998), S754.

[76] Djalel Eddine Meskaldji, Marie-Christine Ottet, Leila Cammoun, Patric Hagmann, Reto Meuli, Stephan Eliez, Jean Philippe Thiran, and Stephan Morgenthaler. "Adaptive strategy for the statistical analysis of connectomes". In: *PloS one* 6.8 (2011), e23009.

[77] Andrew Zalesky, Alex Fornito, and Edward T Bullmore. "Network-based statistic: identifying differences in brain networks". In: *Neuroimage* 53.4 (2010), pp. 1197–1207.

[78] Cedric E Ginestet and Andrew Simmons. "Statistical parametric network analysis of functional connectivity dynamics during a working memory task". In: *Neuroimage* 55.2 (2011), pp. 688–704.

[79] Andrew Zalesky, Luca Cocchi, Alex Fornito, Micah M Murray, and ED Bullmore. "Connectivity differences in brain networks". In: *Neuroimage* 60.2 (2012), pp. 1055–1062.

[80] Martha D Kaiser, Caitlin M Hudac, Sarah Shultz, Su Mei Lee, Celeste Cheung, Allison M Berken, Ben Deen, Naomi B Pitskel, Daniel R Sugrue, Avery C Voos, et al. "Neural signatures of autism". In: *Proceedings of the National Academy of Sciences* (2010), p. 201010412.

[81] Gina Rippon, Jon Brock, Caroline Brown, and Jill Boucher. "Disordered connectivity in the autistic brain: challenges for the 'new psychophysiology'". In: *International journal of psychophysiology* 63.2 (2007), pp. 164–172.

[82] Danielle Smith Bassett and ED Bullmore. “Small-world brain networks”. In: *The neuroscientist* 12.6 (2006), pp. 512–523.

[83] Yong Liu, Meng Liang, Yuan Zhou, Yong He, Yihui Hao, Ming Song, Chunshui Yu, Haihong Liu, Zhening Liu, and Tianzi Jiang. “Disrupted small-world networks in schizophrenia”. In: *Brain* 131.4 (2008), pp. 945–961.

[84] Ernesto J Sanz-Arigita, Menno M Schoonheim, Jessica S Damoiseaux, Serge ARB Rombouts, Erik Maris, Frederik Barkhof, Philip Scheltens, and Cornelis J Stam. “Loss of ‘small-world’ networks in Alzheimer’s disease: graph analysis of FMRI resting-state functional connectivity”. In: *PloS one* 5.11 (2010), e13788.

[85] Roger Guimera and Luis A Nunes Amaral. “Functional cartography of complex metabolic networks”. In: *nature* 433.7028 (2005), pp. 895–900.

[86] Alex Fornito, Andrew Zalesky, and Edward T Bullmore. “Network scaling effects in graph analytic studies of human resting-state FMRI data”. In: *Frontiers in systems neuroscience* 4 (2010), p. 22.

[87] Giampiero Bardella, Angelo Bifone, Andrea Gabrielli, Alessandro Gozzi, and Tiziano Squartini. “Hierarchical organization of functional connectivity in the mouse brain: a complex network approach”. In: *Scientific reports* 6 (2016), p. 32060.

[88] Archana Venkataraman, Marek Kubicki, and Polina Golland. “From connectivity models to region labels: identifying foci of a neurological disorder”. In: *IEEE transactions on medical imaging* 32.11 (2013), pp. 2078–2098.

[89] Archana Venkataraman, Daniel Y-J Yang, Kevin A Pelphrey, and James S Duncan. “Bayesian community detection in the space of group-level functional differences”. In: *IEEE transactions on medical imaging* 35.8 (2016), pp. 1866–1882.

[90] Archana Venkataraman, Marek Kubicki, and Polina Golland. “From brain connectivity models to identifying foci of a neurological disorder”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 715–722.

[91] Archana Venkataraman, James S Duncan, Daniel Y-J Yang, and Kevin A Pelphrey. “An unbiased Bayesian approach to functional connectomics implicates social-communication networks in autism”. In: *NeuroImage: Clinical* 8 (2015), pp. 356–366.

[92] Andrew Sweet, Archana Venkataraman, Steven M Stufflebeam, Hesheng Liu, Naoro Tanaka, Joseph Madsen, and Polina Golland. "Detecting epileptic regions based on global brain connectivity patterns". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 98–105.

[93] Hariharan Ravishankar, Radhika Madhavan, Rakesh Mullick, Teena Shetty, Luca Marinelli, and Suresh E Joel. "Recursive feature elimination for biomarker discovery in resting-state functional connectivity". In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE. 2016, pp. 4071–4074.

[94] Seok-Jun Hong, Sofie L Valk, Adriana Di Martino, Michael P Milham, and Boris C Bernhardt. "Multidimensional Neuroanatomical Subtyping of Autism Spectrum Disorder". In: *Cerebral Cortex* (2017), pp. 1–11.

[95] Gagan S Sidhu, Nasimeh Asgarian, Russell Greiner, and Matthew RG Brown. "Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD". In: *Frontiers in systems neuroscience* 6 (2012), p. 74.

[96] Lucina Q Uddin, Kaustubh Supekar, Charles J Lynch, Amirah Khouzam, Jennifer Phillips, Carl Feinstein, Srikanth Ryali, and Vinod Menon. "Salience network–based classification and prediction of symptom severity in children with autism". In: *JAMA psychiatry* 70.8 (2013), pp. 869–879.

[97] Christine Ecker, Vanessa Rocha-Rego, Patrick Johnston, Janaina Mourao-Miranda, Andre Marquand, Eileen M Daly, Michael J Brammer, Clodagh Murphy, Declan G Murphy, MRC AIMS Consortium, et al. "Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach". In: *Neuroimage* 49.1 (2010), pp. 44–56.

[98] Mary Beth Nebel, Ani Eloyan, Carrie A Nettles, Kristie L Sweeney, Katarina Ament, Rebecca E Ward, Ann S Choe, Anita D Barber, James J Pekar, and Stewart H Mostofsky. "Intrinsic visual-motor synchrony correlates with social deficits in autism". In: *Biological psychiatry* 79.8 (2016), pp. 633–641.

[99] Mehdi Rahim, Bertrand Thirion, Danilo Bzdok, Ir'ene Buvat, and Gaël Varoquaux. "Joint prediction of multiple scores captures better individual traits from brain images". In: *NeuroImage* 158 (2017), pp. 145–154.

[100] Joana Cabral, Morten L Kringelbach, and Gustavo Deco. "Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: Models and mechanisms". In: *NeuroImage* 160 (2017), pp. 84–96.

[101] Barnaly Rashid, Eswar Damaraju, Godfrey D Pearlson, and Vince D Calhoun. "Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects". In: *Frontiers in human neuroscience* 8 (2014), p. 897.

[102] True Price, Chong-Yaw Wee, Wei Gao, and Dinggang Shen. "Multiple-network classification of childhood autism using functional connectivity dynamics". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 177–184.

[103] Biao Cai, Pascal Zille, Julia M Stephen, Tony W Wilson, Vince D Calhoun, and Yu Ping Wang. "Estimation of dynamic sparse connectivity patterns from resting state fMRI". In: *IEEE transactions on medical imaging* 37.5 (2017), pp. 1224–1234.

[104] Heather Shappell, Brian S Caffo, James J Pekar, and Martin A Lindquist. "Improved state change estimation in dynamic functional connectivity using hidden semi-Markov models". In: *NeuroImage* 191 (2019), pp. 243–257.

[105] Robert Engle. "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models". In: *Journal of Business & Economic Statistics* 20.3 (2002), pp. 339–350.

[106] Martin Lindquist. "Dynamic Connectivity: Pitfalls and Promises". In: (2016).

[107] Martin A Lindquist, Yuting Xu, Mary Beth Nebel, and Brain S Caffo. "Evaluating dynamic bivariate correlations in resting-state fMRI: a comparison study and a new approach". In: *NeuroImage* 101 (2014), pp. 531–546.

[108] Massimiliano Caporin and Michael McAleer. "Ten things you should know about the dynamic conditional correlation representation". In: *Econometrics* 1.1 (2013), pp. 115–126.

[109] Gian Piero Aielli. "Dynamic conditional correlation: on properties and estimation". In: *Journal of Business & Economic Statistics* 31.3 (2013), pp. 282–299.

[110] Limin Xia and Zhenmin Li. "A new method of abnormal behavior detection using LSTM network with temporal attention mechanism". In: *The Journal of Supercomputing* 77.4 (2021), pp. 3223–3241.

[111] Marlies E Vissers, Michael X Cohen, and Hilde M Geurts. "Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links". In: *Neuroscience & Biobehavioral Reviews* 36.1 (2012), pp. 604–625.

[112] Lisa Weyandt, Anthony Swentosky, and Bergljot Gyda Gudmundsdottir. "Neuroimaging and ADHD: fMRI, PET, DTI findings, and methodological limitations". In: *Developmental neuropsychology* 38.4 (2013), pp. 211–225.

[113] SD Roosendaal, Jeroen JG Geurts, Hugo Vrenken, HE Hulst, Keith S Cover, JA Castelijns, Petra JW Pouwels, and Frederik Barkhof. "Regional DTI differences in multiple sclerosis patients". In: *Neuroimage* 44.4 (2009), pp. 1397–1403.

[114] Talia M Nir, Neda Jahanshad, Julio E Villalon-Reina, Arthur W Toga, Clifford R Jack, Michael W Weiner, Paul M Thompson, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. "Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging". In: *NeuroImage: clinical* 3 (2013), pp. 180–195.

[115] MARIKA Urbanski, M Thiebaut De Schotten, S Rodrigo, C Oppenheim, E Touzé, J-F Méder, K Moreau, C Loeper-Jeny, B Dubois, and Paolo Bartolomeo. "DTI-MR tractography of white matter damage in stroke patients with neglect". In: *Experimental brain research* 208.4 (2011), pp. 491–505.

[116] Philip W Kuchel, Guilhem Pagès, Kaz Nagashima, Sendhil Velan, Vimalan Vijayaragavan, Vijayasarathi Nagarajan, and Kai Hsiang Chuang. "Stejskal–tanner equation derived in full". In: *Concepts in Magnetic Resonance Part A* 40.5 (2012), pp. 205–214.

[117] Mark A Horsfield and Derek K Jones. "Applications of diffusion-weighted and diffusion tensor MRI to white matter diseases–a review". In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 15.7-8 (2002), pp. 570–577.

[118] Marek Kubicki, Robert McCarley, Carl-Fredrik Westin, Hae-Jeong Park, Stephan Maier, Ron Kikinis, Ferenc A Jolesz, and Martha E Shenton. “A review of diffusion tensor imaging studies in schizophrenia”. In: *Journal of psychiatric research* 41.1-2 (2007), pp. 15–30.

[119] Ben Jeurissen, Maxime Descoteaux, Susumu Mori, and Alexander Leemans. “Diffusion MRI fiber tractography of the brain”. In: *NMR in Biomedicine* 32.4 (2019), e3785.

[120] Timothy EJ Behrens et al. “Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?” In: *NeuroImage* 34.1 (2007), pp. 144–155.

[121] Chang-hyun Park, Soo Yong Kim, Yun-Hee Kim, and Kyungsik Kim. “Comparison of the small-world topology between anatomical and functional connectivity in the human brain”. In: *Physica A: statistical mechanics and its applications* 387.23 (2008), pp. 5958–5962.

[122] Yu Sun, Qihua Yin, Rong Fang, Xiaoxiao Yan, Ying Wang, Anastasios Bezerianos, Huidong Tang, Fei Miao, and Junfeng Sun. “Disrupted functional brain connectivity and its association to structural connectivity in amnestic mild cognitive impairment and Alzheimer’s disease”. In: *PloS one* 9.5 (2014).

[123] Fei Wang, Jessica H Kalmar, Yong He, Marcel Jackowski, Lara G Chepenik, Erin E Edmiston, Karen Tie, Gaolang Gong, Maulik P Shah, Monique Jones, et al. “Functional and structural connectivity between the perigenual anterior cingulate and amygdala in bipolar disorder”. In: *Biological psychiatry* 66.5 (2009), pp. 516–521.

[124] Qifeng Wang, Tung-Ping Su, Yuan Zhou, Kun-Hsien Chou, I-Yun Chen, Tianzi Jiang, and Ching-Po Lin. “Anatomical insights into disrupted small-world networks in schizophrenia”. In: *Neuroimage* 59.2 (2012), pp. 1085–1093.

[125] Archana Venkataraman, Yogesh Rathi, Marek Kubicki, Carl-Fredrik Westin, and Polina Golland. “Joint modeling of anatomical and functional connectivity for population studies”. In: *IEEE transactions on medical imaging* 31.2 (2011), pp. 164–182.

[126] Ixavier A Higgins, Suprateek Kundu, and Ying Guo. “Integrative Bayesian analysis of brain functional networks incorporating anatomical knowledge”. In: *Neuroimage* 181 (2018), pp. 263–278.

[127] Chong-Yaw Wee, Pew-Thian Yap, Daoqiang Zhang, Kevin Denny, Jeffrey N Browndyke, Guy G Potter, Kathleen A Welsh-Bohmer, Lihong Wang, and Dinggang Shen. "Identification of MCI individuals using structural and functional connectivity networks". In: *Neuroimage* 59.3 (2012), pp. 2045–2056.

[128] Jing Sui et al. "Combination of resting state fMRI, DTI, and sMRI data to discriminate schizophrenia by N-way MCCA+ jICA". In: *Frontiers in human neuroscience* 7 (2013), p. 235.

[129] Maryam Akhavan Aghdam, Arash Sharifi, and Mir Mohsen Pedram. "Combination of rs-fMRI and sMRI data to discriminate autism spectrum disorders in young children using deep belief network". In: *Journal of digital imaging* 31.6 (2018), pp. 895–903.

[130] Jeffrey D Rudie, JA Brown, D Beck-Pancer, LM Hernandez, EL Dennis, PM Thompson, SY Bookheimer, and MJNC Dapretto. "Altered functional and structural brain network organization in autism". In: *NeuroImage: clinical* 2 (2013), pp. 79–94.

[131] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[132] John Muschelli, Mary Beth Nebel, Brian S Caffo, Anita D Barber, James J Pekar, and Stewart H Mostofsky. "Reduction of motion-related artifacts in resting state fMRI using aCompCor". In: *Neuroimage* 96 (2014), pp. 22–35.

[133] Rastko Ciric, Adon FG Rosen, Guray Erus, Matthew Cieslak, Azeez Adebimpe, Philip A Cook, Danielle S Bassett, Christos Davatzikos, Daniel H Wolf, and Theodore D Satterthwaite. "Mitigating head motion artifact in functional connectivity MRI". In: *Nature protocols* 13.12 (2018), pp. 2801–2826.

[134] Robert W Cox. "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages". In: *Computers and Biomedical research* 29.3 (1996), pp. 162–173.

[135] Mark Jenkinson et al. "Fsl". In: *NeuroImage* 62.2 (2012), pp. 782–790.

[136] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. "The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism". In: *Journal of autism and developmental disorders* 30.3 (2000), pp. 205–223.

[137] Sven Bölte, Fritz Poustka, and John N Constantino. "Assessing autistic traits: cross-cultural validation of the social responsiveness scale (SRS)". In: *Autism Research* 1.6 (2008), pp. 354–363.

[138] Stewart H Mostofsky et al. "Developmental dyspraxia is not limited to imitation in children with autism spectrum disorders". In: *Journal of the International Neuropsychological Society: JINS* 12.3 (2006), p. 314.

[139] MA Dziuk et al. "Dyspraxia in autism: association with motor, social, and communicative deficits". In: *Dev. Medicine & Child Neurology* 49.10 (2007), pp. 734–739.

[140] Lauren R Dowell, E Mark Mahone, and Stewart H Mostofsky. "Associations of postural knowledge and basic motor skill with dyspraxia in autism: implication for abnormalities in distributed connectivity and motor learning." In: *Neuropsychology* 23.5 (2009), p. 563.

[141] Simon B Eickhoff, BT Thomas Yeo, and Sarah Genon. "Imaging-based parcellations of the human brain". In: *Nature Reviews Neuroscience* 19.11 (2018), pp. 672–686.

[142] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain". In: *Neuroimage* 15.1 (2002), pp. 273–289.

[143] Nematollah K Batmanghelich, Ben Taskar, and Christos Davatzikos. "Generative-discriminative basis learning for medical imaging". In: *IEEE transactions on medical imaging* 31.1 (2012), pp. 51–69.

[144] Harini Eavani, Theodore D Satterthwaite, Roman Filipovych, Raquel E Gur, Ruben C Gur, and Christos Davatzikos. "Identifying sparse connectivity patterns in the brain using resting-state fMRI". In: *Neuroimage* 105 (2015), pp. 286–299.

[145] Harini Eavani, Theodore D Satterthwaite, Raquel E Gur, Ruben C Gur, and Christos Davatzikos. “Discriminative sparse connectivity patterns for classification of fMRI data”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 193–200.

[146] Niharika Shimona D’Souza, Mary Beth Nebel, Nicholas Wymbs, Stewart Mostofsky, and Archana Venkataraman. “A generative-discriminative basis learning framework to predict clinical severity from resting state functional MRI data”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 163–171.

[147] Neal Parikh, Stephen Boyd, et al. “Proximal algorithms”. In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239.

[148] Dimitri P Bertsekas and Athena Scientific. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[149] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. “A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data”. In: *Neuroimage* 45.1 (2009), S163–S172.

[150] Charles J Lynch, Lucina Q Uddin, Kaustubh Supekar, Amirah Khouzam, Jennifer Phillips, and Vinod Menon. “Default mode network in childhood autism: posteromedial cortex heterogeneity and relationship with social deficits”. In: *Biological psychiatry* 74.3 (2013), pp. 212–219.

[151] Diane L Williams, Gerald Goldstein, and Nancy J Minshew. “The profile of memory function in children with autism.” In: *Neuropsychology* 20.1 (2006), p. 21.

[152] Devarajan Sridharan, Daniel J Levitin, and Vinod Menon. “A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks”. In: *Proceedings of the National Academy of Sciences* 105.34 (2008), pp. 12569–12574.

[153] Mingrui Xia, Jinhui Wang, and Yong He. “BrainNet Viewer: a network visualization tool for human brain connectomics”. In: *PloS one* 8.7 (2013), e68910.

[154] Adriana Di Martino, David O’connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. “Enhancing studies of the connectome in autism using the autism brain imaging data exchange II”. In: *Scientific data* 4.1 (2017), pp. 1–15.

[155] Jayaraman J Thiagarajan et al. “Multiple kernel sparse representations for supervised and unsupervised learning”. In: *IEEE transactions on Img. Proc.* 23.7 (2014), pp. 2905–2915.

[156] Mayssa Soussia and Islem Rekik. “High-order connectomic manifold learning for autistic brain state identification”. In: *International Workshop on Connectomics in Neuroimaging*. Springer. 2017, pp. 51–59.

[157] Niharika Shimona D’Souza, Mary Beth Nebel, Nicholas Wymbs, Stewart Mostofsky, and Archana Venkataraman. “A Coupled Manifold Optimization Framework to Jointly Model the Functional Connectomics and Behavioral Data Spaces”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 605–616.

[158] Bernard N Flury. “Asymptotic theory for common principal component analysis”. In: *The annals of Statistics* (1986), pp. 418–430.

[159] Stephen Wright et al. “Numerical optimization”. In: *Springer Science* 35.67-68 (1999), p. 7.

[160] Niharika Shimona D’Souza, Mary Beth Nebel, Nicholas Wymbs, Stewart Mostofsky, and Archana Venkataraman. “Integrating neural networks and dictionary learning for multidimensional clinical characterizations from functional connectomics data”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 709–717.

[161] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[162] Christopher M Bishop. *Mixture density networks*. Tech. rep. Citeseer, 1994.

[163] Liron Rabany, Sophy Brocke, Vince D Calhoun, Brian Pittman, Silvia Corbera, Bruce E Wexler, Morris D Bell, Kevin Pelphrey, Godfrey D Pearlson, and Michal Assaf. “Dynamic functional connectivity in schizophrenia and autism spectrum disorder: Convergence, divergence and classification”. In: *NeuroImage: Clinical* 24 (2019), p. 101966.

[164] Niharika Shimona D’Souza, Mary Beth Nebel, Deana Crocetti, Nicholas Wymbs, Joshua Robinson, Stewart Mostofsky, and Archana Venkataraman. “A Deep-Generative Hybrid Model to Integrate Multimodal and Dynamic Connectivity for Predicting Spectrum-Level Deficits in Autism”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 437–447.

[165] Niharika Shimona D'Souza, Mary Beth Nebel, Deana Crocetti, J Robinson, N Wymbs, Stewart H Mostofsky, and Archana Venkataraman. "Deep sr-DDL: Deep structurally regularized dynamic dictionary learning to integrate multimodal and dynamic functional connectomics data for multidimensional clinical characterizations". In: *NeuroImage* 241 (2021), p. 118388.

[166] Anirban Banerjee and Jürgen Jost. "On the spectrum of the normalized graph Laplacian". In: *Linear algebra and its applications* 428.11-12 (2008), pp. 3015–3022.

[167] Jiahao Pang and Gene Cheung. "Graph Laplacian regularization for image denoising: Analysis in the continuous domain". In: *IEEE Transactions on Image Processing* 26.4 (2017), pp. 1770–1785.

[168] Chun-Mei Feng, Ying-Lian Gao, Jin-Xing Liu, Chun-Hou Zheng, and Jiguo Yu. "PCA based on graph Laplacian regularization and P-norm for gene selection and clustering". In: *IEEE transactions on nanobioscience* 16.4 (2017), pp. 257–265.

[169] Rémi Cuingnet, Joan Alexis Glaunès, Marie Chupin, Habib Benali, and Olivier Colliot. "Spatial and anatomical regularization of SVM: a general framework for neuroimaging data". In: *IEEE transactions on pattern analysis and machine intelligence* 35.3 (2012), pp. 682–696.

[170] Jonathan H Manton, Robert Mahony, and Yingbo Hua. "The geometry of weighted low-rank approximations". In: *IEEE Transactions on Signal Processing* 51.2 (2003), pp. 500–514.

[171] Robert B Schnabel and Ph L Toint. "Forcing sparsity by projecting with respect to a non-diagonally weighted Frobenius norm". In: *Mathematical Programming* 25.1 (1983), pp. 125–129.

[172] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[173] Richard Everson. "Orthogonal, but not orthonormal, procrustes problems". In: *Advances in computational Mathematics* 3.4 (1998).

[174] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[175] Diederik P Kingma and Jimmy Lei Ba. "Adam: A Method for Stochastic Optimization". In: (2015).

[176] Van Essen et al. “The WU-Minn human connectome project: an overview”. In: *Neuroimage* 80 (2013), pp. 62–79.

[177] Joelle Zimmermann, John D Griffiths, and Anthony R McIntosh. “Unique mapping of structural and functional connectivity on cognition”. In: *Journal of Neuroscience* 38.45 (2018), pp. 9658–9667.

[178] Stephen M Smith et al. “Resting-state fMRI in the human connectome project”. In: *Neuroimage* 80 (2013), pp. 144–168.

[179] G Kiar et al. “ndmg: Neurodata’s mri graphs pipeline”. In: *Zenodo* (2016).

[180] John Duncan. “Frontal lobe function and general intelligence: why it matters.” In: *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior* (2005).

[181] Warren B Bilker et al. “Development of abbreviated nine-item forms of the Raven’s standard progressive matrices test”. In: *Assessment* 19.3 (2012), pp. 354–369.

[182] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011, pp. 315–323.

[183] Carlo Sestieri et al. “Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses”. In: *J. Neuroscience* 31.12 (2011).

[184] Jessica R Andrews-Hanna. “The brain’s default network and its adaptive role in internal mentation”. In: *The Neuroscientist* 18.3 (2012), pp. 251–270.

[185] Lucina Q Uddin, BT Thomas Yeo, and R Nathan Spreng. “Towards a universal taxonomy of macro-scale functional human brain networks”. In: *Brain topography* (2019), pp. 1–17.

[186] Vinod Menon. “Large-scale brain networks and psychopathology: a unifying triple network model”. In: *Trends in cognitive sciences* 15.10 (2011), pp. 483–506.

[187] David R Euston, Aaron J Gruber, and Bruce L McNaughton. “The role of medial prefrontal cortex in memory and decision making”. In: *Neuron* 76.6 (2012), pp. 1057–1070.

[188] Luke J Hearne, Jason B Mattingley, and Luca Cocchi. “Functional brain networks related to individual differences in human intelligence at rest”. In: *Scientific reports* 6 (2016), p. 32328.

[189] Oliver Y Chén et al. “Resting-state brain information flow predicts cognitive flexibility in humans”. In: *Scientific reports* 9.1 (2019), pp. 1–16.

[190] Carissa Cascio, Francis McGlone, Stephen Folger, Vinay Tannan, Grace Baranek, Kevin A Pelphrey, and Gregory Essick. “Tactile perception in adults with autism: a multidimensional psychophysical study”. In: *Journal of autism and developmental disorders* 38.1 (2008), pp. 127–137.

[191] Lucinda BC Pouw, Carolien Rieffe, Lex Stockmann, and Kenneth D Gadow. “The link between emotion regulation, social functioning, and depression in boys with ASD”. In: *Research in Autism Spectrum Disorders* 7.4 (2013), pp. 549–556.

[192] Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. “Tracking whole-brain connectivity dynamics in the resting state”. In: *Cerebral cortex* 24.3 (2014), pp. 663–676.

[193] Soham Gadgil, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Ehsan Adeli, and Kilian M Pohl. “Spatio-Temporal Graph Convolution for Resting-State fMRI Analysis”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 528–538.

[194] Antonis D Savva, Georgios D Mitsis, and George K Matsopoulos. “Assessment of dynamic functional connectivity in resting-state fMRI using the sliding window technique”. In: *Brain and behavior* 9.4 (2019), e01255.

[195] Rushil Anirudh and Jayaraman J Thiagarajan. “Bootstrapping graph convolutional neural networks for autism spectrum disorder classification”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3197–3201.

[196] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. “Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer’s disease”. In: *Medical image analysis* 48 (2018), pp. 117–130.

[197] Sidong Liu et al. "Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders". In: *Brain informatics* 2.3 (2015), pp. 167–180.

[198] Wen Zhang, Liang Zhan, Paul Thompson, and Yalin Wang. "Deep representation learning for multimodal brain networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 613–624.

[199] Gloria Castellazzi et al. "A machine learning approach for the differential diagnosis of alzheimer and vascular dementia fed by MRI selected features". In: *Frontiers in neuroinformatics* 14 (2020), p. 25.

[200] Lan Lin et al. "Predicting healthy older adult's brain age based on structural connectivity networks using artificial neural networks". In: *Computer methods and programs in biomedicine* 125 (2016), pp. 8–17.

[201] Eleanor Wong, Jeffrey S Anderson, Brandon A Zielinski, and P Thomas Fletcher. "Riemannian regression and classification models of brain networks applied to autism". In: *International Workshop on Connectomics in Neuroimaging*. Springer. 2018, pp. 78–87.

[202] Jiahao Liu et al. "Community-preserving graph convolutions for structural and functional joint embedding of brain networks". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 1163–1168.

[203] Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. "Topology-aware generative adversarial network for joint prediction of multiple brain graphs from a single brain graph". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 551–561.

[204] Tero Karras et al. "Training generative adversarial networks with limited data". In: *arXiv preprint arXiv:2006.06676* (2020).

[205] Zhiwu Huang and Luc Van Gool. "A riemannian network for spd matrix learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 2017.

[206] Zhen Dong et al. "Deep manifold learning of symmetric positive definite matrices with application to face recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 2017.

[207] Chang Wang et al. “Manifold alignment using procrustes analysis”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1120–1127.

[208] Niharika Shimona D’Souza, Mary Beth Nebel, Deana Crocetti, Nicholas Wymbs, Joshua Robinson, Stewart Mostofsky, and Archana Venkataraman. “A Matrix Autoencoder Framework to Align the Functional and Structural Connectivity Manifolds as Guided by Behavioral Phenotypes”. In: *arXiv preprint arXiv:2105.14409* (2021).

[209] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[210] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. “Efficient backprop”. In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[211] Bernhard K Flury. “Two generalizations of the common principal component model”. In: *Biometrika* 74.1 (1987), pp. 59–69.

[212] Toni Duras. “Aspects of common principal components”. PhD thesis. Jönköping University, Jönköping International Business School, 2017.

[213] Hongming Li, Theodore D Satterthwaite, and Yong Fan. “Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks”. In: *2018 ieee 15th international symposium on biomedical imaging (isbi 2018)*. IEEE. 2018, pp. 101–104.

[214] Gazi F Azad, Erica Reisinger, Ming Xie, and David S Mandell. “Parent and teacher concordance on the Social Responsiveness Scale for children with autism”. In: *School mental health* 8.3 (2016), pp. 368–376.

[215] Katherine Gotham, Susan Risi, Andrew Pickles, and Catherine Lord. “The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity”. In: *Journal of autism and developmental disorders* 37.4 (2007), pp. 613–627.

[216] Cynthia A Molloy, Donna S Murray, Rachel Akers, Terry Mitchell, and Patricia Manning-Courtney. “Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting”. In: *Autism* 15.2 (2011), pp. 143–162.

[217] Eleanor Wong, Sourabh Palande, Bei Wang, Brandon Zielinski, Jeffrey Anderson, and P Thomas Fletcher. “Kernel partial least squares regression for relating functional brain network topology to clinical measures of behavior”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 1303–1306.

[218] Juntang Zhuang, Nicha C Dvornek, Xiaoxiao Li, Pamela Ventola, and James S Duncan. “Prediction of Severity and Treatment Outcome for ASD from fMRI”. In: *International Workshop on PRedictive Intelligence In MEdicine*. Springer. 2018, pp. 9–17.

[219] Alexandre Abraham, Michael P Milham, Adriana Di Martino, R Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. “Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example”. In: *NeuroImage* 147 (2017), pp. 736–745.

[220] Qingyu Zhao, Dongjin Kwon, Eva M Müller-Oehring, Anne-Pascale Le Berre, Adolf Pfefferbaum, Edith V Sullivan, and Kilian M Pohl. “Longitudinally consistent estimates of intrinsic functional networks”. In: *Human brain mapping* 40.8 (2019), pp. 2511–2528.

[221] Timothée Lesort, Massimo Caccia, and Irina Rish. “Understanding Continual Learning Settings with Data Distribution Drift Analysis”. In: *arXiv preprint arXiv:2104.01678* (2021).

[222] Baochen Sun and Kate Saenko. “Subspace distribution alignment for unsupervised domain adaptation.” In: *BMVC*. Vol. 4. 2015, pp. 24–1.

[223] Xin Chen, Lili Huang, Qing Ye, Dan Yang, Ruomeng Qin, Caimei Luo, Mengchun Li, Bing Zhang, and Yun Xu. “Disrupted functional and structural connectivity within default mode network contribute to WMH-related cognitive impairment”. In: *NeuroImage: Clinical* 24 (2019), p. 102088.

[224] Alison E Lane, Simon J Dennis, and Maureen E Geraghty. “Brief report: further evidence of sensory subtypes in autism”. In: *Journal of autism and developmental disorders* 41.6 (2011), pp. 826–831.

[225] Gerald J August and Barry D Garfinkel. “Behavioral and cognitive subtypes of ADHD”. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 28.5 (1989), pp. 739–748.

[226] Tara Stevens, Lei Peng, and Lucy Barnard-Brak. “The comorbidity of ADHD in children diagnosed with autism spectrum disorder”. In: *Research in Autism Spectrum Disorders* 31 (2016), pp. 11–18.

[227] Markus D Schirmer, Archana Venkataraman, Islem Rekik, Minjeong Kim, Stewart H Mostofsky, Mary Beth Nebel, Keri Rosch, Karen Seymour, Deana Crocetti, Hassna Irzan, et al. "Neuropsychiatric disease classification using functional connectomics-results of the connectomics in neuroimaging transfer learning challenge". In: *Medical image analysis* 70 (2021), p. 101972.

[228] Yujiro Yoshihara, Giuseppe Lisi, Noriaki Yahata, Junya Fujino, Yukiko Matsumoto, Jun Miyata, Gen-ichi Sugihara, Shin-ichi Urayama, Manabu Kubota, Masahiro Yamashita, et al. "Overlapping but asymmetrical relationships between schizophrenia and autism revealed by brain connectivity". In: *Schizophrenia bulletin* 46.5 (2020), pp. 1210–1218.

[229] Oliver Y Chén, Hengyi Cao, Huy Phan, Guy Nagels, Jenna M Reinen, Jiangtao Gou, Tianchen Qian, Junrui Di, John Prince, Tyrone D Cannon, et al. "Identifying neural signatures mediating behavioral symptoms and psychosis onset: High-dimensional whole brain functional mediation analysis". In: *NeuroImage* 226 (2021), p. 117508.

[230] Martin A Lindquist. "Functional causal mediation analysis with an application to brain connectivity". In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1297–1309.

[231] Dushyant Sahoo, Theodore D Satterthwaite, and Christos Davatzikos. "Hierarchical extraction of functional connectivity components in human brain using resting-state fmri". In: *IEEE Transactions on Medical Imaging* 40.3 (2020), pp. 940–950.

[232] Yueying Zhou, Limei Zhang, Shenghua Teng, Lishan Qiao, and Dinggang Shen. "Improving sparsity and modularity of high-order functional connectivity networks for MCI and ASD identification". In: *Frontiers in neuroscience* 12 (2018), p. 959.

[233] Şeymanur Aktı, Doğay Kamar, Özgür Anıl Özlü, Ihsan Soydemir, Muhammet Akcan, Abdullah Kul, and Islem Rekik. "A Comparative Study of Machine Learning Methods for Predicting the Evolution of Brain Connectivity from a Baseline Timepoint". In: *arXiv preprint arXiv:2109.07739* (2021).

[234] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. "Identification of autism spectrum disorder using deep learning and the ABIDE dataset". In: *NeuroImage: Clinical* 17 (2018), pp. 16–23.

[235] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. "Graph neural networks: A review of methods and applications". In: *arXiv preprint arXiv:1812.08434* (2018).

[236] Tzu-An Song, Samadrita Roy Chowdhury, Fan Yang, Heidi Jacobs, Georges El Fakhri, Quanzheng Li, Keith Johnson, and Joyita Dutta. "Graph convolutional neural networks for Alzheimer's disease classification". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 414–417.

[237] Thomas N Kipf and Max Welling. "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907* (2016).

[238] Niharika Shimona Dsouza, Mary Beth Nebel, Deana Crocetti, Joshua Robinson, Stewart Mostofsky, and Archana Venkataraman. "M-GCN: A Multimodal Graph Convolutional Network to Integrate Functional and Structural Connectomics Data to Predict Multidimensional Phenotypic Characterizations". In: *Medical Imaging with Deep Learning*. 2021.

[239] Joan Bruna et al. "Spectral networks and locally connected networks on graphs". In: *arXiv preprint arXiv:1312.6203* (2013).

[240] Jun Qi, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. "On mean absolute error for deep neural network based vector-to-vector regression". In: *IEEE Signal Processing Letters* 27 (2020), pp. 1485–1489.

[241] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.