Deep Learning and Structural MRI: In Pursuit of Improved Self-Regulatory Behavior and Mental Health Prediction

Written Dissertation

Gareth Harman[1,2]

Dissertation Advisory Committee:

Angelica Morales[1]
Jayashree Kalpathy-Cramer[4]
Michael Mooney[2]
Bonnie Nagel[1,3]

[1]Oregon Health & Science University, Department of Psychiatry

[2]Oregon Health & Science University, Department of Medical Informatics and Clinical Epidemiology

[3]Oregon Health & Science University, Behavioral Neuroscience

[4]University of Colorado Anschutz, Department of Ophthalmology

Acknowledgements

Few people have offered the amount of encouragement, compassion, and unwavering belief in my abilities that Dr. Bonnie Nagel provided during my graduate tenure. It would not be hyperbolic to say that I would not be the person and researcher I am today without her guidance and support as my chair, role model, and perpetual advocate. I must also thank my dissertation advisor, Dr. Michal Mooney, for his advice, support, and expertise in a field that is ever-changing, as well as my other dissertation committee members, Dr. Angelica Morales and Dr. Jayashree Kalpathy-Cramer. I cannot express how fortunate I feel to have found a group of such brilliant researchers and wonderful human beings to guide and shape my academic development.

I must also thank my friends, family, and support structures that kept me afloat during the long days of writing and failed analyses. To my sisters, Sophie and Kira, my father, Bruce, and my mother, Shari, I am forever thankful for your encouragement, good-natured ribbing, and food during the challenging periods. To my parents, my mother specifically, for her undying love and endless frosted desserts, and for instilling the value of kindness and shaping me into the person I am today. To my great friend and fellow self-compassion advocate, Lucas, for his support and assistance in creating and maintaining a life full of present moments, intentional living, and self-compassion above all else. A heartfelt thank you to my fellow Olympic weightlifting community and friends at the gym.

Finally, to my colleagues, coworkers, and collaborators: I must express the absolute pleasure it was to be part of a cohort comprised of such wonderful students and people, and for the late nights, takeout, and ceaseless brainstorming.

Contents

Table of figures

Acronyms

DEEP LEARNING

FFNN	Feed Forward Neural Network
FCNN	Fully Connected Neural Network
ANN	Artificial Neural Network
AE	Autoencoder
VAE	Variational Autoencoder
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network

STATISTICS

LR	Linear Regression
PCA	Principal Components Analysis
PLS (R)	Partial Least Squares (Regression)
MSE	Mean Squared Error
MAE	Mean Absolute Error
r	Pearson Correlation Coefficient

IMAGING

MRI	Magnetic Resonance Imaging
fMRI	Functional Magnetic Resonance Imaging

sMRI	Structural MRI
rs-fMRI	Resting State fMRI
BOLD	Blood Oxygen Level Dependent (signal)
rsfc	Resting State Functional Connectivity
PET	Positron Emission Tomography
FD	Framewise Displacement
TR	Time to Repetition
(t)SNR	(temporal) Signal-to-noise Ratio

CLINICAL PSYCHIATRY

ABCD	Adolescent Brain and Cognitive Development Study
EF	Executive Function
WM	Working Memory
IC	Inhibitory Control
SS	Set Shifting
EXT	Externalizing Disorders
INT	Internalizing Disorders
HC	Healthy Control
MHD	Mental Health Disorders
MDD	Major Depressive Disorder
GAD	Generalized Anxiety Disorder
SCZ	Schizophrenia
ASD	Autism Spectrum Disorder

AD(H)D	Attention Deficit (Hyperactivity) Disorder
PD	Parkinson's Disease
AD	Alzheimer's Disease
NIHTbx	NIH Toolbox
MCI	Mild Cognitive Impairment

Chapter 1: Emergent psychopathology and the developing brain

The Burden of Mental Health in Childhood

Nearly one in five children (17%) aged 2-8 were diagnosed with either a mental, emotional, developmental, or behavioral disorder in 2016 in the US. Furthermore, within those aged 10-19, 15% possess a mental health disorder (MHD), making MHD nearly 13% of the global burden of disease in this age bracket ("2022 National Healthcare Quality and Disparities Report," 2022). Most critically, suicidal thoughts and behaviors rose 40% from 2009-2019 among high school students. Recently, the AAP (American Academy of Pediatrics), AACAP (American Academy of Child and Adolescent Psychiatry) and the Children's Hospital Association have declared "unmet youth mental health needs" a national emergency (Sorter et al., 2023), leading to the Biden-Harris administration to pledge a 300-million dollar fund to support mental health services in schools (Biden-Harris | SAMHSA, 2022.).

Early identification of MHD is crucial, as many mental health conditions first appear during these formative years, offering a unique window for effective intervention and improving healthy transition into adulthood (Scheiner et al., 2022). The impact of early MHD extends beyond health, influencing education, future employment, and social relationships, with effects not only on the individual but their families and communities (Ruggero et al., 2019). Thus, prioritizing our understanding, characterization, and treatment of MHD in young populations extends beyond individual well-being, becoming a top public health priority (Malla et al., 2018). Additional research has the potential to inform public policy, influence resource allocation, and improve both school-based and community health programs. Furthermore, continued research and increased awareness is critical for reducing stigma associated with MHD, encouraging early

treatment seeking behavior, and ensuring the modification of age-appropriate therapeutic approaches that are more effective for younger populations (Sheikhan et al., 2023; Villatoro et al., 2022). Moreover, the comorbidity of, and heterogeneity within MHD in children and adolescents presents additional challenges, necessitating a comprehensive approach to treatment and care.

Dimensional models of psychopathology

The discretization of symptoms into binary diagnoses, i.e. presence or absence of major depressive disorder, has led to considerable difficulties when attempting to model these disorders (Caspi & Moffitt, 2018). For example, the presence of "difficulty concentrating" is a known symptom of generalized anxiety disorder (GAD), major depressive disorder (MDD), and attention deficit disorder (ADD)(Riglin et al., 2021). This discretization via symptom thresholding creates substantial heterogeneity within these disorders. In a similar fashion, research has clearly established that there is great utility in understanding variability between individuals that endorse no symptoms of a disorder and those that fall just below this diagnostic threshold referred to as sub-threshold diagnoses (Caspi et al., 2014; Caspi & Moffitt, 2018). Adding to the complexity of characterizing these disorders, classic clinical nosology fails to capture unique variability of individual disorders due to frequent comorbidity (possessing multiple diagnoses), creating additional challenges for understanding the unique characteristics and traits of individuals suffering from these disorders in both research and clinical settings. One study found that, of individuals meeting criteria for one diagnosis, 66% met criteria for a second, 51% of those that have a second diagnosis met criteria for a third and so forth (McGrath et al., 2020). These challenges have led researchers to examine potential factor structures of these

disorders and the subsequent creation of *dimensional models of psychopathology*, such as the Hierarchical Taxonomy of Psychopathology (HiTOP) and P-factor of psychopathology (Caspi et al., 2014; Ruggero et al., 2019).

Early factor models of psychopathology revealed symptom profiles, including internalizing symptoms (disorders such as depression, anxiety, and phobias), externalizing (disorders such as alcohol and substance use, conduct disorder, and attention deficit hyperactivity disorder), and thought-disordered symptoms (schizophrenia, obsessive compulsive disorder, and mania) (Magyar & Pandolfi 2018). However, recent factor models have created a *general* factor of psychopathology, the *p-factor*, that accounts for much of the shared variance of all mental health disorders (Caspi et al., 2014; Caspi & Moffitt, 2018). Akin to the *g-factor* of intelligence, a latent variable capturing general intelligence and accounts for the fact that individuals that score high on one intelligence test often score high on the others, the *p-factor* provides several attractive elements as a measure of psychopathology. First, it is a continuous variable that has been shown to be normally distributed in the general population (Caspi & Moffitt, 2018), it accounts for a substantial portion of the variance between mental health disorders, and it has been linked to genetics and neurobiology (Sprooten et al., 2021), treatment outcomes (Cervin et al., 2021), and family history (Caspi et al., 2014; Caspi & Moffitt, 2018). However, while the p-factor provides an appealing measure of psychopathology and removes issues related to discretization and sub-threshold disease presentation, there remains unexplained variance within these disorders that disease specific clinical symptoms alone cannot capture.

Executive Function and Mental Health

Strategies to expound upon or model aspects of remaining variability within these disorders may be a critical prerequisite to comprehensive modeling of mental health disorders. One example of such a strategy, from Marquand and colleagues (Marquand et al., 2019) found that they were able to reveal variability within a sample of individuals with attention deficit hyperactivity disorder (ADHD) by first modeling abnormal patterns of reward response. Yet another promising opportunity, highlighted by Snyder et al., (2015), is the evaluation of impairments and dysfunction within components of executive function and their role in psychopathology. Aberrant components of executive function have been tied to essentially all forms of psychopathology (Eisenberg et al., 2009; Espy et al., 2011a; Martel & Nigg, 2006; Nigg, 2017a). Furthermore, executive function dysfunction has been implicated as a predictor for other risk factors of psychopathology, including worrying, rumination, and issues with using emotional regulation tactics (Andreotti et al., 2013; Crowe et al., 2007; De Lissnyder et al., 2012; Whitmer & Banich, 2007; Zetsche et al., 2012). Further evidence from the Adolescent Brain and Cognitive Development (ABCD) study found that dysfunctions within components of executive function at baseline were prospective predictors of psychopathology two years later (Romer & Pizzagalli, 2021). Additionally, researchers found evidence for leveraging transdiagnostic brain-based measures of cognition that characterized variability in the development of MHD in early adolescence within the ABCD sample (Xiao et al., 2023). Finally, a recent study by Cordova et al, found that not only were features of executive function such as working memory, response inhibition, and cognitive flexibility, predictive of autism spectrum disorder and ADHD, but these features also revealed distinct subtypes within these disorders (Cordova et al., 2020). All together

these findings highlight the important role executive function plays within psychopathology and the utility of examining distinct factors of executive function within these disorders.

Importance of Neurobiological Markers

The field of biological psychiatry asserts that disease manifestation does not solely consist of clinical symptoms and highlights the importance of establishing biomarkers to better understand and profile mental health disorders (Brückl et al., 2020). Important to clearly describe, the *Biomarkers Definitions Working Group* define a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathologic processes, or biological responses to a therapeutic intervention". Furthermore, the period of childhood and early adolescence is one of significant neurodevelopment, referred to as highly neuroplastic, with the brain prioritizing the development of distal functional networks and an increase in white-matter myelination. Neuroimaging derived biomarkers have been identified across the range of MHD including emerging depression (Kliamovich et al., 2021), autism spectrum disorder (Sen et al., 2018), attention deficit disorder (Albajara Sáenz et al., 2019), and many others.

Neuroimaging, specifically magnetic resonance imaging (MRI), provides a means to evaluate brain structure and function, with current and potential utility to aid in differential diagnosis, prognosis, clinical management, and targeted intervention development (Brückl et al., 2020; Filippi et al., 2012; Malhi & Lagopoulos, 2008; Osuch & Williamson, 2006). As mentioned previously, there is substantial heterogeneity within mental health diagnoses, however, researchers have leveraged neuroimaging biomarkers to uncover putative subgroups within disorders such as autism spectrum disorder (ASD) and ADD (Cordova et al., 2020), mood

and anxiety disorders (Trombello et al., 2018), MDD (Liu et al., 2021), and symptom trajectories within schizophrenia (SCZ) (Jiang et al., 2023). Perhaps one of the most sought-after applications for biomarker-to-clinical translation is to personalized treatment response prediction for mental health disorders, and while there are promising avenues of on-going research, others highlight current limitations of this application (Cohen et al., 2021). However, while there are substantial challenges, most notably relating to computational demands and model training time, remaining in the use of biomarker identification to clinical translation, understanding changes in the brain are a piece of the puzzle of the aggregation that is our mental health.

Feature Extraction: Neuroimaging

Despite these promising biomarker applications, the low signal-to-noise ratio and high-dimensionality of the neuroimaging data provides substantial challenges in elucidating these clinically relevant MHD biomarkers. The 3- or 4-dimensional structure of neuroimaging data provides a considerable input size with over 7 million 3D pixels (voxels) often fraught with significant confounds, including motion, scanner artifacts, and unwanted biological signal such as respiration (Kang et al., 2016). Additionally, these data provide substantial challenges for traditional statistical approaches using hypothesis testing made even more difficult due to the smaller sample sizes, further contributing to the "needle in a haystack" problem (Eklund et al., 2016; Kang et al., 2016; Smith & Nichols, 2018).

Structural Features: T1 sMRI

With T1 structural MRI (sMRI) data, software allows us to evaluate aspects of both morphology and pixel intensity (reflecting tissue type). Typically, an atlas is used to segment the

brain into distinct cortical and subcortical regions based on known neurobiological architecture. For each of these regions we obtain high-level estimates of morphology such as volume, mean thickness, surface area, and indices related to curvature (see Figure 1) using the FreeSurfer software package (Dale et al., 1999). Additionally, we obtain measures of intensity, in which fluid appears dark, regions with more fat appear bright white, gray matter appears as darker grey, and white matter as lighter gray (Taylor et al., 2016).

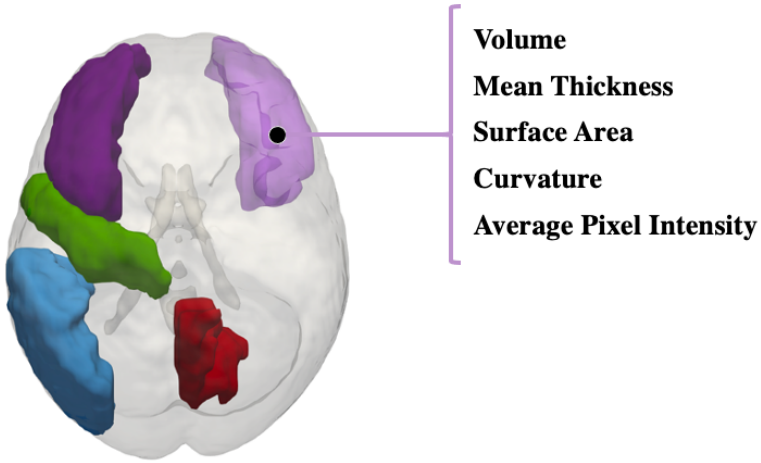


Figure 1: T1 Features

However, when reducing these cortical or subcortical structures, sometimes containing millions of voxels, into only a few summary statistics to capture high level information of morphology and intensity, we must ask the question; *Is there pertinent information from the original image I am failing to capture?* In Figure 2 below, if we compute the relationship between the summary statistics from the superior frontal cortex of two individuals, we obtain a nearly perfect correlation. However, by simple visual inspection we can infer notable differences within the intricate patterns of cortical folding and morphology of these two individuals. There is a trade-

off that we must make when running traditional analyses, often times with large sample sizes. First, many traditional modeling strategies do not allow the use of input data that is not 1-dimensional, and running models with millions of features becomes quickly intractable due to the large amount of memory and computing power that would be required. While there are methods that allow for the examination of more nuanced elements within the cortex, such as the local gyrification index (LGI), these models are immensely difficult if not impossible to run using large sample sizes and have shown limited utility in prediction of cognition (Mathias et al., 2020).

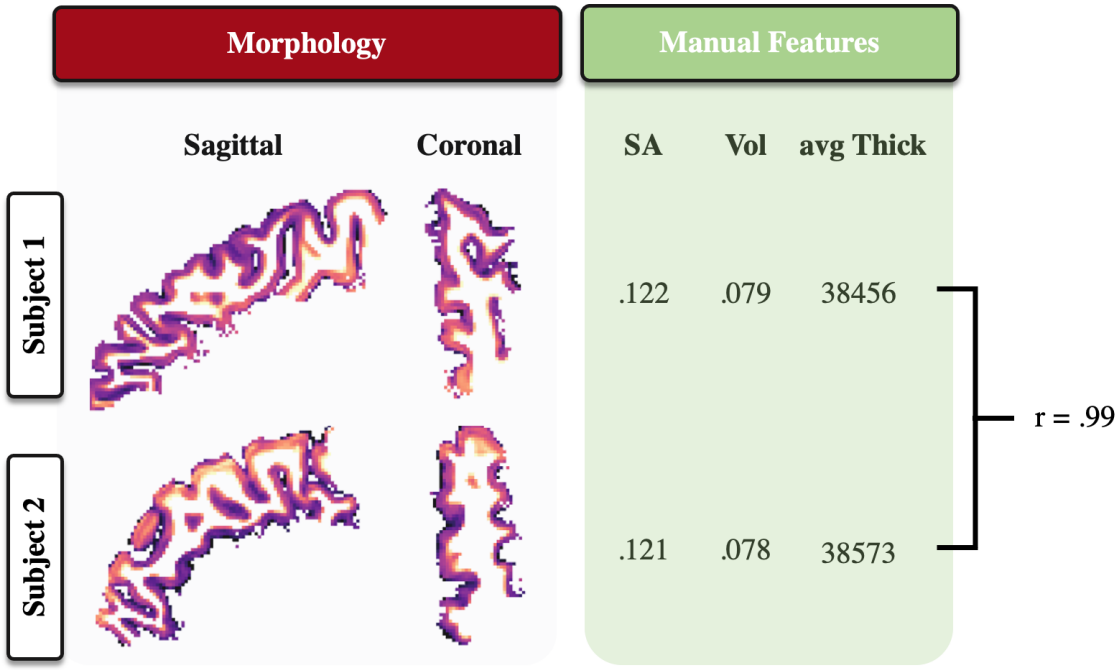


Figure 2: Cortical Similarity

Functional Features: rsfMRI

There are many different methods to capture information related to functional aspects of the brain. In this document we will be focusing on resting-state functional MRI (rsfMRI), a

modality in which we examine the organization and communication of regions of the brain *at-rest* or in the absence of a specific task via coordinated blood oxygen level dependent (BOLD) activity. Typically, the brain is segmented (also referred to as parcellated) into different similar regions. The BOLD signal time-series of all voxels in each region (parcel) are extracted and averaged. Additionally, we can infer information from the *functional-connectome* which is a correlation matrix in which we compute the relationship of these averaged BOLD signals in the brain by computing the Pearson correlation coefficient between parcel 1 and all other parcels (see Figure 3, first row of the correlation matrix).

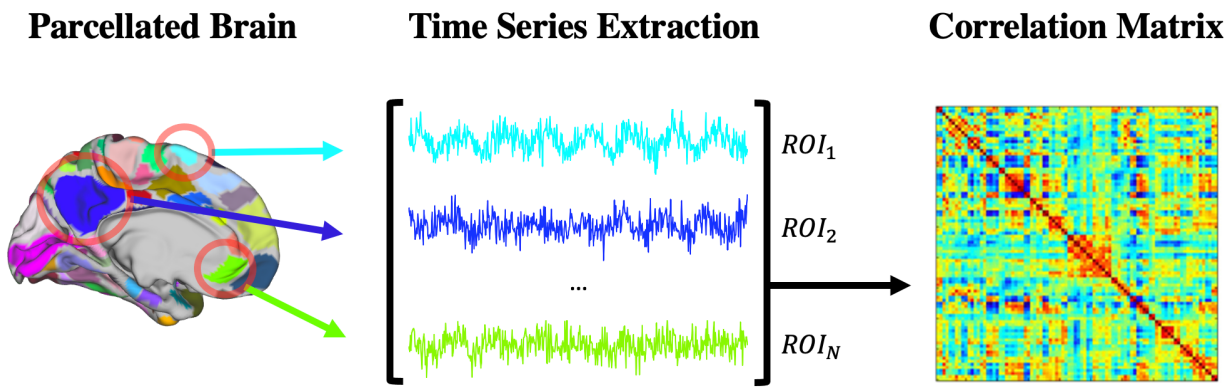


Figure 3: rsfMRI features

There is an inherent lack of granularity arising from examining the relationships of two time courses over such a large temporal window, often 5-10+ minute segments with hundreds of points that may fail to capture subtle localized relationships. Additional techniques attempt to deal with this problem by leveraging a "sliding window" approach, examining smaller segments of the time course. Nevertheless, this method seldom examines relationships between unique windows and suffers from a higher computational demand and can greatly alter results when

using different window lengths (Shakil et al., 2016). Furthermore, despite the averaging of temporal information from hundreds of thousands of voxels into a set of N predefined parcels, when computing pairwise relationships, we yield $\frac{N^2}{2} - N$ features. Thus, a set of 352 parcels yields over 124,000 values, quickly making large sample analyses enormously challenging.

Exploring other avenues: Artificial Intelligence

Artificial Intelligence (AI) has gained a lot of attention in both research settings and the general population, and for good reason. Large language models such as *OpenAI's* generative pretrained chatGPT are capable of impressive prompt-response generation, writing code with better adherence to PEP8 (python style guide) standards than most programmers, and formulating derivative, but passable poetry. Large vision models like Google's DeepDream can instantly generate stunning and unique visualizations of goats on a mountainside. It is important to delineate AI; a broad field in computer science focused on leveraging computations to emulate human intelligence, from machine learning (ML); a subcategory of AI, in which algorithms learn and improve from experience (data) to complete tasks, and finally deep learning (DL); a branch of ML, that utilizes neural networks comprised of many (typically more than four) layers that learn from massive amounts of training data (Ray et al., 2022). Deep learning specifically, excels at unearthing important signal from high-dimensional data without being explicitly told what to extract. Both chatGPT and DeepDream are expressly able to achieve their impressive results by leveraging exceedingly deep neural networks (DNN) with billions of trainable parameters trained using immense amounts of data.

Theoretical Advantages

Complex Data Structures and Feature Extraction

Classical statistical methods have struggled significantly to handle input with complex data structures (i.e. anything that isn't 1D data). Data such as time series, graphs, or imaging data previously required some type of reduction through feature extraction. Specifically, those in the realm of neuroimaging are familiar with methods of feature extraction that we mentioned previously, and the potential shortcomings of these methods of feature extraction in Figure 1. Additionally problematic is the case of image flattening/reshaping. For smaller images such as the handwritten digit prediction dataset MNIST, the 28x28 size images are often reshaped into 784 1D vectors for prediction. Not only intractable for models using larger imaging inputs, such as the 182x212x182 sized T1 sMRI images, but this process results in the loss of contextual information that is embedded in the native dimensionality of the original image. This importance is illustrated clearly in

Figure 4 below. Handwritten digits from the MNIST dataset are used to predict which number they represent. Again, a 2D image cannot be used natively through traditional statistical methods such as linear regression, random forest, etc., thus our first option is to flatten this image. However, when we do this, we can see that we lose information in this 2D structure, visually we as humans can easily identify the number in 2D image but would not be able to identify the flattened version of this image.

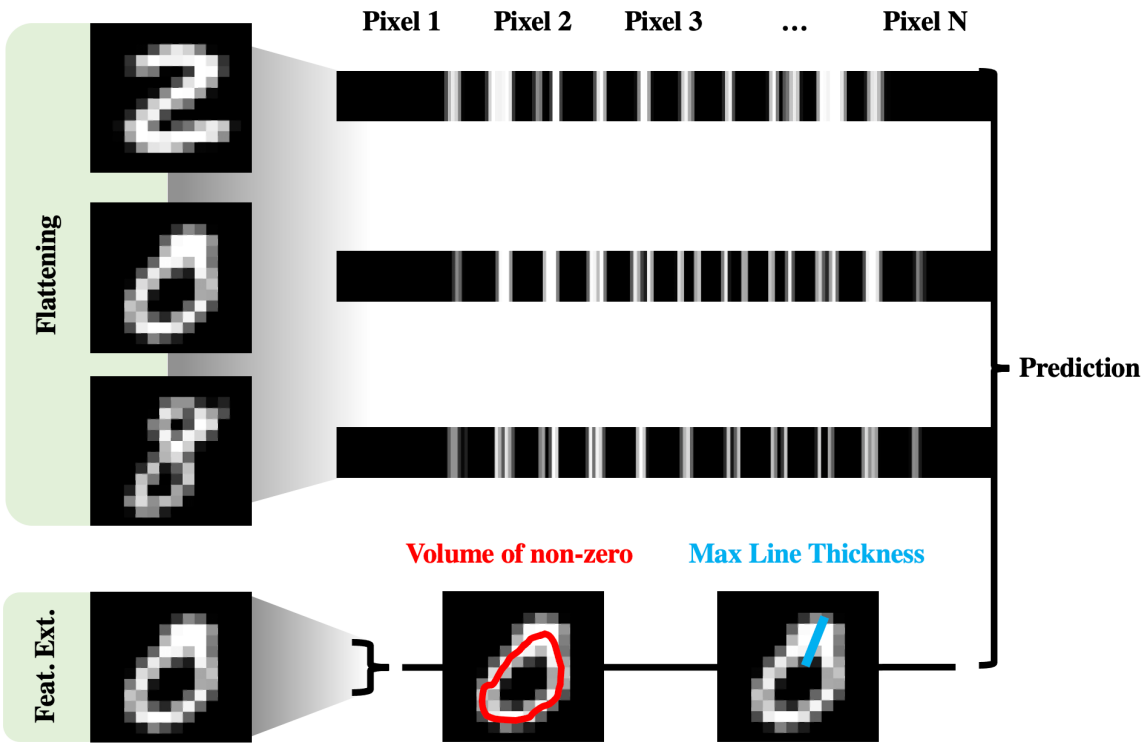


Figure 4: Image Feature Extraction

Perhaps the neural network architecture most essential for the use of imaging data is that of the convolutional neural network (CNN). These models are inherently designed to process data with multiple levels of abstraction. In the image processing space, these models can *extract* features from the native dimensionality of the image. This capability is critical for applications like medical image analysis or autonomous driving, where understanding complex inputs is essential. The foundation of the CNN architecture is the convolution operation. This operation is simply the combination of two signals to make a third signal. In the example below in Figure 5, the first signal is the T1 image and the second is a matrix (kernel) of weights. The two matrices are multiplied, and the resulting outputs are displayed. This example highlights extraction of fundamental "low-level" features from the original image such as vertical (top left) and

horizontal (top right) lines or outlines (lower right) and inverted outlines (lower left). This operation allows these networks to hierarchically extract low-level information (in the early layers) and then aggregate this information to unearth higher-level features (in the later layers of the network).

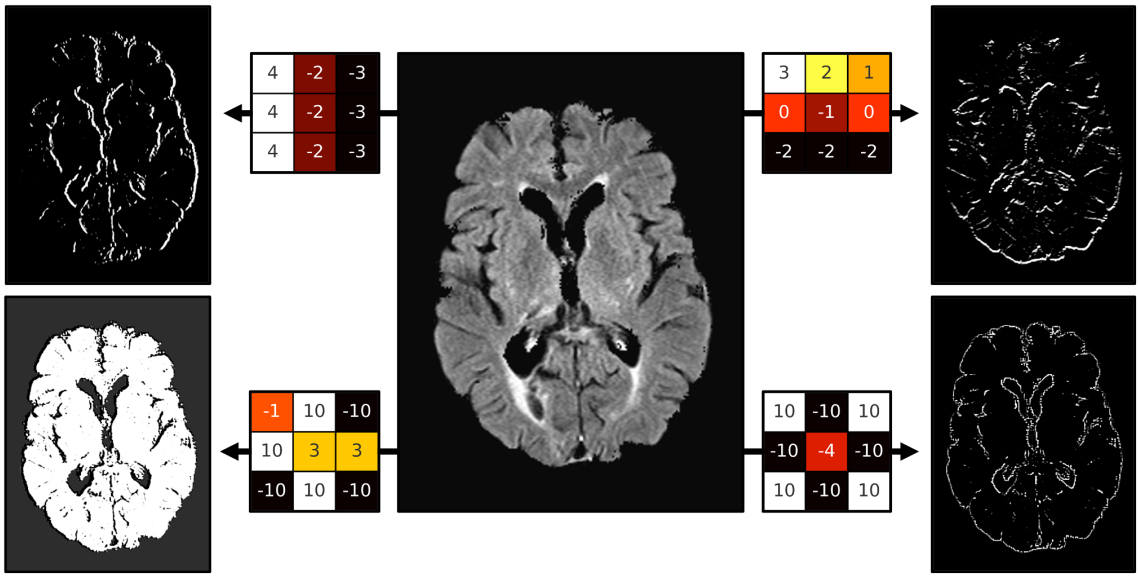


Figure 5: The Convolution

Adaptability and Continuous Learning

Yet another substantial advantage is that DL models, once deployed, can be designed to learn from new data continually (A. Li et al., 2024). This is particularly beneficial in dynamic environments like clinical health systems where the model can adapt to changes over time. These models can also automate or aid in complex decision-making tasks, such as diagnosing diseases from medical scans, which traditionally require expert human analysis and suffer from poor inter-grader reliability (Ulloa et al., 2015).

Computational Advantages

GPU Usage and Parallel Processing and Memory Efficiency

Perhaps the single most important trait of DL models is their ability to leverage graphics processing units (GPUs) to train models using parallel computing. These models, especially those involving large neural networks and/or large imaging data, are well suited for parallelization, meaning they can process many computations concurrently. Continuous advancements in GPU hardware and software utilizing parallel processing, are vital for handling these large, often high-dimensional datasets, dramatically speeding up the learning process.

Furthermore, these networks are highly efficient with their memory usage. Neural networks update the weights using highly optimized learning algorithms that provide avenues for much faster convergence than traditional weight estimation or gradient descent methods. Techniques like network pruning, where insignificant neural network weights are removed, and weight sharing, where weights are reused, make deep learning models more memory-efficient. This is crucial for deploying lightweight models in resource-limited environments (Narang et al., 2017).

One of the more substantial limitations of traditional machine learning involves how models are trained. To learn the weights of a standard regression model, all the data is loaded into memory and coefficients are jointly estimated. Different subsets of data yield different parameter estimates, thus, the addition of new data means entirely retraining the model and a new set of parameter estimates. However, when working with large-scale neuroimaging data (~10,000 subjects) it is intractable if not impossible to load the full dataset into memory. DL

circumvents this by updating the weights of the model in *batches*, with the number of observations in a *batch* being a hyperparameter that is tuned, allowing the model to train using the entire dataset by iterating through N/b times (where N = number of subjects and b = batch size).

Deep Learning in Neuroimaging

Naturally, the computational advancements in DL have gained the interests of neuroimaging researchers to evaluate their utility in numerous processing and prediction domains, including biomarker-disease modeling. The ability of deep neural networks (DNNs) to extract complex patterns and relationships from large-scale neuroimaging datasets and possibly provide insights into brain structure and function that may have been previously inaccessible. Advanced techniques could uncover novel biomarkers for early detection of disease. Furthermore, these algorithms have the potential to automate labor-intensive tasks such as segmentation and feature extraction.

Image Processing and Quality Control

Large-scale DL segmentation models for neuroimaging data have several advantages over traditional non-DL segmentation strategies. First, these models are themselves incredibly efficient. A DL T1 segmentation tool FastSurferCNN (Henschel et al., 2020), similar to that of the MRI software package FreeSurfer, is capable of performing equally accurate semantic segmentation in under one minute, a process that takes FreeSurfer roughly seven hours without parallel CPU utilization (Dale et al., 1999). Additionally, DL models are well suited and efficient at assessing image quality, a process that is enormously time-consuming, requires training, and

can suffer from poor inter-grader score reliability. These models have found success in assessing quality control (QC) metrics for T1 sMRI, diffusion tensor imaging (DTI), as well as denoising and image registration (Garcia et al., 2023; Keshavan et al., 2019; Samani et al., 2019, 2020).

Synthetic Data and Generative Networks

One of the more interesting applications of DNNs is their ability to generate highly plausible synthetic examples of new data given a training set of some existing data. Small sample size is perhaps one of the largest current issues in the field of neuroimaging. The process of obtaining structural and functional neuroimaging data is both time consuming and expensive. The ability of DNNs to generate credible synthetic data has the potential to address issues of data scarcity, augmenting training data to enhance the generalizability of DNNs, and enhanced imputation. These architectures typically use generative adversarial networks (GANs), variational autoencoders (VAEs), and other generative models to produce convincing data along with known ground truth labels or characteristics. By synthesizing diverse and representative neuroimaging data, researchers can mitigate challenges associated with limited sample sizes, data heterogeneity, and privacy concerns, thereby enabling more effective model training and validation (Goceri, 2023; Yin et al., 2019). Moreover, synthetic data augmentation techniques have been shown to improve the generalization and transferability of deep learning models across different imaging modalities and clinical populations (Zhao et al., 2020). Furthermore, the creation of explicitly defined synthetic images enables researchers to explore hypothetical scenarios, emulate disease progression, and examine the effects of interventions in a controlled environment (Dimitriadis et al., 2022).

Prediction

Traditional modeling approaches typically rely on mass univariate hypothesis testing or manual feature extraction. As mentioned previously, the manual extraction of features, such as cortical thickness or volume, creates a more manageable feature space but discards a large amount of the original spatial information which may contain important disease-specific nuances in morphology (Abrol et al., 2021). Deep learning has found recent success in the neuroimaging literature outperforming traditional modeling strategies using manual feature extraction, with applications such as predicting age (Abrol et al., 2021; Lee et al., 2021; Peng et al., 2021) and sex assigned at birth (Leming & Suckling, 2021), as well as the classification of Alzheimer's disease (Ayyar et al., 2021; Lian et al., 2020; Nanni et al., 2020) and Parkinson's disease (Huseyn, n.d.; Kaur et al., 2021; Mozhdehfarahbakhsh et al., 2021). However, the most common applications of deep learning with neuroimaging data typically involves marked developmental or disease-specific neuroanatomical changes. Therefore, the advantages of using deep learning to model emerging mental health disorders remains to be further examined.

Multimodal Fusion Networks

An interesting aspect of deep neural networks is their ability to leverage data from multiple sources in a single network architecture. In traditional analyses data from multiple modalities are simply concatenated; however, this can be tricky in situations in which one modality has many more or different types of features. For example, in a situation where we want to include clinical data with T1 sMRI data in a linear model, several million features for pixels from the T1 sMRI data would be concatenated with, often, small amounts of clinical predictors. The other strategy would be to perform dimensionality reduction on the T1 sMRI data

before concatenating with the clinical features, however, with a multimodal neural network we are able to extract imaging features in the context of the additional clinical features, this is because all features are included within the same network. This is highlighted in toy example in Figure 6 below. In this example, we have a large amount of data in our T1 sMRI image, therefore we construct convolutional layers in the network with the sole purpose of "extracting" information from this input directly and then create constrained sets of neurons in our hidden layer that reflect distinct combinations of our input data (green: T1 sMRI only, purple: T1 sMRI + clinical data, red: clinical data only). Furthermore, we can leverage possible covariates by concatenating them to neurons later in the network. This is only a toy example; we could include covariates (gray) earlier in the network if we believe them to be directly related to low-level features of the input and allow the model to integrate covariates with the raw input data from the beginning. The purpose of this example is to highlight the enormous customizability and flexibility of these networks.

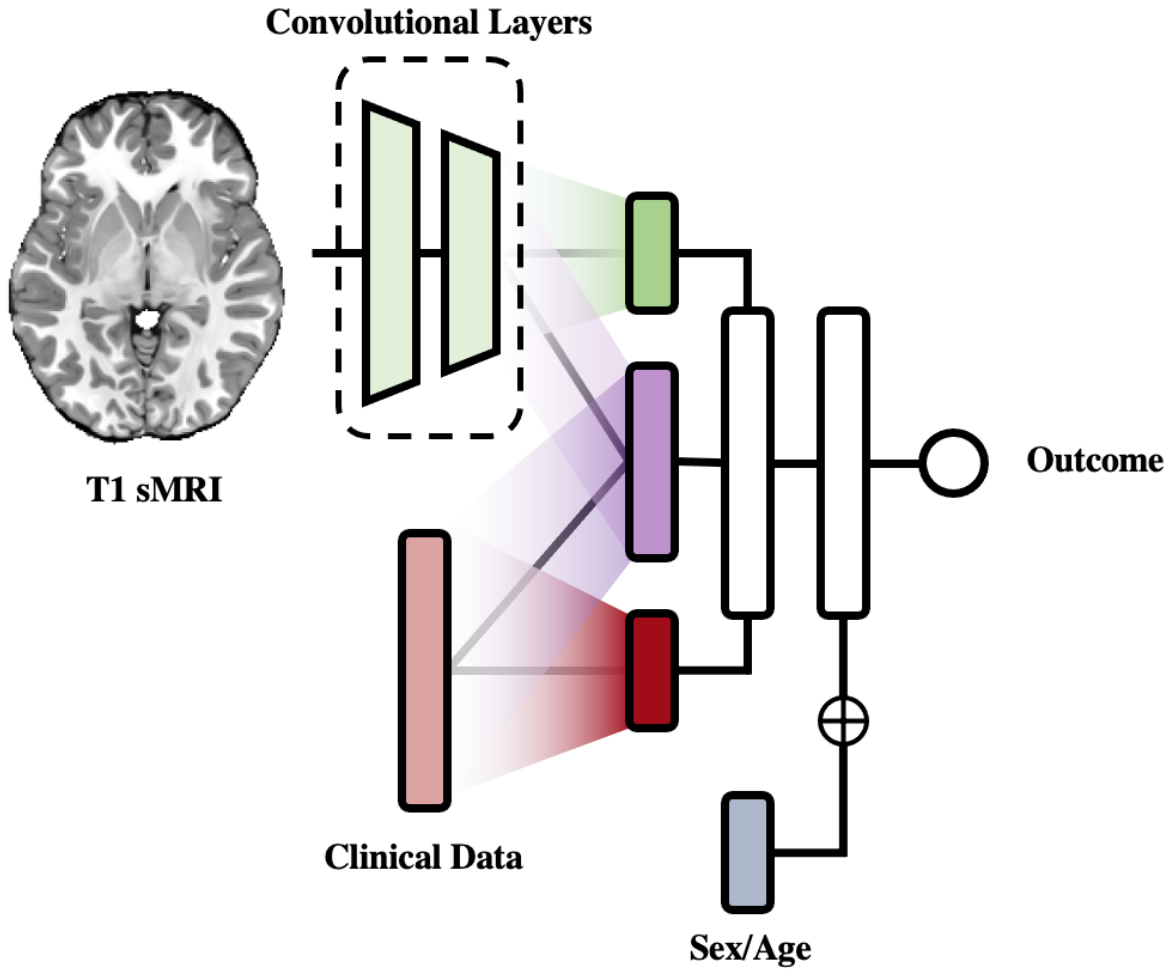


Figure 6: Multimodal Networks

These multimodal neural networks have found novel application in the neuroimaging literature. Leveraging sMRI, genotype, and clinical data in a single multimodal network best delineates Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy controls (HC) (Lu et al., 2018). Researchers were also able to predict early AD diagnoses using fusion sMRI and positron emission tomography (PET) networks (Venugopalan et al., 2021). One of the more challenging facets of neuroimaging is the ability to find and use relationships between structure and function to understand both normative functioning and disease states.

Challenges associated with DL

Infinite Architectures

However, these same properties that make these models appealing, specifically their vast ability of customization, also make them inherently challenging to train and optimize. Networks comprised of billions of trainable parameters can be structured using infinite manners. Determining the number of layers and parameters in each layer, choosing the types of layers (e.g., linear, convolutional, recurrent, or combinations of types), are all time-consuming stages of model evaluation. Additional elements such as evaluating different optimizers to update model weights, the rate at which weights are updated (learning rate), the number of samples in each pass through the network (batch size), and number of times to pass the data through the network (epochs) are just a few of the items in each network that must be finely tuned (Emmert-Streib et al., 2020). While these elements are precisely what make these networks so flexible and able to learn highly intricate patterns in complex data structures, thoughtful consideration is required to avoid becoming stuck attempting to evaluate millions of architecture/hyperparameter combinations.

Neuroimaging specific issues

While this dissertation focuses on the evaluation of deep neural networks under the of assumption that there is information or relationships within these massive neuroimaging modalities that we are not able to readily manually extract, identifying predictive features from images containing millions of pixels can be compared to “finding a needle in a haystack”.

Models capable of loading these data require immense computational resources and large amounts of memory.

Furthermore, neuroimaging data involves a substantial amount of preprocessing. Variability in how researchers process data used as input for these models such as the amount of registration to a standardized template, denoising strategies, motion correction, intensity normalization, and artifact removal (some of which have poorly agreed upon strategies), all effect model performance and reproducibility (Aurich et al., 2015; Dular et al., 2023a). Substantial batch effects also exist within large multisite consortium studies such as different brands of MRI scanner, a fact some researchers leverage to create specific architectures to remove these confounding elements (Bento et al., 2022; Hu et al., 2023).

Furthermore, recent publications such as the landmark Marek et al., (2022) study have highlighted the considerable sample size required to identify meaningful and replicable effects within brain-wide association studies. This information calls into the question the large number of studies that predict clinical outcomes using neuroimaging data with sample sizes fewer than 100 subjects - an effect that is potentiated by the complexity and size of the data being used. However, there is still a robust body of literature examining the potential utility of these emerging methodologies with neuroimaging datasets (Table 1).

Author	Application	Subjects	Modality	Architecture	Perf
(Kuang & He, 2014)	ADHD	449	fMRI	DBN	.41-.81
(Ulloa et al., 2015)	SCZ	198	fMRI	DFCNN	.75
(Pinaya et al., 2016)	SCZ	191/198	sMRI	DBN	.73
(Zou et al., 2017)	ADHD	336	fMRI	DBN	.64-.79
(Farzi et al., 2017)	ADHD	429/239	fMRI	3D CNN	.66
(Sen et al., 2018)	MDD	24/24	fMRI	Autoenc. + CNN	.95
(Sen et al., 2018)	ADHD ASD	491/279 573/538	fMRI + sMRI	Autoencoder	.64-.67
(Matsubara et al., 2019)	SCZ	122/50	fMRI	DFNN	.76
(Pinaya et al., 2019)	SCZ	40/35	sMRI	Autoencoder	.64-.71
Rahman 2021(76)	ASD SCZ AZD	255/314 151/160 186/186	rs-fMRI	2D Conv + LSTM	~64% ~78% ~65%
He 2021(77)	Age	16,705	T1	3D Resnet	3.00 / .98
Peng 2021(78)	Age	14,503	T1	3D CNN	2.14
Kim 2021(79)	ADHD	776	rs-fMRI	CNN	71%
Lee 2021(15)	Age	2,349	FDG-PET T1	3D Densenet	0.85 0.804
Ayyar 2021(18)	AZD	58/48	T1	3D CNN	
Leming 2021(80)	Sex	14,683	rs + task fMRI	CNN Ensembles	0.84
Si 2021(81)	Epilepsy	30/33	DTI	Inception Resnet	92%
Kang 2020(82)	EMCI	50/70	sMRI, DTI	VGG-16	94%
D'Souza 2021(83)	Multi	150	rs-fMRI, DTI	LSTM-ANN	-
Joo 2023	Age	3004	T1	CNN	.93
Qiu 2022	AZD, MCI	971/369	T1	CNN	.95

Table 1: A Snapshot into the Deep Learning Literature

Transfer Learning and Domain Adaptation

The overarching definition of transfer learning is simply re-using (i.e., transferring) knowledge learned in one model to another. This "knowledge" is most commonly in the form of learned representations, in the form of trainable parameters (weights). There are several applications for this methodology. In domain adaption a model trained in one domain is adapted to perform well in another, such as training a model to classify images containing cats or Hyundai Elantra's to the similar but different task of classifying bears from semi-trucks. The benefits could arrive in multiple properties, if the primary goal is the classifier predicting bears from semi-trucks, but we have a relatively small sample size for this task, we can leverage the, hypothetically, larger dataset containing images of cats and Elantra's. That is, we believe that learned representations and relationships within the low-level features of each dataset may be similar (i.e. both contain tires, animal fur, claws, door handles, etc.).

In a recent example of this application, researchers found that self-supervised pre-training within different fMRI tasks both improved final model performance, and the pretrained models managed to converge using roughly 10% of the data required to achieve convergence using models without pretraining. In another example researchers found enhanced classification accuracy of classifying patients with schizophrenia (SCZ) patients from healthy controls (HC) using resting-state functional connectivity (rsfc) by initializing weights for the supervised model using those from unsupervised stacked autoencoders (AE) (J. Kim et al., 2016).

Yet another theoretical opportunity of transferring knowledge is that of learning tasks of *iterative complexity*, sometimes called curriculum learning or continued learning. In the seminal work "Curriculum Learning" (Bengio et al., 2009) researchers highlight the ability to improve

generalization and speed of model convergence using deep deterministic and stochastic neural networks in both language models and object recognition by learning gradually more complex concepts in a manner emulating that of how humans learn. Furthermore, by using a top-down design and decomposing a single complex problem into more granular tasks of lower complexity, we reveal new opportunities for embedded interpretation. For an illustrative example, if we have the task of assessing the longevity of a vehicle, we may benefit from having individually trained specialist mechanics (models) assessing, in isolation, distinct subsystems within our global system such as the suspension system, fuel delivery system, and engine. Subsequently, we can aggregate the assessments (predictions) from each specialist to both achieve the overall goal and determine the importance of each in arriving at that assessment. This embedded interpretability can act as a sort of "consolation prize" in situations in which those individual assessments are themselves not interpretable, as is the case of the "black box", or lack of inherent interpretability, nature of deep learning.

In summation, while there has been a lot of excitement surrounding DL and its potential utility in these large-scale complex neuroimaging modeling problems, it is critical to evaluate the positive aspects as well as the limitations of these emerging methodologies.

Chapter 2: Multimodal Analyses

Materials and Methods

The Adolescent Brain and Cognitive Development Study

The data used in these analyses comes from the landmark Adolescent Brain and Cognitive Development Study (ABCD). The ABCD study comprises 11,872 children aged 9-11 at baseline, enrolled in the 21-sites across the US (https://abcdstudy.org, Release 3.0). All caregivers and children provided written informed consent/assent for participation. All study procedures were approved by an Institutional Review Board. Sampling, recruitment, inclusionary/exclusionary criteria, and assessment measures for the ABCD Study have been described in detail previously (Auchter et al., 2018; Garavan et al., 2018; Volkow et al., 2018). This large multi-site study is comprised of a sample that reflects the socioeconomic status, racial identity and ethnicity, and sex assigned at birth of each study site city. Measures collected at baseline include structural and functional neuroimaging scans, as well as a variety of demographic, neurocognitive, and behavioral information. Participants without either structural or functional neuroimaging data that passes quality control, measures from the Child Behavior Checklist (CBCL), and items for the NIH toolbox were not included in these analyses.

Subject Demographics

	N	**%**
N	6037	
Age	9.8	*0.6 (SD)*
	Sex	
Male	3032	*50.2*

Female	3005	*49.8*
	Race/Ethnicity	
Asian	108	*1.8*
Black	748	*12.4*
Hispanic	1123	*18.6*
White	3470	*57.5*
Other	588	*9.7*
	Parents Married	
Yes	4246	*70.3*
No	1752	*29.0*
	Parent Highest Edu.	
< Highschool	210	*3.5*
Highschool/GED	483	*8.0*
Some College	1574	*26.0*
Bachelors	1634	*27.1*
Graduate	2132	*35.3*
	Household Income	
Income <= 50k	1884	*31.2*
50k < Income < 100k	1643	*27.2*
Income >= 100k	2457	*40.7*
	MRI Manufacturer	
Siemens	3952	*65.4*
GE	1440	*23.9*
Phillips	645	*10.7*

Table 2: Participant Demographics Chapter 2

The NIH-Toolbox and EF

Executive functioning, sometimes referred to as cognitive control, comprises a set of processes that includes working memory (WM), set-shifting (SS), and inhibitory control (IC). These processes work together to allow individuals to complete tasks, as well as set and achieve

goals. Additionally, dysfunction within components of executive function have been consistently implicated in a variety of mental health disorders, including depression, bipolar disorder, attention deficit disorder, conduct disorder, schizophrenia, autism, and obsessive-compulsive disorder (Cordova et al., 2020; Espy et al., 2011b; Flores et al., 2022; Friedman & Robbins, 2021; Nigg, 2017b; Shahrokhi et al., 2017; Strauman, 2017). The *NIH Toolbox for Assessment of Neurological and Behavioral Function* was created to provide a consistent and reliable way to assess neurocognitive functioning (Achenbach & Ruffle, 2000; Magyar & Pandolfi, 2017). The toolbox comprises the work of over 250 researchers from around the world to provide a critical resource for neuroscience researchers. It seeks to establish large-scale standardized methods of collecting measures of cognitive, emotional, sensory, and motor function. The NIH Toolbox Flanker and Inhibitory Control Test©, a measurement of "visuospatial inhibitory attention" represents inhibitory control, or the ability to suppress immediate desires or habitual responses in favor of more appropriate or goal-oriented behaviors. Sometimes referred to as "switching", set shifting refers to an individual's ability to alternate focus and attention between tasks and rule-sets. This construct is being assessed by the NIH Toolbox Dimensional Change Card Sort Test©. Weintraub et al., (2013) describe four components of working memory including, the ability to aggregate and process information from a given set of tasks, subsequently retain this information in a "short-term buffer", maneuver and modify the information, and hold the said modified information in the buffer. Working memory is assessed by the NIH Toolbox List Sorting Working Memory Test©.

Measures of Psychopathology

The instrument used to measure psychopathology for this study is the 119-item parent reported Child Behavior Checklist (CBCL). Designed to measure behavioral and emotional problems in youth (Achenbach & Ruffle, 2000), it contains questions about the child's mental and physical health over the past 6 months using a rating of 0: Not True, 1: Somewhat True, or 2: Very/Often True. From these questions composite scales are created for withdrawn, somatic complaints, anxious/depressed, delinquent behavior, aggressive behavior, social problems, thought problems, and attention problems. To calculate the composite dimensional p-factor score we used existing structural models validated within the ABCD Study sample. Specifically, we used a bifactor CFA p-factor model identified within the literature (Clark et al., 2021; Sripada et al., 2021). In this factor structure, a broader internalizing factor loads onto withdrawn, somatic complains, and anxious/depressed subscales, and an externalizing factor onto delinquent behavior and aggressive behavior. The P-factor loads onto each of the eight CBCL subscales (see Figure 7 below).

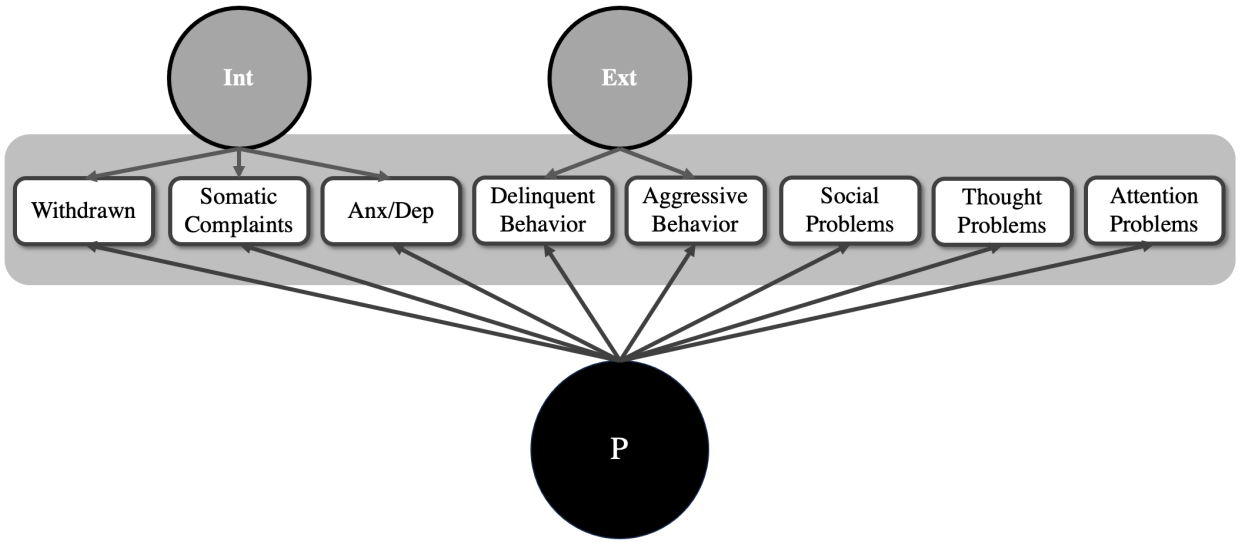


Figure 7: Structure of the P-factor

It is important to note that in both our analyses and the literature, the relationship between the constructed p-factor and a measure of "total problems" (sum of all item level responses from the CBCL) is strong. To approach a more normally distributed distribution of generated p-factor scores, we log normalized the p-factor. Important to note (Figure 8 below) is the large portion of zeroes from the CBCL responses. Puzzling, this large (~500) sample of participants endorsing not a single item from the entire 119-item CBCL questionnaire creates problems both for any non-zero inflated modeling strategies and conceptually, as it is unlikely an individual is entirely free from all mental, behavioral, or physical disorder symptoms. Thus, we evaluate both models in which we include and exclude this large portion of zero response subjects.

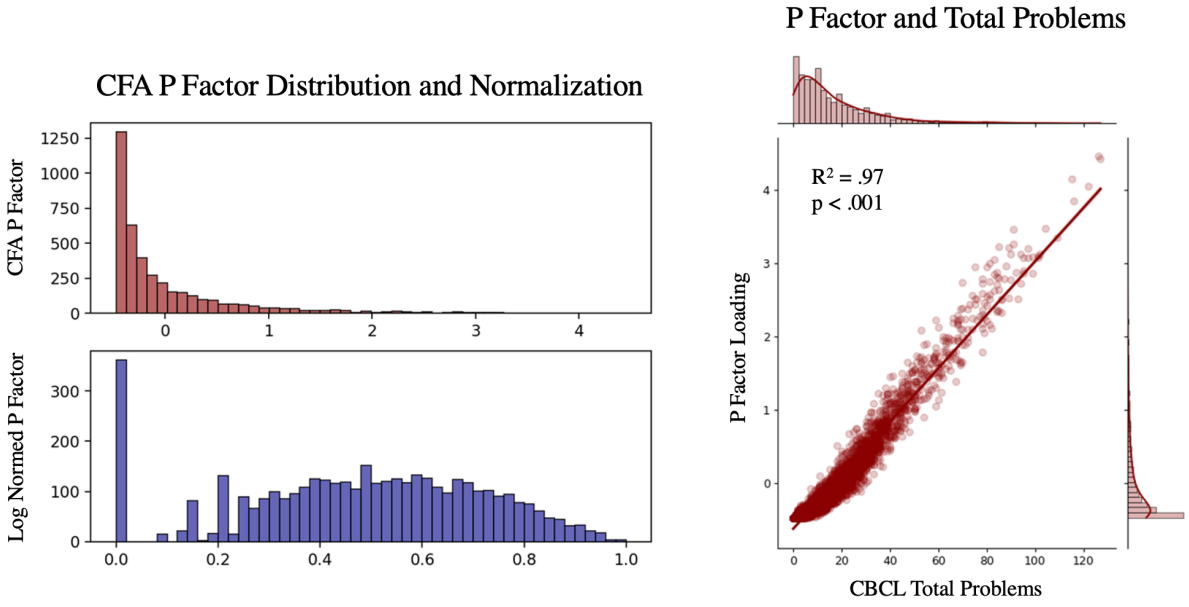


Figure 8: P-factor and CBCL Total Problems

Top left shows the distribution of the CFA p-factor loading both before and after log normalization (bottom left). Additionally, we see that the relationship between the P factor loadings and total CBCL problems is exceedingly high.

One of the primary underlying assertations in these analyses is that measures of executive function (EF) are correlated with psychopathology. While heavily supported by the literature (*see Executive Function and Mental Health: Chapter 1)*, it is important to evaluate this hypothesis early in our sample. Figure 9 below portrays that there are *weak ($p < .001$),* but statistically significant negative relationships between the p-factor and set-shifting (SS), working memory (WM), and inhibitory control (IC) indicating that greater scores from these EF measures are associated with lower p-factor scores, supporting findings from the literature. Furthermore, distributions for SS and IC are relatively normal, with SS having slight skew for larger values, whereas WM is almost bi-modal in nature. In addition to examining the predictive performance of a single composite measure psychopathology, we will also be examining our ability to predict the broader internalizing (INT) and externalizing (EXT) dimensions using the CBCL derived t-scores to determine if a specific subset of symptom categories yields higher predictive accuracy.

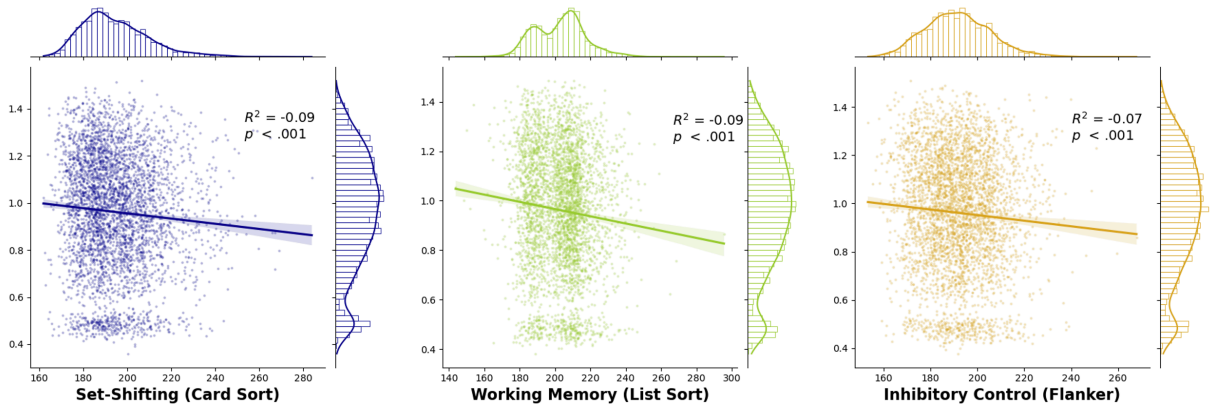


Figure 9: Relationship between EF and P-factor

Image Acquisition

Three different scanner platforms (3T Siemens Prisma, General Electric MR750, and Philips instrument) were used across the 21 sites within the ABCD study. Parameters for each scan protocol were harmonized across platforms (Casey et al., 2018). Individual scan sessions included a localizer, 3D T1-weighted MRI, 3-4 five minute runs of resting-state fMRI (rs-fMRI), diffusion weighted imaging (DWI), 3D T2-weighted MRI, and three task-fMRI scans. For rs-fMRI scans individuals fixated on a single focal target, and head motion was assessed in real-time using Framewise Integrated Real-time Motion Monitoring (FIRMM) (Dosenbach et al., 2017). Depending, on the amount of motion during the rs-fMRI scans, participants completed either three or four scans.

The data release used in these analyses came from the ABCD-BIDS Community Collection (ABCC; NDA Collection 3165). To promote accessibility in accordance with FAIR (findability, accessibility, interoperability, and reusability) data principles and support reproducibility, both raw and processed imaging data, adhering to the Brain Imaging Data Structure (BIDS), were provided (Feczko et al., 2021). Processing of fMRI data was accomplished using a modified version of the widely available and utilized Human Connectome Project (HCP) pipeline (Glasser et al., 2013). Modifications include the use of Advanced Normalization Tools (ANTs) for denoising and N4 bias field correction, the removal of artifactual motion from respiration, and adaptation of scanners unique to the ABCD study. The code and details are publicly available at https://collection3165.readthedocs.io/en/stable/.

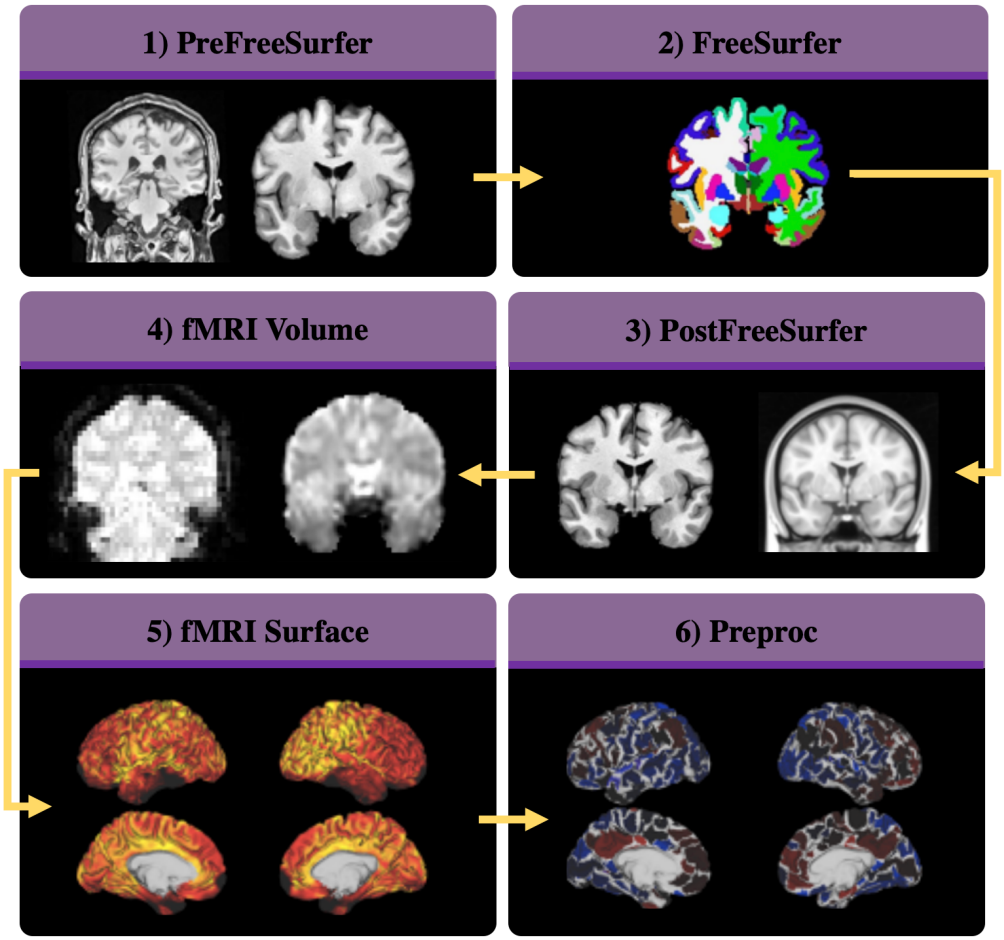


Figure 10: ABCD-BIDS Processing Pipeline (Adapted From: Feczko et al., 2021)*: Overview of the Adolescent Brain and Cognitive Development – Brain Imaging Data Structure (ABCD-BIDS) processing pipeline. 1) Images undergo typical normalization processes including masking, denoising, bias correction, and registration. 2) FreeSurfer, a set of neuroimaging utilities, segments both the cortical and subcortical anatomical data. 3) Conversion to the alternate surface-based cifti file format. 4) Functional data is registered to the standard MNI atlas and (5) subsequently converted into surface files. 6) Finally, the functional data are filtered for standard motion and respiration related motion artifact.*

T1 Structural MRI

For the evaluation of structural data, T1 sMRI was utilized in two ways. First, imaging data were used directly in DNNs and only underwent normalization, skull-stripping, and linear registration to the MNI152 template (Fonov et al., 2011). Additionally, data were downsampled (nearest neighbors' algorithm) using the python package *SciPy* (Virtanen et al., 2020) to evaluate aspects of lower dimensionality and model performance. Moreover, we compare the performance of models using imaging data directly against the more traditional strategy of FreeSurfer derived summary statistics (volume, surface area, mean intensity, etc.). The atlas used to obtain the summary statistics was the widely adopted Desikan-Killiany atlas (see Figure 11) which separates the cortex into 34 distinct regions in each hemisphere with highly accurate segmentation (Desikan et al., 2006). In addition to the cortical segmentations provided by FreeSurfer's primary processing pipeline are several subcortical structures including the hippocampus, caudate, putamen, pallidum, thalamus, and amygdala for each hemisphere. Furthermore, the morphological measurements (volume, surface area, thickness) were standardized using total intra-cranial volume (ICV) of the individual. While researchers have made arguments both for and against the correction of ICV, with what seems the majority correcting for ICV, recent analyses within the ABCD sample found no significant differences in performance predicting cognition, including WM, when correcting and not correcting for ICV within measures of surface area, gray matter volume, and cortical thickness (Dhamala et al., 2022). Thus, while a potentially unnecessary correction, in keeping with the majority of literature, we used ICV-corrected measures of morphology.

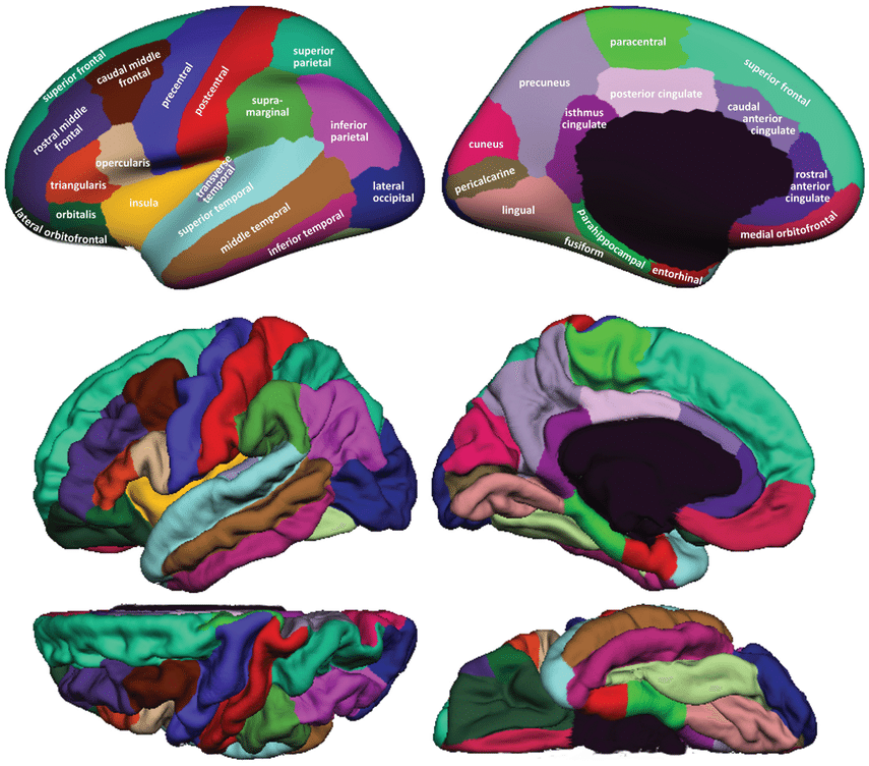


Figure 11: Desikan-Killiany Structural Atlas (Figure Credit: Klein & Tourville 2012)

Sometimes referred to as the Desikan-Killiany-Tourville (DKT) atlas, the Desikan-Killiany atlas partitions the cerebral cortex into 34 unique regions per hemisphere, a total of 68 regions across the entire brain. Segmentation is based on the cortical topography defined by gyral and sulcal patterns via a combination of structural MRI data, anatomical landmarks, and expert neuroanatomical insights.

Resting-State fMRI

As mentioned previously, three to four five-minute runs of resting state are acquired. Initial standard processing of the rs-fMRI runs includes temporal de-meaning and de-trending, the use of a general linear model (GLM) for the denoising of white matter, cerebrospinal fluid (CSF), global signal, and movement regressors. Finally, ACPC-alignment is applied to the

subject's native space data. The respiration motion filter is applied after standard processing (Fair et al., 2020).

A substantial component of rs-fMRI data filtering involves *frame censoring*. This process involves setting a minimum movement threshold of framewise displacement (FD) for each full volume (time to repetition: TR), we use the traditional .2 mm FD (Feczko et al., 2021). If a given TR exceeds this movement threshold that entire TR is removed (censored) from the resulting rs-fMRI time series. Finally, the remaining time from the multiple rs-fMRI runs are concatenated. Furthermore, we included only subjects that had at least five minutes of rs-fMRI data *after* the application of frame censoring. The choice to use five minutes as the minimum threshold was selected in accordance with general guidelines of required resting state time necessary to unearth meaningful effects and minimize the number of subjects lost to exclusion based on this requirement (see Figure 12) (Van Dijk et al., 2010).

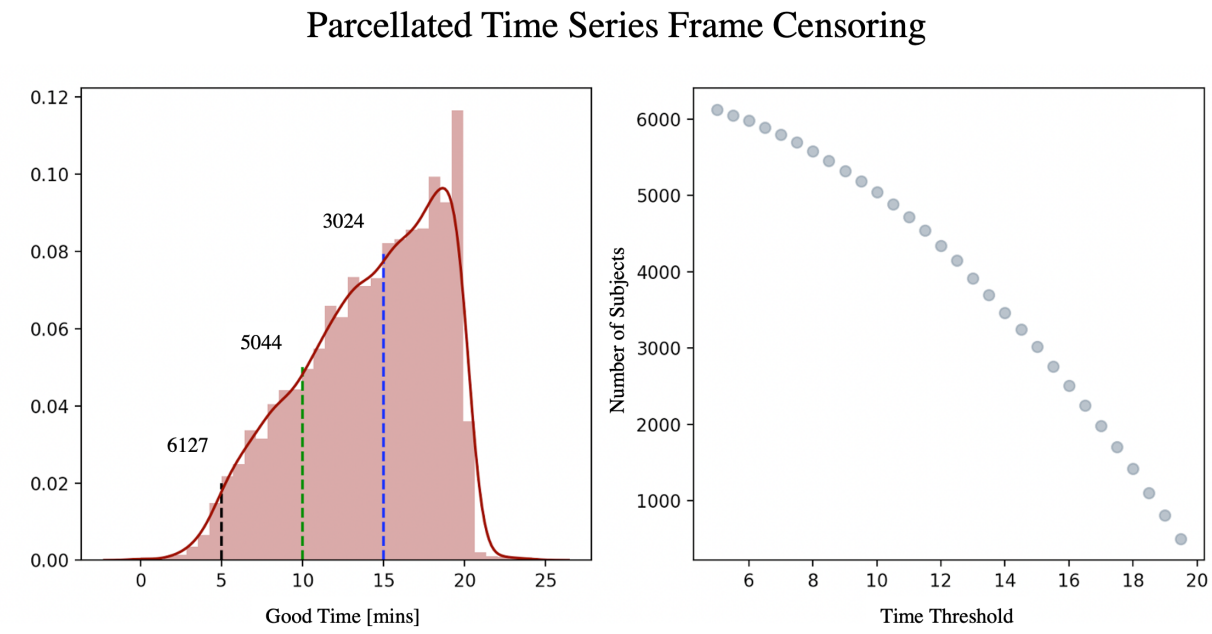
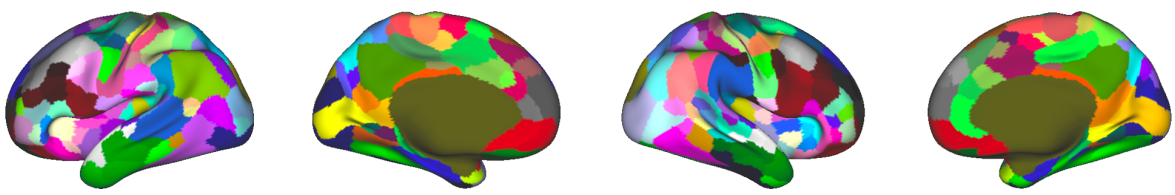


Figure 12: Frame censoring of rs-fMRI

Left displays the distribution of subject rs-fMRI "good-time", time left after frame censoring at a framewise-displacement (FD) threshold of .2mm. Black, green, and blue lines indicate the number of subjects left at 5, 10, and 15 minute thresholds respectively. Right also highlights the relationship between the number of subjects remaining after different time thresholds.

The final aspect of the rs-fMRI processing pipeline is the generation of parcellated timeseries (PT-series). This involves the use of an atlas to parcellate or split the brain into distinct regions and averages the signal of voxels within each region at each TR. The atlas selected for use in this project is the Gordon-Subcortical atlas (Gordon et al., 2016). The Gordon atlas organizes the distinct cortical ROIs into *functional networks*. Representing large-scale patterns of connectivity and brain activity, these networks are comprised of structures involved in distinct cognitive and functional process. Specifically relevant to our analyses is that of the prefrontal cortex and structures within networks involved in higher-order cognitive functioning and attention, such as the frontoparietal as well as ventral and dorsal attention networks. The atlas we chose to use combines the 333 ROI Gordon cortical atlas in addition to several subcortical regions such as the thalamus, ventral striatum, putamen, and caudate as well the cerebellum (Seitzman et al., 2020).

Gordon Atlas



L **R**

Figure 13: Gordon Resting State Atlas

Feature Selection

Due to the large number of features obtained through the processing of the neuroimaging data, we elect to evaluate three different strategies for reducing the number of features fed into each model. With the high multicollinearity (features being highly correlated themselves such as surface area and volume) and enormous computational demand required to train these large datasets, we believe it is prudent to evaluate and justify inclusion of the full feature set. Given n observations and p features yields a training time in the order of $O(np)$ or $O(np^2)$, depending on hyperparameters for the linear models and larger for the DNNs. Furthermore, rs-fMRI has an inherently low temporal signal-to-noise ratio (tSNR) and could, if not cautious, obfuscate models with unwanted noise.

Meta-Analytic Prioritization

The first method of feature selection involves selecting only regions identified by a meta-analytic prioritization strategy. Our method of meta-analytic prioritization is captured in Figure 14 below. A search term (or query) is entered for both NeuroQuery and NeuroSynth meta-analytic databases. We chose to leverage both platforms and their regional agreement due to the different strategies when used to associate queries with brain regions, ideally returning higher confidence regions. In NeuroSynth, each voxel (3D pixel) in the brain map represents the likelihood of activation associated with a query. The values are z-scores or probabilities that indicate how likely it is to observe activation at that voxel for studies associated with the queried term (Yarkoni et al., 2011). In contrast, NeuroQuery uses a predictive model to generate brain

activation predictions. The values in a NeuroQuery brain map represent the predicted level of activation for a specific term or concept based on trained models (Dockès et al., 2021). These values are often normalized and can be interpreted as the model's confidence in the predicted activation.

Once the raw query maps are obtained from each platform, we applied a threshold to retain the upper half of the NeuroQuery activation map, and binarized each query map before combining the two. The resulting parcellation schema, Gordon + Subcortical for rsfMRI and Desikan-Killiany, was leveraged to identify parcels, in this illustration, those with at least 50% overlap between a parcel and the query map yields a meta-analytic parcel.

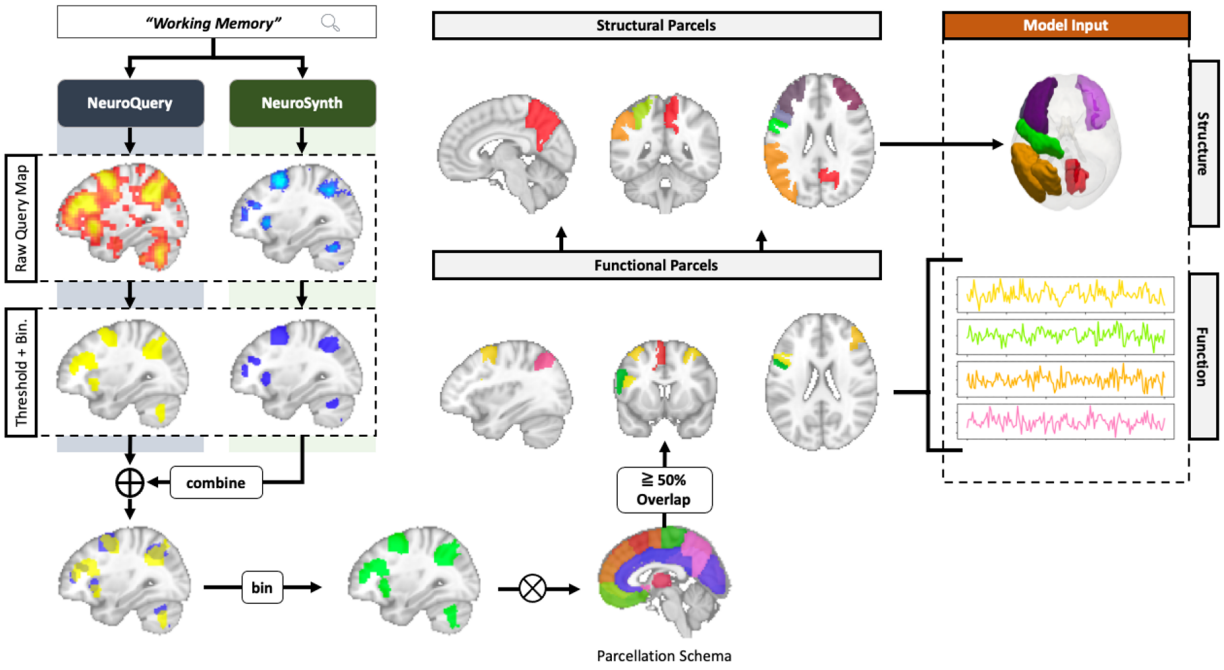


Figure 14: Meta-analytic Prioritization

In Figure 15 below we illustrate the effect of the selection of overlap thresholding between the query map and selected parcels, we evaluate several thresholds within the range displayed by the red rectangular box (dashed red line indicating the "elbow" of the plot). Only data from these meta-analytically prioritized structural parcels and functional regions are used in the "meta-analytic prioritization" set of selected features.

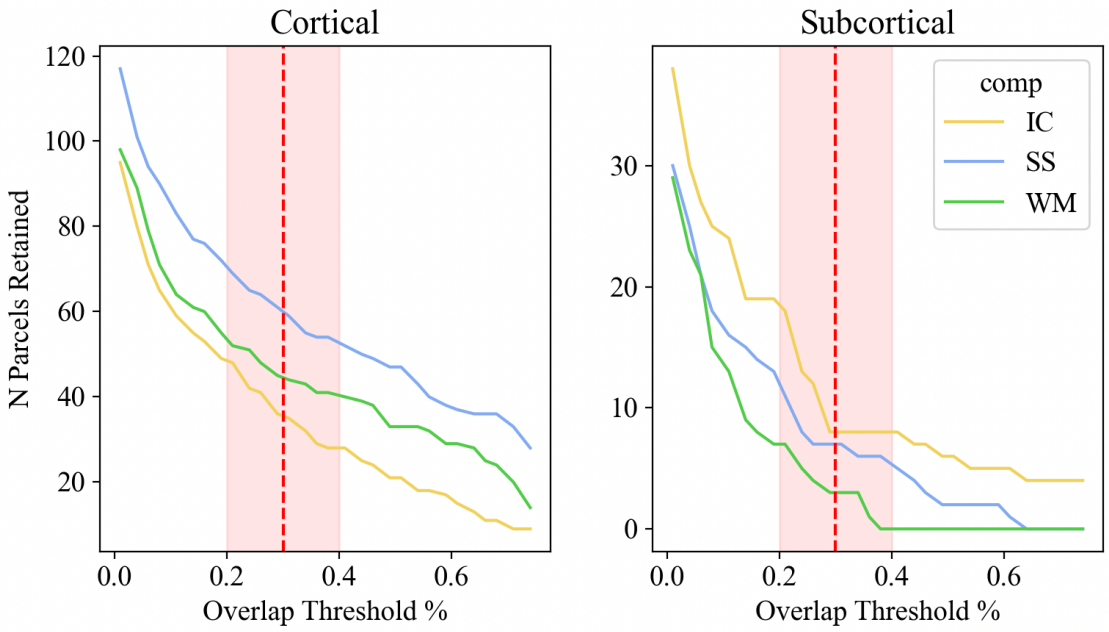


Figure 15: Refining the Meta-Filter

The number of cortical and subcortical regions remaining for each of the components of executive function (EF), working-memory (WM, green), inhibitory control (IC, yellow), and set-shifting (SS, blue). The light red patch indicates a range of values we evaluated for the overlap threshold centered around a roughly visualized "elbow" (red vertical dashed line).

Emerging Meta-Analytic Regions

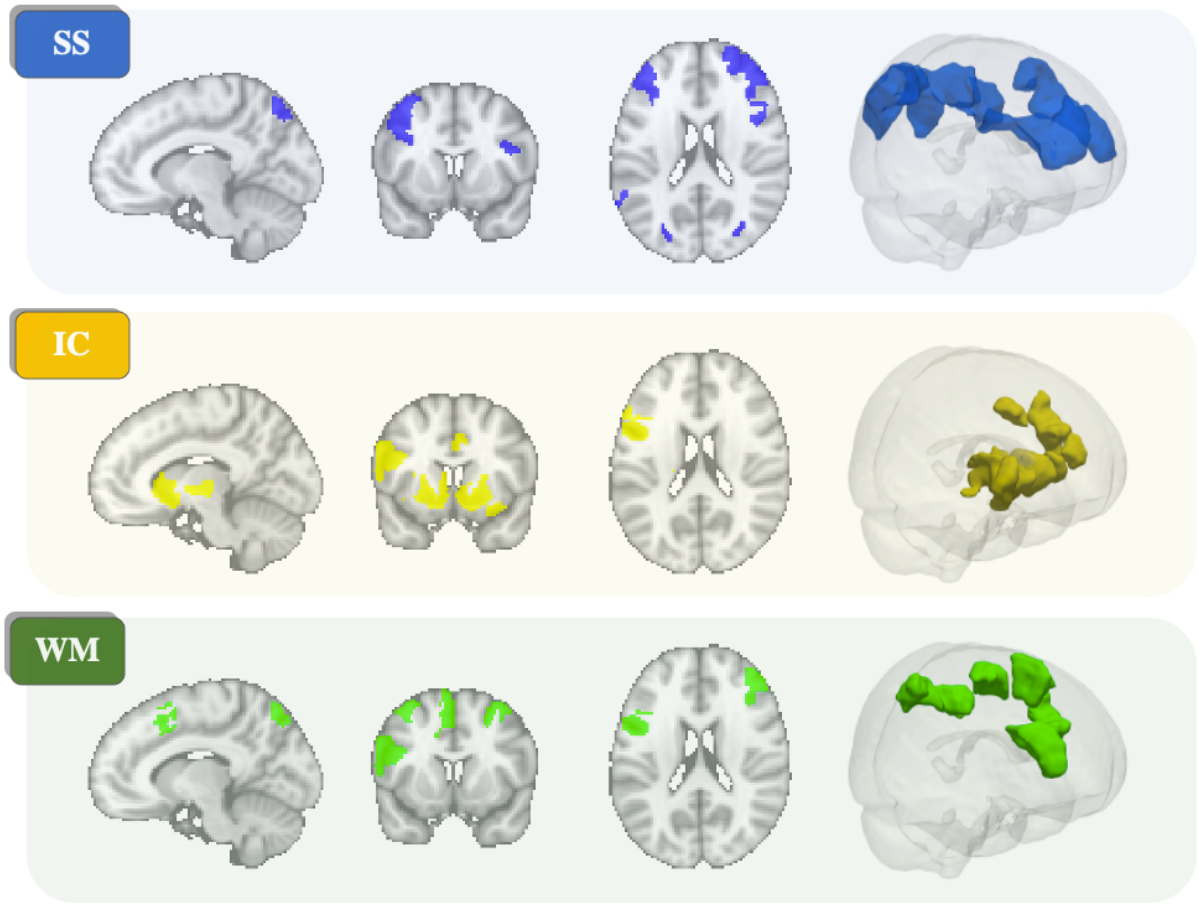


Figure 16: Meta-Analytic Regions of EF

Top row: Emerging regions from set-shifting are the superior frontal gyrus, inferior parietal lobule, and lateral occipital regions. Middle row: Inhibitory control includes regions of the basal ganglia including the nucleus accumbens, global pallidus and ventral caudate in addition to the thalamus, and precentral gyrus. Bottom row: WM regions include the middle frontal gyrus, inferior parietal lobule, and inferior frontal regions.

Variance and Random Filter

In addition to the meta-analytic prioritization feature selection method, we used both a variance filter and a *random* filter (arbitrarily selecting features) to select the same number of features identified by the meta-analytic prioritization feature selection method, assuring all feature selection subsets had the same number of features. While the meta-analytic feature selection method is sample agnostic (not informed by a given subset of the data), the variance- and random-filters were obtained at each split of the training data.

Data Partitioning and Normalization

The train, validation, and test sets were created using a 4:1:1 ratio (n's = 4003, 1017, 1017) for model training, hyperparameter tuning, and evaluation respectively. The test set contains a group of subjects that is consistent throughout the entire analyses (global test), but the train and validation were created by randomly sampling from the dataset ten-times to obtain different splits of train and validation sets. All splits were created and verified to have no significant differences between sex assigned at birth, race/ethnicity, MRI scanner manufacturer, the NIHTbx EF measures, and the established p-factor. Additionally, all data were z-scored (removing the mean and scaling by the unit variance). This data partitioning schema was chosen to allow for the evaluation of final model performance using a distribution of test performances. The ten models trained on unique samples of training data result in ten global test performance metrics, allowing to statistically (through ANOVA's) evaluate *significant* differences in performance from different methods of both feature selection and model type (see Figure 17 below).

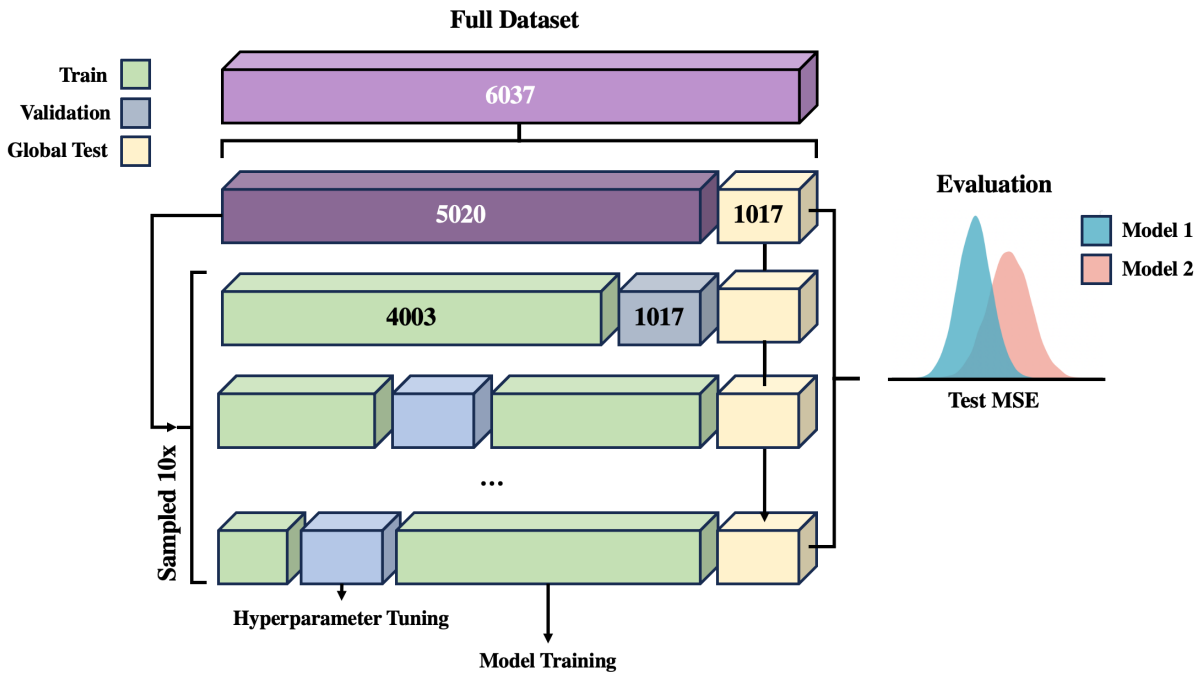


Figure 17: Data Partitioning

Modeling Strategies

Traditional Statistical Methods

In the context of these analyses, standard machine learning (SML) refers to anything that is not a DNN trained with backpropagation (the method used to update weights in DNNs). All standard methods used are linear (no non-linear functions or kernel basis functions). Each selected method accomplishes a different modeling strategy and objective. All SML models were created using the scikit-learn (*sklearn*) python software (Pedregosa et al., 2012) . Standard linear regression (LR) was the first method to evaluate, and subsequently penalized Lasso (LSS) regression. Additionally, we used both principal components analysis (PCA) and partial least squares regression (PLSR) as methods to achieve unsupervised (PCA) and supervised

projections into a reduced dimensionality before prediction. In the case of PCA the resulting components were then fed into linear regression, evaluated with and without regularization, after transformation, each of these methods are summarized below. Each method used accomplishes a slightly different goal in model prediction. LR is the fundamental model minimizing loss via mean squared error (MSE, Equation 1) with no constraints.

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2$$

Equation 1: Mean Squared Error

Additionally, Lasso includes a regularization parameter on the L1 norm of the regression coefficients (Equation 2) encouraging "sparsity" within the coefficients, acting as a sort of feature selection and penalizes models with many predictors, pushing some of the coefficients toward zero. The objective is to minimize loss (MSE) with an additional error term, controlled by parameter λ, that penalizes the absolute sum of coefficients in the model.

$$\frac{1}{n}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{M}|\widehat{\beta_j}|$$

Equation 2: Lasso Penalty

The tunable parameter λ penalizes the absolute sum of the coefficients β_j and is added on to the standard MSE loss term.

Both PCA and PLS project the original feature space into a lower dimensionality of *N* components (selected hyperparameter) using a linear combination of the original feature space. PCA seeks to identify latent projections that both maximize the variance within said components while also being uncorrelated. For our purposes we will be evaluating two methods of selecting the number of latent components, N_{comp}, first by selecting $\sqrt{n_{features}}$ and also N_{comp} such that we retain ~50% of the variance within the data.

$$\max_{W_1}(W_1^T X^T X W_1),$$

$$s.t: W_2^T W_1 = 0$$

Equation 3: PCA Objective

The goal is to maximize the variance explained by the first principal component using the vector of weights W_1 of the dataset X, subject to the constraint of maintaining orthogonality between components. This forces the latent components themselves to be uncorrelated and helps to remove multicollinearity within a dataset.

In the supervised PLS, the objective is to instead maximize the covariance between the latent projections of the data, X, and target Y, under the constraint that each latent variable (LV) is unit-normalized and orthogonal to all other LVs.

Deep Multimodal Neural Networks

Yet another capability we will be evaluating is the capacity of neural networks to handle data from multiple sources in a joint manner. Described previously in this document, Transfer Learning and Domain Adaptation: Chapter 1, data from both T1 sMRI and the rs-fMRI

parcellated timeseries are used in a single multimodal network. Each modality passes through a series of either 2- or 3D convolutional layers before being flattened and passed into the subsequent linear layers and target prediction (see Figure 18). Additionally, we place constraints on the connections between the layers immediately after flattening such that activations from delineated sets of neurons represent T1 data embeddings only, rsfMRI embeddings only, and a combined *structure-function* set of embeddings. This constraint reduces the number of trainable parameters, constricts the flow of information in a modality-informed manner, and allows for modality specific latent representations (embeddings). This methodology is similar to that of the *visible neural network* (VNN) DCell from (Ma et al., 2018), placing constraints on the traditional fully connected neural network (FCNN) using prior knowledge. Ma and colleagues found that these constrained networks achieved nearly equivalent performance using a fraction of trainable parameters, while also allowing for inherent aspects of interpretability by creating biologically constrained latent representations of the input data.

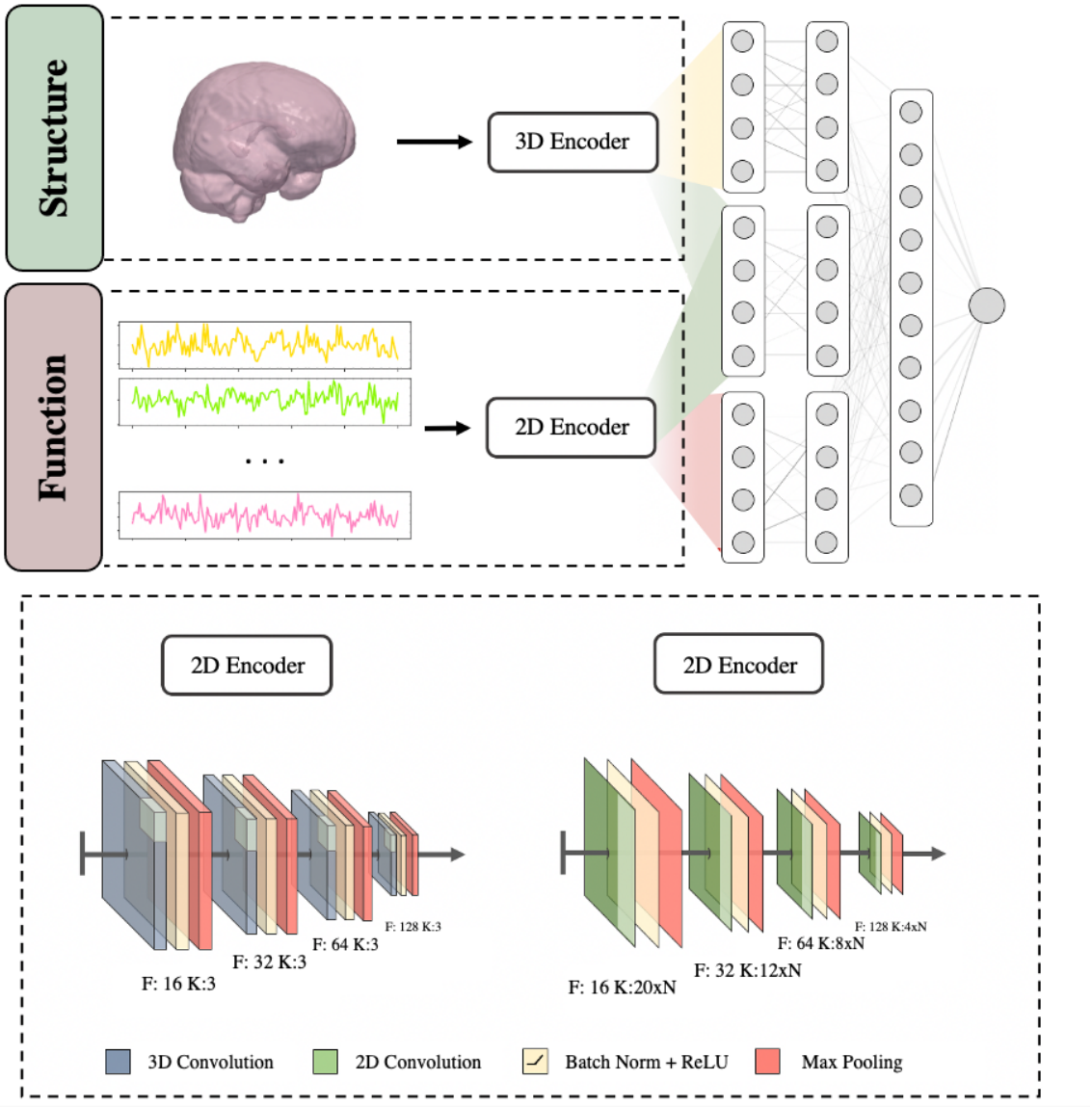


Figure 18: Architecture of the Multimodal Fusion Network

F represents the number of filters (sometimes called channels) being used, and K refers to the kernel size. In the 3D encoder kernels of size K, in the 2D encoder K is 2D with the size x by N (number of parcels in the resting-state fMRI parcellation atlas).

For the rs-fMRI data, time series from each parcel, undergoes "network-only" convolutions, meaning we extract the lowest level features from within each function resting

state network (i.e., default mode, frontoparietal, or salience networks). Constrained networks include the 1-dimensional time series of each of the *N* parcels included in that network. Thus, the size of the weight matrix is N_{TR} *x* $N_{parcels}$. The activation matrices from these operations are reduced using MaxPooling, and finally these network specific activations are concatenated, and joint convolutions are applied across these abstracted representations of all rs-fMRI networks, with the objective of identifying relationships among the abstracted representations of features between the networks. We chose to use convolutions for the timeseries data instead of recurrent neural networks (RNN) and long-term short-term memory networks (LSTM), because they are typically more computationally efficient than LSTMs, in most architectures requiring less memory and faster training times. Furthermore, the translational invariant nature of CNNs allows for the learning of common sets of local patterns and dependencies within the timeseries data.

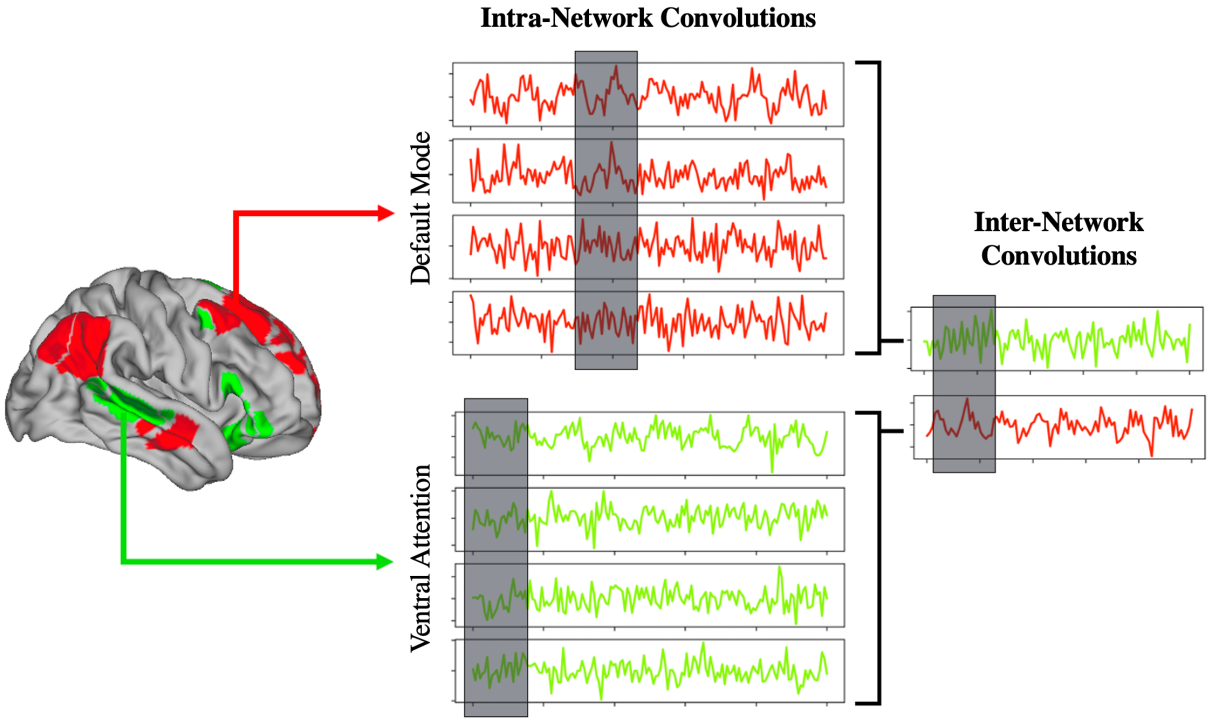


Figure 19: rs-fMRI Network Convolutions

A more detailed visual representation of the rs-fMRI network specific convolutions. The second dimension of the filter applied to each resting-state functional network is equal to the number of parcels in that established functional network only ("intra-network convolutions). After initial convolutional layers and MaxPooling the reduced set of embeddings is concatenated and undergoes joint or "inter-network" convolutions. This allows restricts the extraction of information by leveraging prior established knowledge from the field.

Statistic Derived Fully Connected Neural Networks

In addition to evaluating neural networks utilizing multidimensional convolutions on either the 2- or 3D imaging data, we also examine fully connected neural networks created to use the processed set of the neuroimaging features via either the FreeSurfer statistics of morphology and intensity or rsfc matrices. This strategy is an evaluation seeking to answer the questions of,

"Can DL better identify complex layered nonlinear relationships among extracted features?"

and

"Can DL identify features in data that we are unaware of or unable to manually extract?".

To further evaluate *knowledge informed,* i.e., constraining connections of layers based on *a prior* knowledge, we include an architecture in which the early layers of the network are connected only to other regions within each resting-state brain network, for rs-fMRI, and another set of layers contain connections only for the FreeSurfer derived summary statistics of T1 sMRI.

Activations from the constrained operations are concatenated later in the network into two fully connected layers before the prediction of the outcome (see Figure 20).

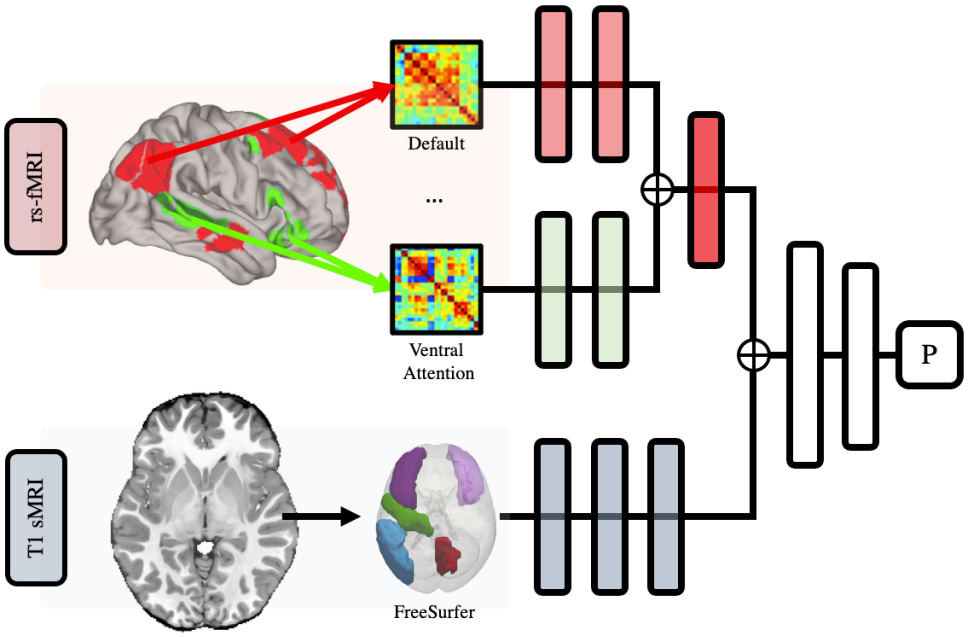


Figure 20: Statistic Derived Neural Network

Transfer Learning

Additionally, we evaluate the ability of transfer learning with DNNs. This evaluation strategy is captured in Figure 21 below. Individual models are trained to predict each of the components of executive function WM, IC, and SS. The primary assumption relies on the extant body of research highlighting the associations between each of these elements of EF and both structural and functional neuroimaging data (highlighted previously in Chapter 1). The aspect we are utilizing is again that of *curriculum learning*, the idea that we provide problems of increasing complexity and *transfer* that information learned from models predicting less heterogeneous outcomes and use them to try and improve prediction in a more challenging task. In this situation, aspects of EF have more replicable and reliable neuroimaging correlates, whereas

findings related to psychopathology are less consistent. Furthermore, EF targets lower level aspects of cognitive processes, unlike psychopathology which, despite evaluating via the dimensional p-factor, has enduring heterogeneity and variable presentation of symptoms. Thus, we seek to evaluate the ability to improve performance in predicting the p-factor of psychopathology by leveraging information from the models trained to predict EF.

The actuality of this strategy is the concatenation of each of the trained EF models and subsequent retraining to predict the p-factor. The hypothesis being this *transfer of knowledge*, via the initialization of weights, in the p-factor model performs better than that of a model in which we randomly initialize the weights. Additionally, we will evaluate the strategy of both *freezing* the weights in the early layers, meaning the weights are not updated and act as a sort of predefined feature extraction, in addition to initializing but allowing the weights in the early layers to be updated with the new objective of predicting the p-factor. Finally, we will also evaluate models using *no transfer learning* and architectures equivalent to that of the EF DNNs.

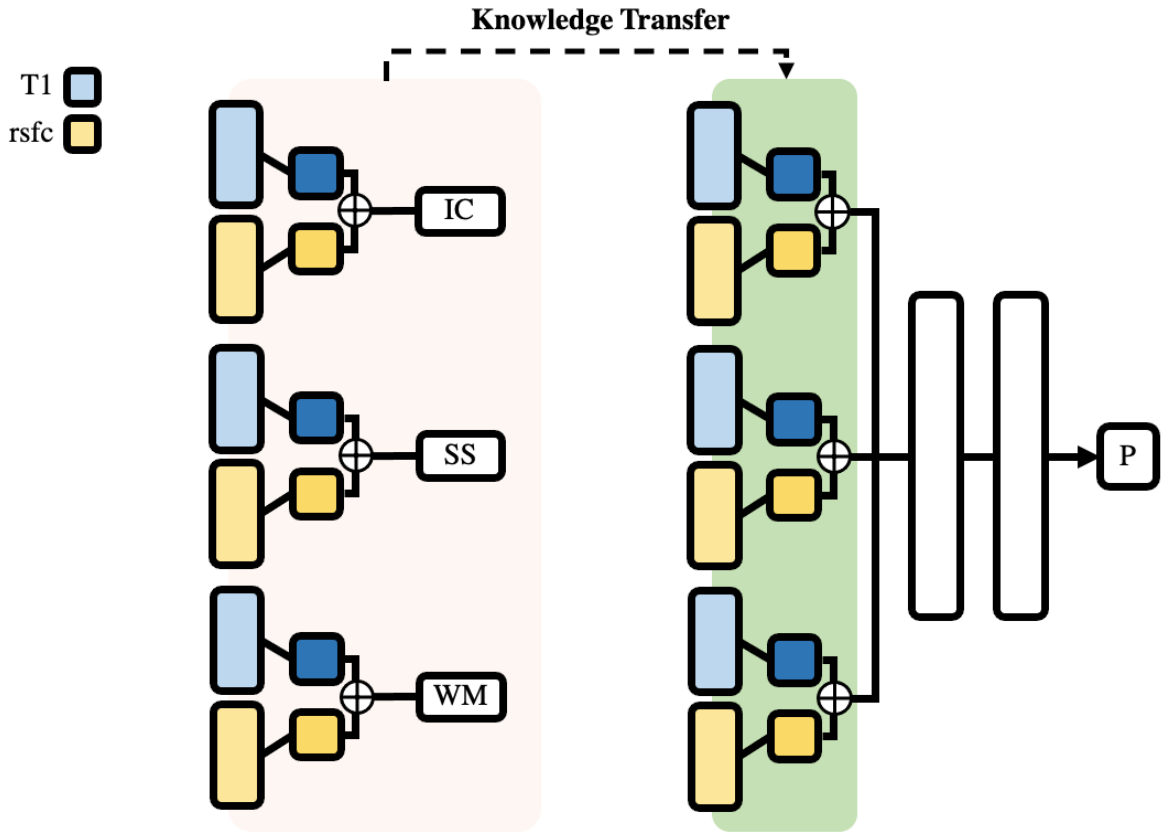


Figure 21: Transfer Learning

Individual models trained to predict each of the components of executive function (EF) are concatenated into a single network of these individual networks and two additionally fully connected linear layers are appended prior to the prediction of the p-factor.

Evaluation

To evaluate the performance and compare different modeling strategies, feature selection methods, and transfer learning tactics we, evaluate the performance of the global test set from the 10-splits of training data. Furthermore, we present both the Pearson correlation coefficient (r), and mean squared error (MSE), to keep consistent with previous research (Abrol et al., 2021;

Chen et al., 2022). We use both, as each method describes performance from a different perspective, and each has unique assumptions and benefits. MSE displays how close on average, the predicted values are to the actual values; it makes no assumptions about linearity between actual and predicted values but is not a standardized metric which can make it difficult to readily interpret. In contrast, *r*, which describes the linear relationship between the predicted and actual values and is scaled between -1 and 1. Furthermore, as we are particularly interested in comparing the performance of multiple model strategies, it is important to understand the strength and direction of relationship between actual outcomes and predicted outcomes. Additionally, we found MSE a more reliable cost function for model convergence within the DNNs over MAE or RMSE. Furthermore, to evaluate if a model's MSE and observed *r* is significantly better than what would be observed by chance, we compare the test MSE and *r* of the actual models against 1000 models trained using permuted data (shuffling the labels of the training outcome). A summary of both the null (H_0) and alternative (H_a) hypotheses for these methods is reported in Table 3 below. The H_a for *r* states that the slope of our coefficient is non-zero and positive, i.e., as our actual values increase so do our predicted values. The H_a for MSE relies on the permutation testing and states that our actual test MSE should be lower than some predefined percentage (α) of permuted test MSE's.

Pearson Correlation Coefficient (*r*)	**MSE**
$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$	$\text{MSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

H_0	$\beta_y \leq 0$	$\frac{\Sigma(MSE_{act} < MSE_{perm_i})}{n} \geq \alpha$
H_a	$\beta_y > 0$ one-sided	$\frac{\Sigma(MSE_{act} < MSE_{perm_i})}{n} < \alpha$

Table 3: Performance Metrics

Results

Predicting EF

The models predicting EF (WM, IC, SS) with the lowest MSE consistently came from using the FreeSurfer extracted features of the T1 sMRI data and were SML models (MSE's .90, .94, .74 respectively), see Table 4 below. However, for IC and SS when looking at *r* as our performance metric, PCA with both T1 sMRI and rsfc performed best. Additionally, only MSE of WM with the T1 was ***significantly*** ($p < .001$, actual test performance vs 1000 permuted models) lower than rsfc or both modalities combined. Conflictingly, we found that multimodal models (T1 sMRI + rsfc) had the best performance between actual and predicted, but not significantly better than rsfc alone, we explore these variable findings more in the discussion.

Interestingly, when predicting WM, all model/modality combinations performed significantly better ($p's < .001$) better than the distribution of permutation MSE's. However, in the case of predicting IC, only PCA + LR was significant ($p's <. 0001$) for each modality, and DNNs with rs-fMRI and both modalities were also significant. The same was true for SS with the addition of PLS with the T1 data being also significantly better than the distribution of permuted performances.

	MSE	*Model/Modality*	*r*	*Model/Modality*
WM	.90 ± .01	PLS/T1*	.28 ± .02	PLS*/T1*
IC	.94 ± .01	PCA*/T1*	.12 ± .01	PCA*/Both
SS	.74 ± .01	PCA/Both	.22 ± .01	PCA*/Both

Table 4: Top EF Prediction Performance

Top predictive performance (mean and standard deviation) via MSE and r for each of the components of executive function within the global test set. Asterisks after model or modality indicate that either that modeling strategy or modality had significantly (ANOVA $p < .001$) better performance than the other modeling or modalities being evaluated.

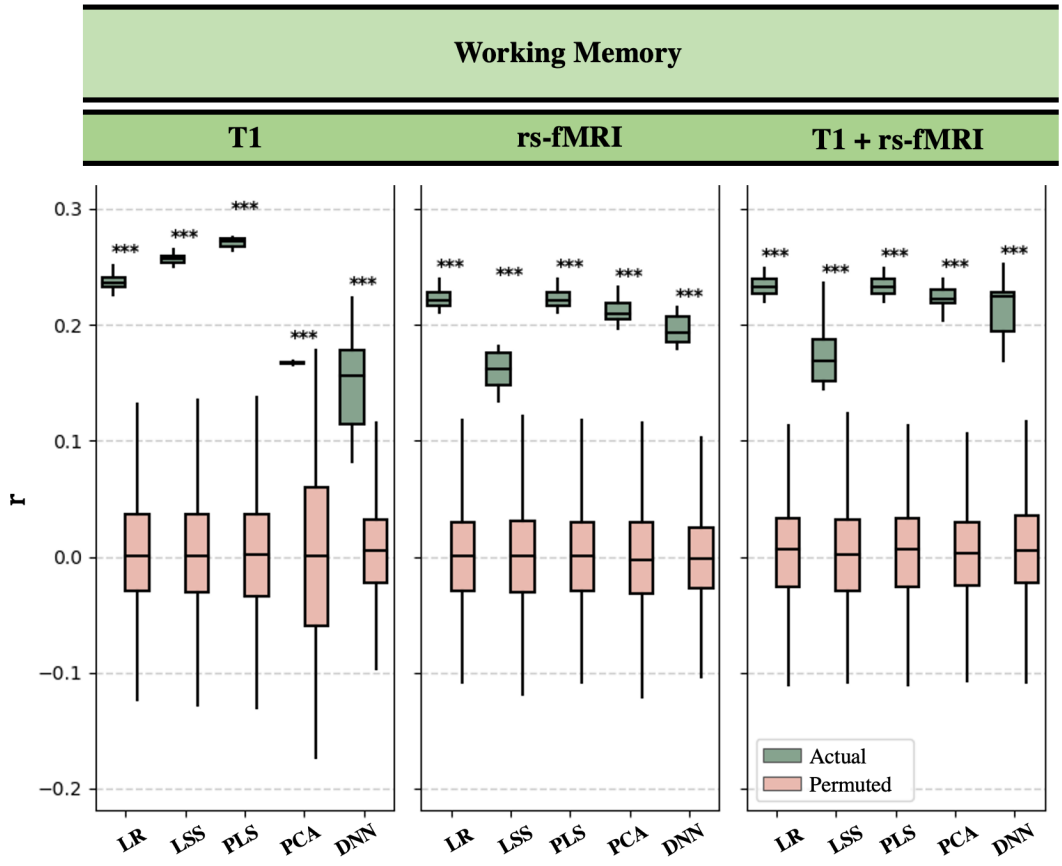


Figure 22: Multimodal WM Prediction

Global test performance of predicting working memory (WM) over the 10-folds of training data using unimodal (T1 or rs-fMRI) and multimodal (T1 + rsfMRI) data. Bottom axis indicates the various modeling strategies of LR (linear regression), LSS (lasso regression), PLS (partial least squares regression), PCA (principal components analysis followed by linear regression), and DNN (deep neural network). Y-axis represents model performance via the Pearson Correlation Coefficient (r) of predicted vs actual outcome. Green colored boxes represent actual test performance and peach colored boxes represent test performance from 1000 permuted models created using a training set after shuffling the outcome. All modeling strategies and modalities achieved lower MSE and higher r (p's < .001) than the permuted distributions (indicated by

****). Note that subsequent boxplots with olive and peach colored boxes all represent actual and permuted test performance.*

EF and Feature Selection

Despite several methods of evaluated feature selection, we found the best predictive performance when using the entire feature space, often most significantly when coupled with dimensionality reduction. Moreover, the variance filter obtained the best performance of the feature selection methods evaluated, and we saw no significant differences between the meta-analytic filter and *randomly* selected feature sets of the same size. Furthermore, among the models trained with feature selection, the DNN's performed significantly (p's $< .0001$) better than the SML models, but DNN's were not significantly better in predicting any component of EF when using the full feature set.

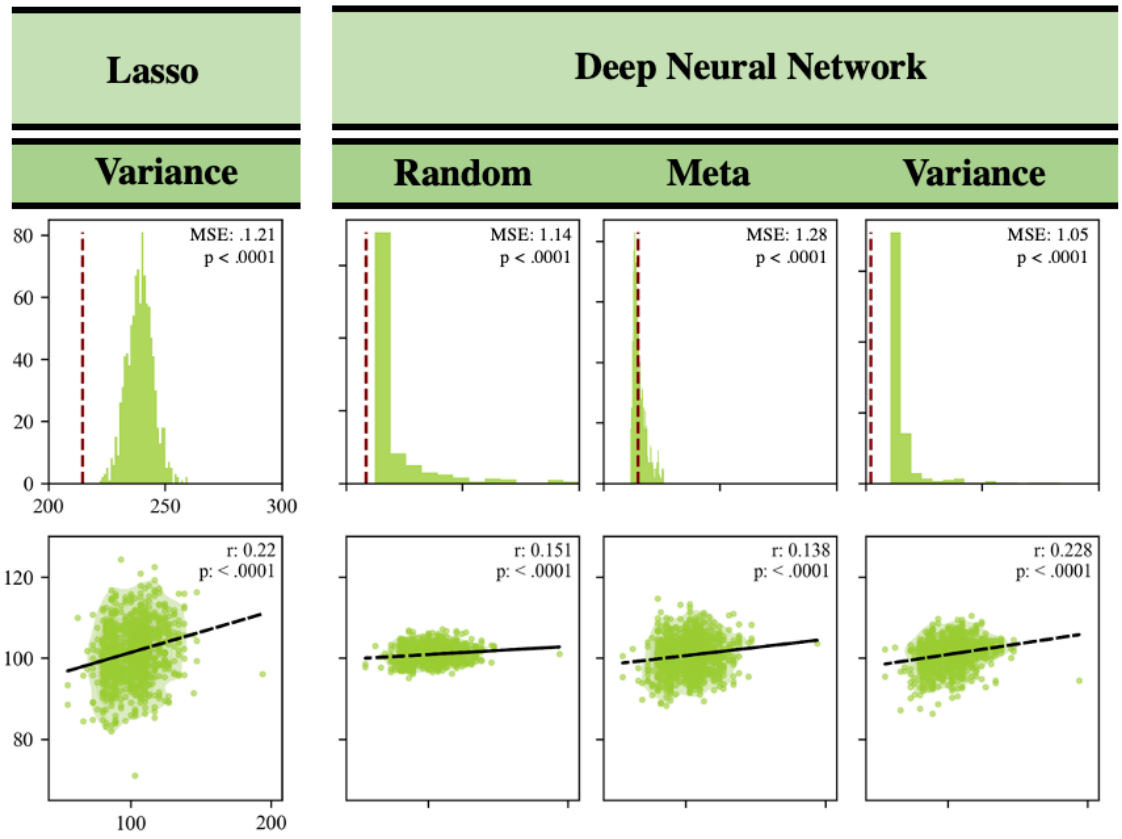


Figure 23: Feature Selection Type Predicting WM

Evaluation of SML (Lasso) and deep neural networks (DNN's) with different methods of feature selection (variance-, random-, and meta-analytic filters) in the prediction of working memory (WM). Axis ranges are shared across plots in each row. Top row: Mean global test MSE (red dashed line) and the distribution of 1000 permuted test MSE's, models in which the labels of the training data were shuffled. All feature selection methods in predicting WM using Lasso and DNNs were better than the permuted distribution (p's < .001). Bottom row: Actual (x-axis) vs predicted (y-axis) WM.

It is important to note the *range* of predicted values differed substantially between the SML and DNN models. While both Lasso and the DNNs utilize MSE as the loss function in

learning model weights, it is possible that the different optimizer algorithms, coordinate descent (Equation 4) used in Lasso and a modified form of gradient descent (Equation 5) (Adaptive Moment Estimation: ADAM) in the DNNs create these disparities. While this may seem a trivial distinction, these underlying elements of operation and algorithmic objectives are important in understanding variability in the performance of different modeling strategies.

A critical finding captured from these results is the utility provided by unsupervised dimensionality reduction. Figure 24 below displays MSE for each of the components of executive function (WM, IC, and SS) using both rsfc matrices and T1 sMRI FreeSurfer metrics. Of the modeling strategies, PCA followed by linear regression achieved performance significantly lower than any other modeling strategy. There was no difference in using N_{comp} retaining ~50% of the variance or $\sqrt{n_{feat}}$ components, thus we used the smaller method of $\sqrt{n_{feat}}$ which was ~10 for T1 and ~250 for rsfc data. Interestingly, even the distribution of permuted performance measures (peach colored box in Figure 24) is significantly lower than the second best actual (olive green) modeling strategy. These results indicate a possible benefit and further exploration of unsupervised methods of dimensionality reduction in high-dimensional spaces (possibly the multidimensional imaging data directly) before targeted prediction.

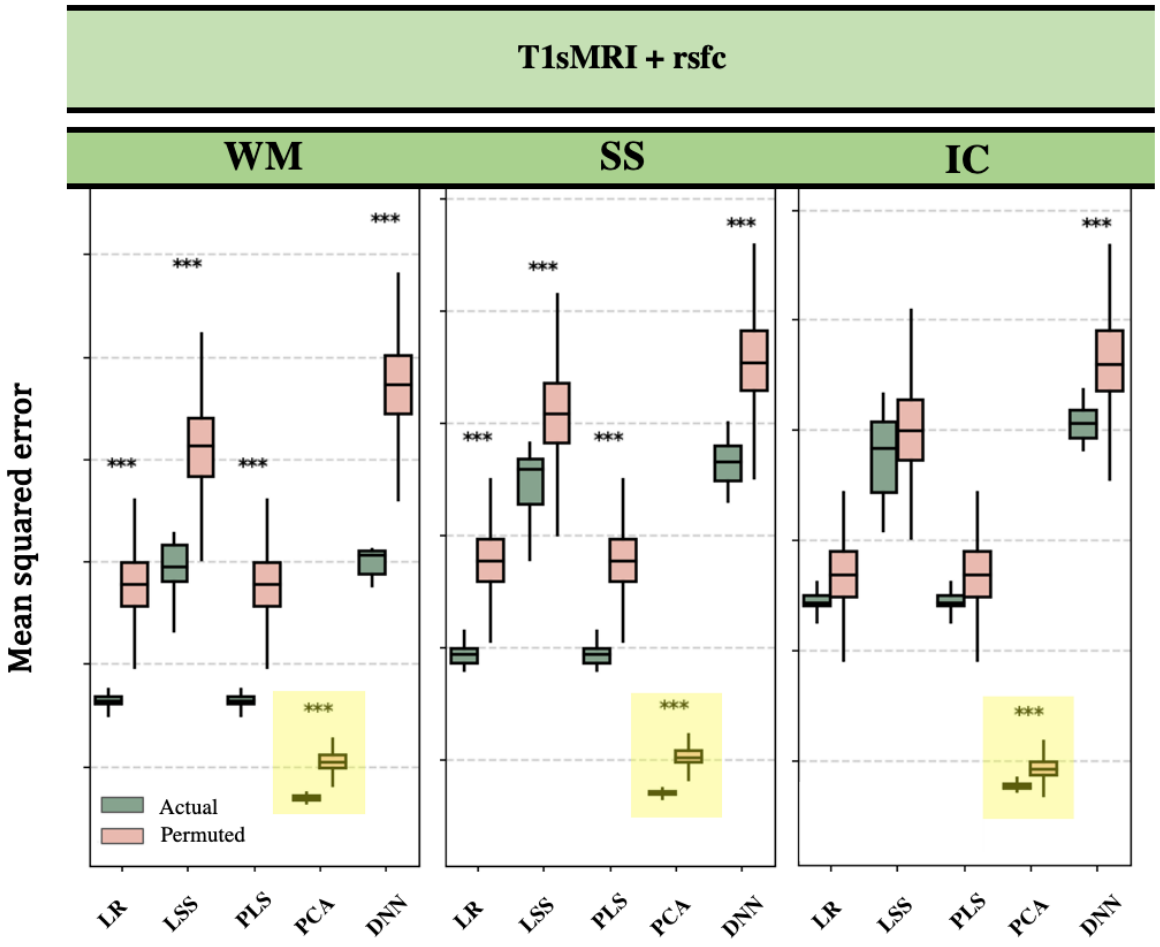


Figure 24: The case for dimensionality reduction

Mean squared error (MSE) of models predicting EF using the multimodal dataset. In these models using the largest set of features, PCA followed by regression achieved significantly lower MSE than any other method. Additionally, the distribution of permuted MSE's (peach) was also lower than any actual MSE from the other modeling strategies.

Psychopathology

The best performing model in predicting the p-factor of psychopathology was Lasso with the T1 data (MSE: .98, *r:* .08). Lasso was not significantly better than other modeling strategies, namely linear regression, but the T1 sMRI data only models significantly performed better than

models using either rsfc matrices or both modalities. Similar to EF prediction results, we did not achieve improved performance from utilizing the various feature selection methods when predicting psychopathology.

To try and further elucidate variability within the prediction of a single measure of psychopathology, we also examined the ability to predict the uniquely presenting internalizing (INT) and externalizing (EXT) disorders. As in the case of the p-factor, Lasso best predicted INT (MSE: 1.11, *r:* .06) and EXT (MSE: .92, *r:* .13) with the T1 sMRI data only.

	MSE	***Model/Modality***	***r***	***Model/Modality***
P	.98 ± .01	Lasso/T1*	.08 ± .01	Lasso*/T1*
INT	1.11 ± .02	Lasso/T1*	.06 ± .02	Lasso*/T1*
EXT	.92 ± .01	Lasso*/T1*	.13 ± .02	Lasso*/T1*

Table 5: Top Psychopathology Prediction

Top predictive performance as mean and standard deviation (MSE and r) for each of the components of executive function. Asterisks after model or modality indicate that modeling strategy or modality was significantly (ANOVA $p < .001$) better than the other modeling or modalities being evaluated.

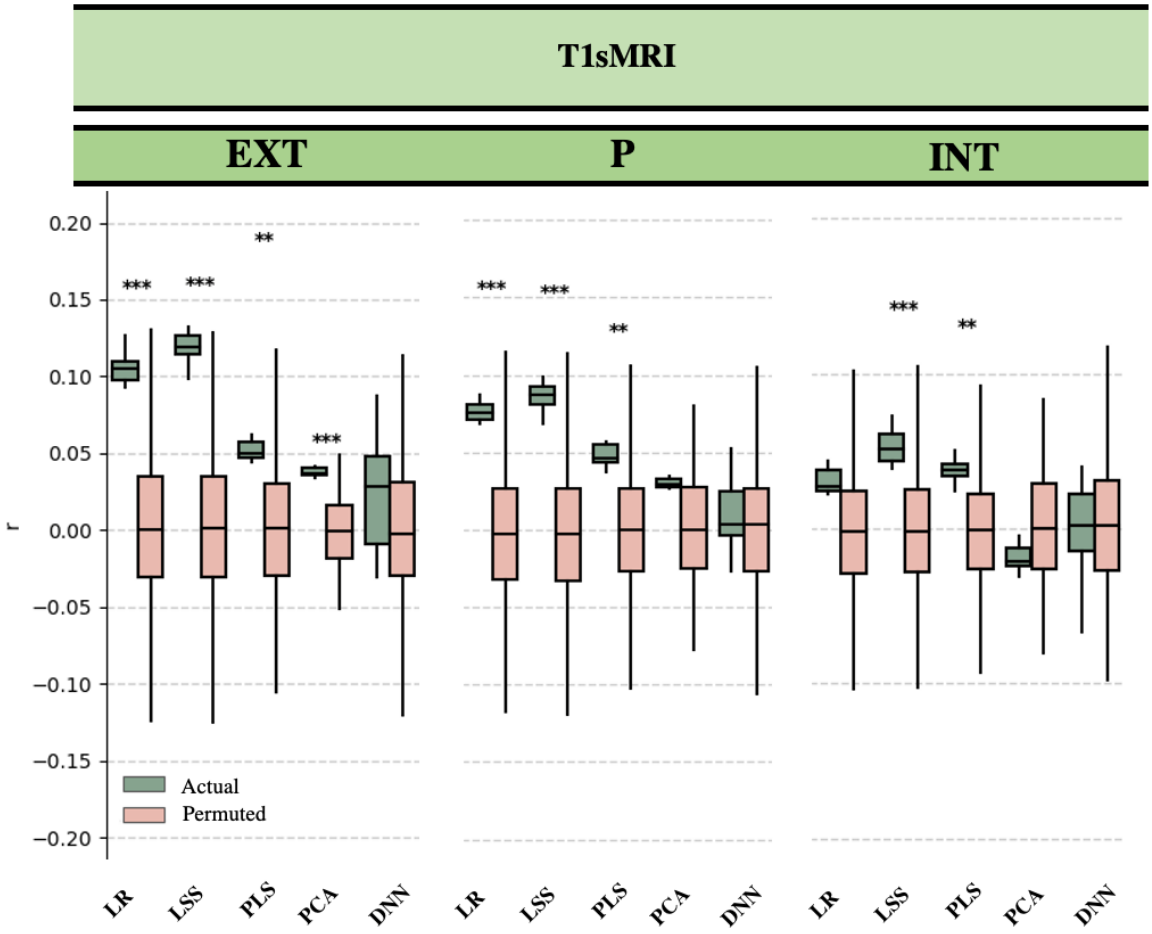


Figure 25: T1 Predicting Psychopathology

As mentioned in the methods section, the targeted outcome (p-factor CFA scores) had a large sample of zeros, that is, individuals endorsing no symptom item from the CBCL checklist. As we discussed previously, this creates issues both logistically (when evaluating traditional linear models) and conceptually, as the likelihood/validity of not a single mental or physical health symptom is low. Figure 26 below displays the performance via the distribution of actual and permuted MSEs, and actual vs predicted regression of both the inclusion of zero endorsement and when removing them from the dataset. Using only the T1 sMRI structural metrics and linear regression, we obtain a test MSE lower than that of the permutation

distribution only in the situation in which we remove the zeroes from the analyses. Furthermore, the same is true for producing a positive non-zero coefficient ($p < .0001$) in the actual vs predicted performance. It is interesting to note that the inclusion of zeros significantly shifts the entire distribution of predicted p-factor down, this is illustrated in the bottom row of Figure 26. While there are arguments for both the inclusion or removal of these individuals, it is important to note results from both situations and the effect this has on modeling and performance.

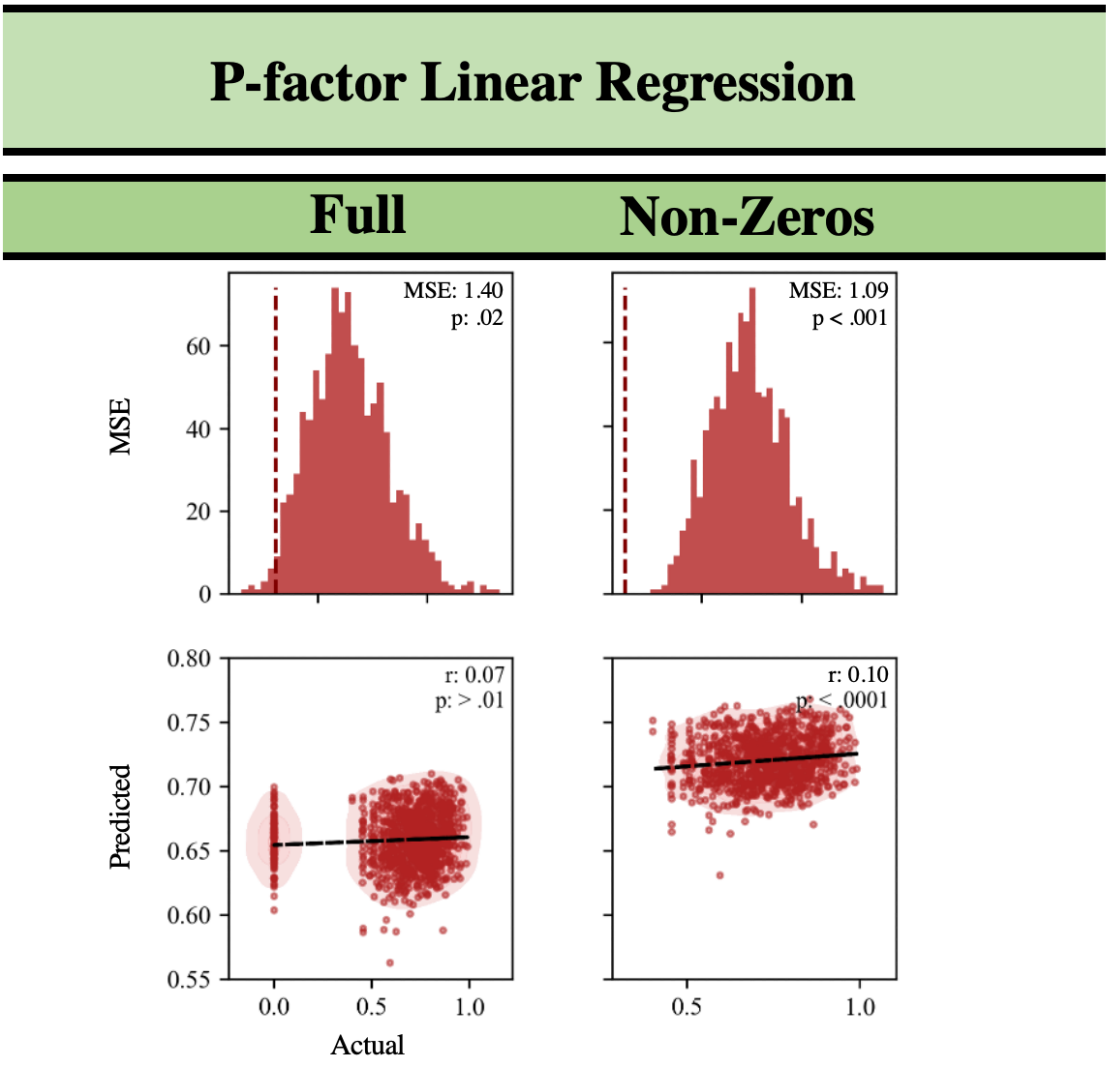


Figure 26: Effect of Non-zero Endorsement

Linear regression of the T1 sMRI data predicting the p-factor of psychopathology and the effect of the inclusion or exclusion of the large sample (~500) of participants endorsing no item from the Child Behavior Checklist (CBCL). Top row: actual test statistic (MSE, red dashed line) and the distribution of 1000 permuted test statistics. Bottom row: predicted vs actual and the regression line (black dashed line).

Multimodal Fusion Imaging Performance

Despite numerous attempts at exhaustive hyperparameter tuning, layer and weight pruning, and evaluating both unimodal and multimodal-fusion networks, the supervised DNN models with 3D T1 sMRI images and 2D rs-fMRI timeseries data never converged, meaning the error never stabilized. Additionally, predictions for each component of executive function (EF), as well as the p-factor were never better than that of random chance, understandable given the non-convergence. We evaluated models ranging from ~50,000 to billions of trainable parameters with no success in either case. The smaller networks never fit the training data, and the larger networks perfectly overfit the training data despite inclusion of methods to combat this including regularization, dropout, and batch-normalization. It was only when using the extracted imaging statistics, FreeSurfer for T1 sMRI and rsfc matrices for rs-fMRI, did we see convergence with the fully-connected DNNs.

Transfer Learning

Unfortunately, this pattern extends to the evaluation of transfer learning with the raw imaging data. Because the individual networks of EF never converged, there was no *knowledge*

to *transfer*. Despite non-convergence, we attempted to use weights from the EF DNNs as the initialized weights in the p-factor networks but saw no improvement in prediction performance. Furthermore, when evaluating the potential use of transfer learning for the DNNs trained using the extracted imaging statistics, we saw no improvements in predictive performance when initializing the weights of the networks predicting psychopathology using the EF networks or when concatenating the multiple EF networks and freezing the early (first two) layers of said networks.

Discussion

Altogether these results underline the performance in predicting measures of EF and psychopathology under various feature selection methods, unimodal and multimodal neuroimaging data, standard machine learning (SML) and deep neural networks (DNNs), and aspects of transfer learning. In this section will we expand upon the interpretation of these results in detail.

This study's performance in predicting executive function (EF) matches or surpasses existing research using neuroimaging data for EF prediction. While primarily evaluating aspects of ICV correction in T1 sMRI, Dhamala et al., (2022) predicted several components of EF and cognition using measures of surface area, GM volume, and cortical thickness in both the ABCD and Human Connectome Project (HCP) studies. They reported Pearson correlation coefficients (r) of ~.20, .05, and .05 for WM, IC, and SS respectively. Similar findings from Chen et al., (2022) used rs-fMRI from the ABCD study to predict multiple measure of cognition (averaged cognition $r = .21$). Additionally, their multimodal analyses comprised of rs-fMRI and multiple task-fMRI scans returned improved performance ($r \sim .29$) when predicting cognition.

Interestingly this multimodal performance closely aligns with our use of T1 sMRI alone. In the same study, the authors also predict aspects of mental health via the Child Behavior Checklist (CBCL) subscales. Using rs-fMRI only resulted in an averaged MHD prediction of $r \sim .05$. Leveraging the same multimodal prediction as noted above, the authors report performance from each of the CBCL subscales comparable to the performance presented in this analysis. Subscales associated with the internalizing (INT) disorders, namely anxious/depressed, somatic complaints, and withdrawn depressed were the most challenging to predict (*r's* ~ .05*)*, supporting our findings that these disorders are immensely difficult to predict using the neuroimaging data. Predicting social and thought problem subscales yielded slightly higher performance (*r's* ~ .10*)* and attention problems the highest performance ($r \sim .17$*)*. Therefore, while our analyses sought to evaluate discrete aspects of DNNs vs SML methods in predictive tasks, the final performance we obtained aligns and in some cases improves upon that of the existing literature.

Performance Evaluation

As mentioned previously in *Chapter 2: Evaluation*, it is critical to contextualize the limitations and assumptions of the performance metrics being evaluated. We found that MSE, possibly the most favored metric of reporting performance of DNN regression tasks, varied as a function of the number of parameters in the model. While it is undisputed that the more parameters in a model the more complicated the loss surface becomes, it is interesting that the mean permutation MSE of the DNNs with T1 data only were significantly lower ($p < .001$) than the actual mean MSE of the models using rs-fMRI or both modalities, despite each DNN distribution of actual MSE being significantly lower than its own respective distribution of permutation MSE (see Figure 27). This phenomenon was seen across every component of EF

and psychopathology. However, the other metric of evaluation, the Pearson Correlation Coefficient (*r*), normalized between -1 and 1, reflects more stability in the permuted test metrics despite the substantially different number of parameters in each model for the different modalities evaluated.

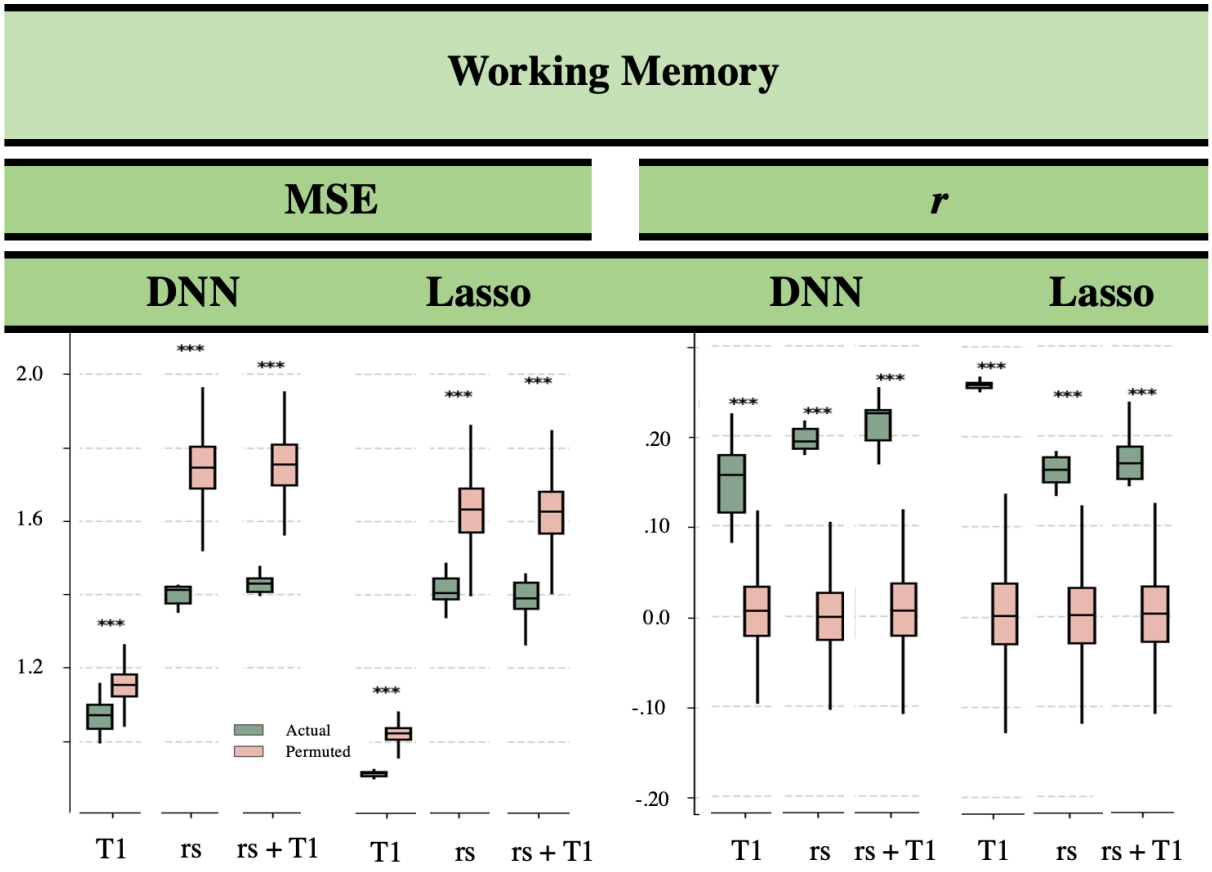


Figure 27: Interpretation of Performance Metrics

In speaking further to the reporting of multiple performance metrics in regression problems, it is important to note that lower MSE may arise from a model prioritizing predictions closer to the mean of the outcome i.e., playing it safe, a situation illustrated clearly in Figure 23. Interestingly, this situation occurred most notably in the DNNs, despite the *sklearn* Lasso

implementation using MSE, the same as our DNN objective function. Furthermore, the most extreme side of this occurrence results in a model in which *only* the mean of the outcome is predicted, a situation that occurred frequently and one we discuss later in detail. There are several possible causes including the generation of the loss surface in addition to the different methods of gradient descent being employed to update the model weights (discussed briefly in the results section). Most likely, is that the greater the number of trainable parameters in the DNN, the more complex the loss surface and task of traversing this surface for weight updating.

$$w_j^{t+1} = argmin_{w_j} f(w_1^t, w_w^2, w_d \ldots, w_d^t)$$

Equation 4: Coordinate Descent

In coordinate gradient descent the weights in the model are learned one at a time while holding all other weights fixed. w_j^t *denotes the parameter j at iteration t.* $argmin_{w_j}$ *is the parameter* w_j *that minimizes the objective function while holding all other parameters fixed.*

$$w^{t+1} = w^t - \eta \nabla f\ (w^t)$$

Equation 5: Gradient Descent

In traditional gradient descent the weights in the model w^t *are iteratively updated at time* (t) *by the learning rate* η *and the gradient of the objective function w.r.t the weights at iteration t,* $\eta \nabla f\ (w^t)$.

Altogether this scenario portrays challenges with reporting model performance and highlights the importance and utility of including multiple performance metrics.

No improvement with DL

Ultimately, we saw no improvements using any of the DNN configurations we evaluated in predicting either components of EF or psychopathology. However, it is critical to note that DL did not outperform SML in the specific context of DNN model architectures and hyperparameter combinations that we evaluated in this analysis. This is not to say that there are not untested combinations of model architectures, hyperparameters, and alternate forms of image preprocessing that would not outperform SML methods. As noted earlier, the infinite configuration possibilities of DNNs, while attractive, makes it impossible to evaluate all, and intractable to evaluate *many*, of the countless architectures, optimizers, and embellishments available to DNNs. We leveraged existing literature to make appropriate generalities and hypotheses to evaluate several different DNN modeling strategies, hyperparameter configurations, and forms of image processing within the scope of this project.

Issues with Input

Perhaps most critical to discuss in relation to the observed model performance are the issues related to the enormously complex data being used as input for each of these analytic strategies. Below we discuss several possible sources of difficulty associated with the use of multi-modal neuroimaging data.

Direct Imaging Models

The primary motivation for using 2- or 3D imaging data relies on the hypothesis that there is signal or information within the data that we are unaware of or unable to manually

extract ourselves. While it is certain that information is lost when reducing this data into coarse summary statistics, such as surface area or thickness for T1 sMRI or a single Pearson Correlation Coefficient to relate parcels within an entire five-minute rs-fMRI run, it may be that the low tSNR and high-dimensionality of the imaging data simply yields too arduous a task. As mentioned in previous sections, while DL has found success in neuroimaging applications, it is most successful when predicting diseases with marked neurodegeneration (e.g., pronounced cortical thinning), such as Alzheimer's Disease and Schizophrenia or in characterizing discrete demographic or developmental changes, such as age and sex at-birth prediction.

In the 2019 Neurocognitive Prediction Challenge (ABCD-NP-Challenge), researchers at SRI International tasked individuals to predict fluid intelligence using T1 sMRI data from the ABCD study. While many groups evaluated various forms of DNNs, the top performing models were comprised of SML methods, including various forms of kernel-based and penalized regression models (Mihalik et al., 2019). This is not to say that other DL architectures would never outperform SML models, but this was not the case in both this highly publicized prediction challenge and within our analyses predicting measures of EF and psychopathology.

Despite having a large sample (+6000 study participants) for neuroimaging standards, researchers posit the incredibly high-dimensionality and low tSNR of neuroimaging data may require tens of thousands of observations to predict complex outcomes such as cognition, EF, and psychopathology. Although some evidence supports DL methods significantly outperforming SML with as few as 1000 participants, the majority of these tasks are again in the realm of age and sex prediction or predicting diseases with low heterogeneity and more marked patterns of neurodegeneration or neuroanatomical patterns. Moreover, top performing computer vision models for object class prediction, such as the groundbreaking AlexNet (Krizhevsky et al.,

n.d.), are trained using over 14-million examples. And while, the images within the ImageNet dataset are also 3D (third channel being RGB), each image in this task contains a total of ~150k pixels, a far cry from the over 7-million pixels in the T1 sMRI data. Ultimately, the underwhelming performance of the models leveraging the multidimensional neuroimaging data, while disappointing, is not entirely unfounded.

On Preprocessing and Noise

Additionally, a particularly difficult factor to consider includes the amount of pre-processing that occurs within each imaging modality going into the DNNs. There is currently minimal consensus on *what* and *how-much*, if any, preprocessing should be performed. For example, the translational invariant nature of feature extraction in DNNs, i.e., the capability of convolutions to extract patterns in different spatial locations within the image, could theoretically allow for less rigorous, i.e., fewer degree of freedom (DOF), registration. If too much registration (non-linear transformations) is applied to the T1 sMRI, we may remove the nuanced morphological aspects we are attempting to evaluate. However, our early attempts replicating sex-prediction performance revealed that linear (affine, 6-DOF) registration was the minimum registration requirement for prediction better than chance. Conceptually this makes sense when considering how convolutions are applied to the image. Traditional CNNs are not rotationally or shear invariant, thus requiring a pre-model registration method that accounts for these orientational elements within the images. There are, as always, additional modifications to traditional CNNs that seek to tackle this specific issue, capsule networks, data augmentation, such as introducing random rotation or shear to images. However, additional evaluation and

added complexity increases both computational demands and training time, making the evaluation of all of these additional methods intractable.

The crux of the previous argument remains the same for almost every other pre-processing procedure, including noise-reduction, artifact removal, frame censoring, and nuisance regression Each of these processes involves a series of decisions, hyperparameter selection, motion cutoff, etc., all of which have been thoroughly evaluated in traditional statistical analyses. However, current research is sparse and at times contradictory as to how preprocessing the neuroimaging data changes prediction outcomes in DL applications. For example, some argue for the use of "minimally-processed" T1 sMRI data for brain age prediction (Dartora et al., 2023), the task of predicting age of an individual using only neuroimaging data, while others argue that "extensive preprocessing" improves model generalizability within the exact same application (Dular et al., 2023b).

Batch Effects and Correction

The point of removing nuisance variables such as motion, extends further into known sources of *batch* within the neuroimaging data within the ABCD study, most notably the effect of scanner manufacturer. Initial models replicating sex at birth prediction using ComBat corrected imaging data, found no improvement in downstream model performance, a finding not uncommon in neuroimaging based disease modeling (Kushol et al., 2023). While it is critical to acknowledge and understand that batch effects of scanner are present in this sample, arguments can and have been made both for (Eshaghzadeh Torbati et al., 2021) and against (Nygaard et al., 2016; Zindler et al., 2020) the use of batch effect correction. Additional research from Dufumier et al 2024, discusses harmonization techniques (including ComBat) in DNNs and how these

methods may "*fail to preserve possible non-linear relationships leveraged by DL*". Ultimately, when analyzing the potential for variable performance across scanners, we found no significant disparities in performance by scanner (see Figure 29, left), and thus did not include the added computational complexity of correcting for scanner manufacturer.

$$Y_{ijp} = \mu_p + X_{ij}\beta_f + \gamma_{ip} + \delta_{if}\epsilon_{ijp}$$

Equation 6: ComBat

The ComBat method of batch effect correction in the predicted features $X_{ij}\beta_p$ are adjusted by applying both additive and multiplicative scaling $(\delta_{ip} + \gamma_{ip})$ to account for sources of batch within a dataset X_{ij}.

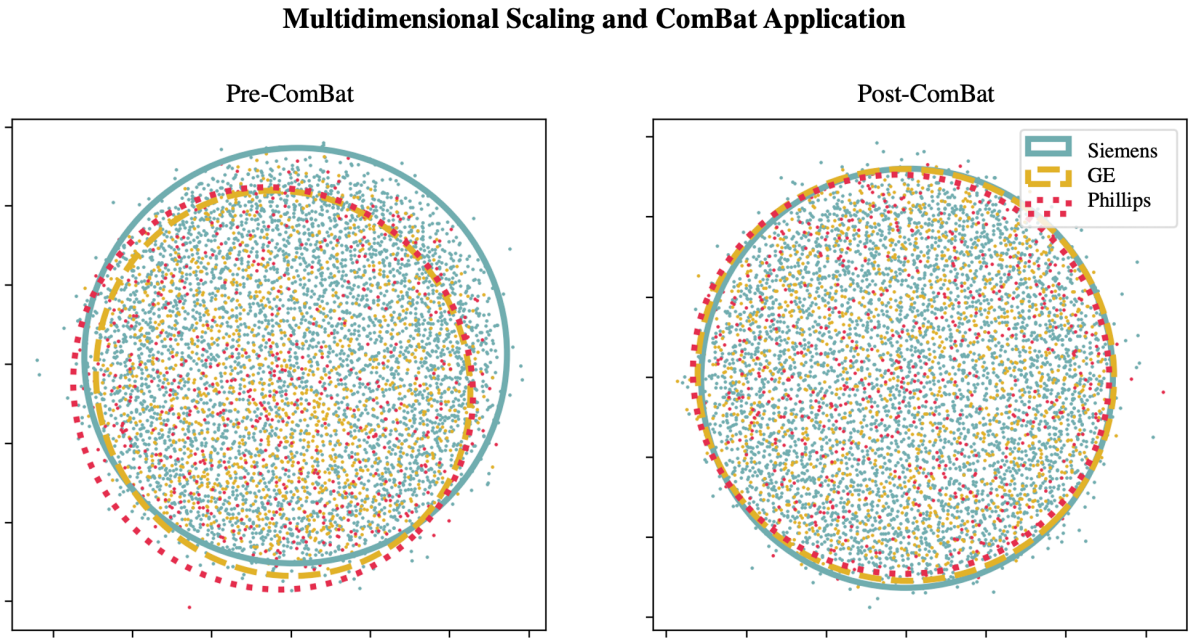


Figure 28: rsfc Scanner and ComBat

Effect of multidimensional scaling (MDS) on the resting state functional connectivity matrices both pre- and post-ComBat application. Aligning with the literature there are notable effects of scanner manufacturer present within the rs-fMRI data.

Altogether, results obtained from the models leveraging neuroimaging data as direct input failed to produce results better than random chance. As noted above, there are numerous possibilities for this outcome, and we contend that additional research is critical to fully understand the challenges associated with high-dimensional low-tSNR imaging data in DL. Subsequent discussion focuses on the evaluation of models using features obtained from additional processing, including the FreeSurfer T1 pipeline for structural features, and the generation of resting state functional connectivity matrices for functional data.

Feature Selection

While the manually created features substantially reduced the dimensionality of the data being used for each model, we also sought to evaluate potential benefits, through reduced model complexity and training time, of various feature selection methods. However, feature selection methods did not yield improvements in model performance. There are several reasons this may have occurred. The variance filter, the method producing the most favorable results of the evaluated feature selection methods, only considers variance of each feature independently, that is in a univariate sense. This is in contrast to the higher performing PCA, which also reduces the number of features, but can robustly capture nuanced variability within the underlying structure of the data by transforming the original features into new, uncorrelated, components. PCA identifies these components using linear combinations of all features to maximize these latent representations of variability within the data. As noted, these latent representations are also themselves uncorrelated, thus removing sources of multicollinearity among features in the data that would not be resolved by a variance filter alone.

Additionally, the meta-analytic feature selection method, while well intended, also has several limitations. Most notably, the neuroimaging modality used in the meta-analytic platforms. These platforms exclusively use results from literature of task-based fMRI studies. In task-fMRI, individuals complete specific activities within the scanner to evoke explicit responses tied to said task. While rs-fMRI does capture brain function, it does so at rest, or in the absence of a task, thus, it may stand that functional processes present differently at rest from signatures evoked during the performance of a task. Furthermore, in relation to T1 sMRI, which was also used to prioritize specific brain regions, the relationship between structure and function within the brain is poorly understood. Some studies have identified underlying relationships between structural connectivity, via DWI, and functional connectivity (Babaeeghazvini et al., 2021; Bennett & Rypma, 2013; Honey et al., 2010), but this modality does not have the same high spatial resolution, and thus, subsequent morphological representation, as the T1 sMRI. This, however, is an inevitable limitation of all current meta-analytic platforms.

In summary, while we maintain the likelihood that signal exists within structural and functional neuroimaging data, it is critical to consider the complexity of the data being used as input for these predictive applications using high-dimensional data to predict complex heterogeneous outcomes. The complexity of this task may require an amount of data that is currently unavailable to researchers in the realm of neuroimaging. Additionally, there are numerous remaining elements within neuroimaging that warrant further research under the context of DL application.

Issues with Outcome

As mentioned previously, the targeted mental health outcomes in this project are inherently difficult to predict using the structural and/or functional neuroimaging measures. While our obtained performance is better than, or equivalent to findings within the existing literature, it is important to highlight exactly what makes this task of modeling components of EF and psychopathology so difficult.

Covariates and Corrections

One of the first steps in the creation of these models is the decision to include covariates. An open question in the field, the decision to correct for aspects of gender, sex-assigned at birth, race, ethnicity, socioeconomic status (SES), and income can further marginalize already vulnerable groups if not examined within the context of larger systems of inequality (Saragosa-Harris et al., 2022). Furthermore, researchers using the Alzheimer's Disease Neuroimaging Initiative (ADNI) data found no differences in model accuracy when correcting or covarying for age or gender in the prediction of AD using multiple modalities of neuroimaging data (Rao et al., 2017). However, we found substantial variability in predictions by sex in some models, but not all. Interestingly, during certain instances of model training the models strongly relied on the aspect of sex as a function of outcome. This effect is illustrated clearly in Figure 29, right. The inclusion of sex in the later layers of the DNNs, and as a feature in the SML models reduces this potential for substantial effects of sex by outcome.

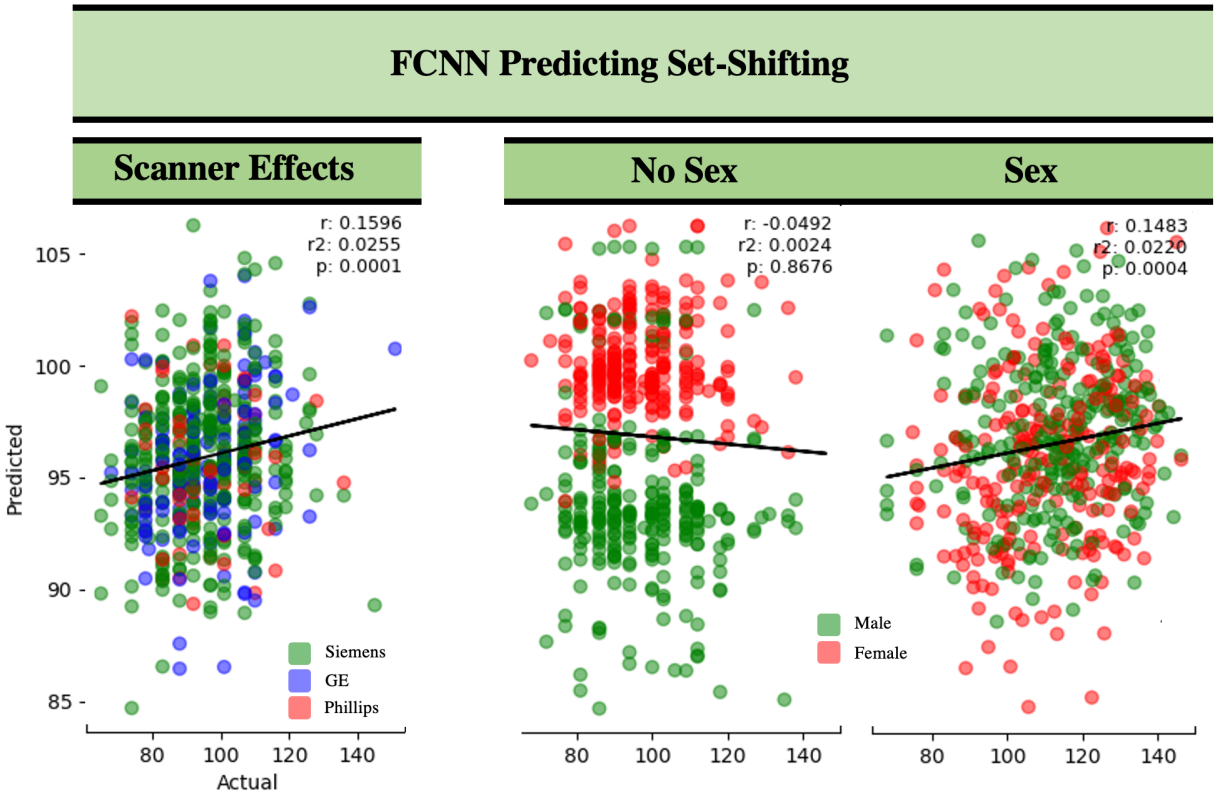


Figure 29: Set-Shifting and Covariates

Predicted vs actual values of set-shifting (SS) using the deep neural networks. Left displays potential effects of scanner indicated by the different colors of observations, we found no effect of scanner in this evaluation (ANOVA, $p > .01$). On the right, however, we can see an instance of model training in which we do not include the binary variable of sex-assigned at birth during model training. Because there are effects of sex within the SS variable, in certain cases of model training, the neural network heavily relied on this single variable which created a large effect of sex present in the predictions and resulted in a model that performed worse than random chance. The right-most plot shows the same model trained including the sex-at-birth variable, removing the highly significant sex effect in the predictions and results in significantly better model performance.

Yet another method some use is "residualization" or the possibility to model and remove unwanted signal from an outcome and predict not the original outcome, but the residual from the linear model after this initial regression. In the previously mentioned *ABCD T1 Cognition Prediction Challenge*, researchers were predicting not raw fluid intelligence, but a residual of this metric after removing effects of total brain volume, data collection site, age at baseline, sex at birth, race/ethnicity, highest parental education, parental income, and parental marital status. This instance of residualization was inherently stringent and left some researchers arguing that the use of a residualized outcome may remove sources of actual biological signal related to the outcome (Oxtoby et al., 2019). Furthermore, researchers in psycholinguistics thoroughly analyze and discuss why residualizing does "not create an improved, purified, or corrected version of the original predictor" (Wurm & Fisicaro, 2014). This is all to say that while there are certainly unmodeled demographic effects within our outcomes, it is critical to ensure that analyses do not further marginalize individuals, and that performance under constraint and consideration of specific covariates are examined with and without methods of correction. This aligns with recommendations from emergent literature discussing that individuals including substantial covariates include results both with and without correction (Hyatt et al., 2020).

Subthreshold disease and Normative Modeling

Despite the provided benefits of evaluating psychopathology via the p-factor, which counters *some* issues related to heterogeneity and comorbidity present in binary mental health disease prediction, there remain inherent challenges in this strategy as well. This methodology makes the inherent assumption that there is a linear relationship between aspects of either

structural or functional neuroimaging data and cognition/psychopathology. However, this aspect is difficult to verify and poorly validated within the field (Greene et al., 2022). Some researchers posit that potential brain-behavior relationships could vary non-linearly across either aspects of neurobiology or psychopathology (Faghiri et al., 2019). Evidence suggests the utility of examining *abnormal* variation (deviations) from normative neurodevelopmental trajectories of individuals over time in relation to psychopathology (Parkes et al., 2021). This framework, dubbed *normative modeling*, while promising, adds substantial complexity, often requiring multi-timepoint longitudinal data and relies on the confidence of the established trajectories. However, researchers have discovered associations between negative deviations from normative cortical thickness and higher general psychopathology in 21 year-olds (Kjelkenes et al., 2022). Further along this line, some believe that the tendency of most models to identify robust relationships centered around the mean of a distribution, fails to capture the most *interesting* relationships between highly atypical brain-behavior relationships. Proposed *multivariate extreme value statistics,* in combination with normative modeling highlight robust relationships between extreme brain deviations and behavior within the UK Biobank sample (Fraza et al., 2022). While normative modeling provides an interesting aspect of disease modeling, it is crucial in the context of these analyses, specifically the evaluation of the utility deep learning, to consider and reduce additional factors that may add to the complexity and scope in which we are examining.

Issues Specific to Deep Learning

Apart from difficulties encountered with model input and outcomes, there were specific problems inherent to deep learning that warrant discussion. The high complexity and size of

some of the DNNs evaluated led to issues of non- or inconsistent convergence, substantial training time, and enormously high demand of computational resources, all of which are sources that make the evaluation and analyses of different architectures and models distinctly demanding.

Computational Demands

While recent computational advances, namely advances in GPU hardware, have sped up model training time in the field of computer vision, large-scale problems, such as the immense size of the neuroimaging data, still take a considerable amount of both time and computational resources. The multi-modal imaging models we evaluated, even after reducing the number of hidden layers from eight or more to only four and reducing the number of neurons in each fully connected layer, took an exorbitant amount of time to train. Furthermore, while initial attempts downsampling the imaging data did speed up model training time substantially, these models also never converged. Some of the larger networks took several hours for a single epoch (one full pass through the training data) of training. While a notable limitation of DL, it is precisely this property that makes it so difficult to train and evaluate many different types of architectures and hyperparameters within the neural network. As mentioned previously, we made hypotheses and leaned on existing literature to evaluate several potentially successful architectures.

Models were trained using a single NVIDIA RTX 3060. This GPU is not part of the newer released GPUs from NVIDIA, and we would have likely seen faster training times with the latest hardware. It is also worth mentioning that the large size of the neuroimaging data requires that models be trained using exceedingly small batch sizes (5-10 observations per batch) due to the limited memory of this and other similar GPUs. Smaller batches are theorized to

create “noisier” gradients and while some research highlights the utility of deliberately adding noise to the gradient to improve model performance (Neelakantan et al., 2015), this is typically in cases of overfitting. However, in the context of neuroimaging data, where high dimensionality and complex structures are prevalent, the balance between batch size and gradient stability becomes a critical factor in model optimization. Thus, while it is possible that larger batches may yield more stable gradients and be more conducive to steady convergence, the computational and memory constraints imposed by large neuroimaging data limit the available options.

Instability of the DNNs

Instability, and expressly model complexity, i.e., the number of parameters and layers in a model, can inhibit model convergence or the likelihood of a model to overfit the training data. This scenario is very much what we saw in these analyses. Specifically, the margin between overfitting and underfitting was razor thin, creating models that either perfectly learned the training data ($r \sim .99$) or learned nothing ($r \sim 0.00$).

One aspect of training multiple instances of DNNs is that of the stability of convergence, even when using the same parameters for different training sets. This point is illustrated below in Figure 30 on the left, which shows actual vs predicted values in the test set for set-shifting using the neural network below. Despite exhaustive hyperparameter tuning there were numerous instances during training different folds of data in which the models simply never converged and only predicted the mean of the outcome, despite assuring there were no significant differences of outcome distributions of each fold. This, in turn, brings down the average relationship between actual and predicted values as seen by the averaged prediction line (bold black line).

In addition to unstable convergence, in situations in which all models do converge, we see inherently higher uncertainty, indicated by higher prediction variance over the 10-folds the

further we get from the mean of the outcome. While this is not uncommon among large samples heavily centered around the mean, it is worth noting, and a phenomenon which is likely the culprit of the limited range of predictions with the DNNs as seen in Figure 23 in the results.

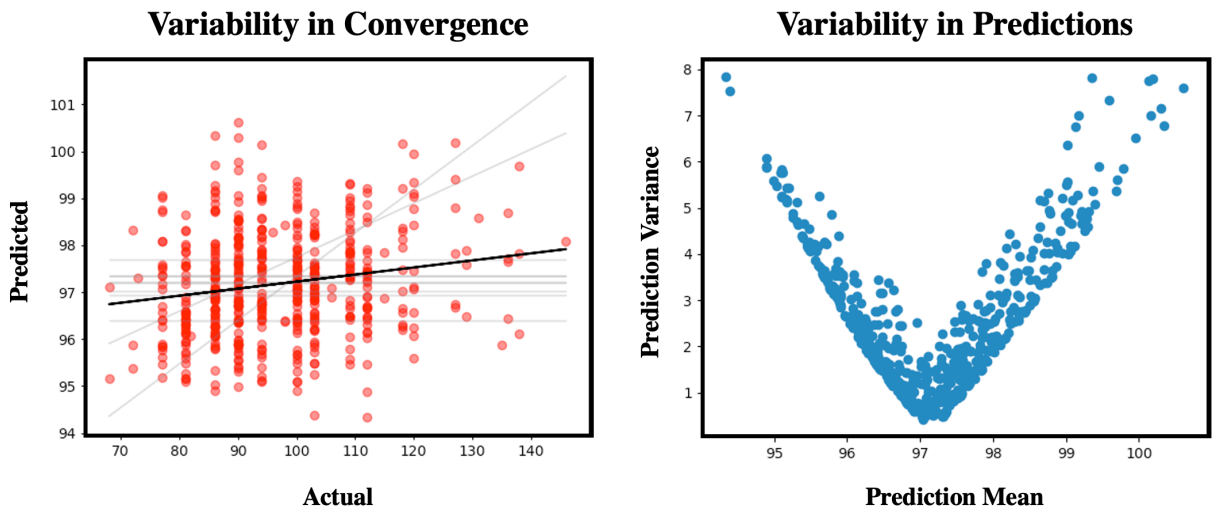


Figure 30: Challenges in Prediction

Left: actual vs predicted values of set-shifting in the test set. Light grey lines indicate regression for a given fold of training and the darker black line represents the regression line fit for predictions over all folds. Right: Prediction variance vs the predicted mean in the test set from the 10 training folds, highlighting uncertainty in predictions deviating far from the mean.

Transfer Learning

Finally, it is important to discuss possible reasons as to why we were unsuccessful in attempts of evaluating strategies of transfer learning. While transfer learning has the potential to yield improved task performance through the *transfer* of knowledge learned in one task to another, it inherently relies on the *information* learned from the model's initial learned task. As

mentioned, the imaging models never converged, and while the extracted feature models obtained competitive results, the transfer of information from models predicting EF did not improve our ability to predict psychopathology over the typical *random* initialization of DNN model weights. While we previously discussed the robust body of literature implicating EF dysfunction in nearly all forms of psychopathology, our initial evaluation of the relationship between EF and psychopathology in the ABCD study, while significant, was weakly correlated. It may stand that, the underlying relationships and information that predict EF are either too weak or dissimilar in their utility for establishing neuroimaging correlates of psychopathology.

Conclusion and Progression

Altogether, these analyses highlight the complexity involved in predicting aspects of executive function and psychopathology via multi-modal neuroimaging modalities under various modeling conditions. While deep learning has shown promise in the realm of neuroimaging, we did not see improvements when using this methodology with our limited selection of architectures, parameters, and tasks. All together we hypothesize several areas, detailed below, in which we can reduce the complexity of the task at hand and expand further upon the successful elements of these analyses.

Narrowing of the Input

Despite a body of literature highlighting the use of rs-fMRI for predicting elements of both cognition and psychopathology, to reduce model complexity and run-time, subsequent analyses will focus solely on the use of T1 sMRI data. The high spatial resolution is well suited for capturing aspects of morphology, whereas the high dimensionality and considerably low

tSNR of rs-fMRI makes it difficult to unearth predictive signal. Furthermore, results obtained from Chen et al., (2022) predicting cognition and psychopathology in the ABCD sample using rs-fMRI data were lower or equivalent to our results using only the T1 sMRI data. Additionally, the performance we obtained in predicting working-memory (highest performing of the EF components) and all components of psychopathology was either higher or not significantly different in models using only the T1 sMRI data. While there are potential benefits for leveraging multi-modal neuroimaging data, the added complexity, training time, and additional processing required to do so comes at a cost, and we will move forward with the T1 modality alone.

Reduction of Outcomes

In addition to the reduction of neuroimaging data being used in these models, we will also be reducing the components of EF predicted and constraining the range of psychopathology being predicted. While we saw promising performance in predicting components of set-shifting and inhibitory control, the performance was lower than that of predicting working-memory. More importantly, homing in on a single component of EF reduces the number of models being trained by two-thirds, and given the significant training time required, this aspect is critical. Additionally, the goal of predicting components of EF is to use the knowledge learned by these models to improve our ability to predict psychopathology. However, our evaluation of this approach did not yield improved model performance.

In addition to reducing the number of components of EF, we also seek to moderate the scope of psychopathology prediction by moving from a single general factor of psychopathology, the p-factor, to a narrower dimension of externalizing psychopathology.

Performance predicting externalizing disorders was significantly better than either the p-factor or internalizing disorders. Furthermore, internalizing disorders proved the most difficult to predict, and as the p-factor encompasses both internalizing and externalizing psychopathology it is likely the inclusion of these internalizing disorders leads to lower performance of p-factor prediction. This finding is not isolated to our analyses, several studies using the ABCD sample have found widely robust brain-based predictors for externalizing disorders. Most notably, the recent work from Xu et al., (2024) used Canonical Correlation Analysis with the rs-fMRI data from ABCD to identify latent profiles of psychopathology. The two primary brain-based symptom dimensions, replicated in an additional study sample, were that of attention problems and rule-breaking and aggression. Furthermore, in bolstering the argument to specifically focus on the relationship between the EF component of WM, the top resting state networks loading onto these symptom dimensions were from the salience, parietal occipital, and medial parietal networks, all of which are associated with top-down aspects of control and include attention and spatial working memory (Cai et al., 2019). Thus, while we are reducing the leverage of using all components of EF, this reduction in context with the sheer number and complexity of the models being trained, allows for a deeper dive into other strategies and architectures to evaluate the potential benefits of DL.

Revision of Modeling

The final aspect to discuss is how we will be moving forward with our modeling strategies and evaluation of deep learning. While the finer aspects of this evolved strategy will be discussed in detail in the following chapter, we seek to improve our ability to leverage DL strategies by focusing on the realm of dimensionality reduction. It is possible that the task of supervised DL

models with the high dimensional and noisy neuroimaging data is simply too complex. Our primary hypotheses for the utility of deep learning in neuroimaging data are that it can,

1) Unearth information from the input we cannot provide via manual feature extraction.
2) Model complex nested relationships we are unaware of.

Thus, we seek to evaluate the use of DL architectures that leverage convolutions not for the prediction of the outcome directly, but for the extraction of deeply embedded features tied to morphology within the structural MRI data.

Chapter 3: The Case for Unsupervised Dimensionality Reduction

Motivation

In this chapter we will be discussing and analyzing results from models using unsupervised dimensionality reduction techniques from both standard machine learning (SML) methods, such as principal components analysis (PCA) and deep neural networks (DNNs) such as the autoencoder (AE) and variational autoencoder (VAE). The primary hypothesis being that low dimensional representations of the structural data, captured by these methodologies, may yield further insight and potential utility for brain-disease modeling than the previous unsuccessful supervised DNN's in the Chapter 2.

Materials and Methods

The clinical and imaging data being used are also from the ABCD study, however, because we are not using the rs-fMRI data, we are no longer required to exclude the large sample of participants not having usable rs-fMRI data (largely due to excess head motion). This allows for the inclusion of an additional ~3800 participants, bringing the total number of participants used in this analysis to 9754, which greatly boosts the power of these large-scale models.

Subject Demographics

	N	**%**
N	9754	
Age	9.9	*0.6 (SD)*
	Sex	
Male	5095	*52.2*
Female	4657	*47.7*

	Race/Ethnicity	
Asian	192	*2.0*
Black	1333	*13.7*
Hispanic	1866	*19.1*
White	5333	*54.6*
Other	1030	*10.6*

	Parents Married	
Yes	6838	*70.1*
No	2916	*29.9*

	Parent Highest Edu.	
< Highschool	488	*5.0*
Highschool/GED	927	*9.5*
Some College	1239	*12.7*
Bachelors	3755	*38.5*
Graduate	3316	*34.0*

	Household Income	
Income <= 50k	3043	*31.2*
50k < Income < 100k	2653	*27.2*
Income >= 100k	3960	*40.6*

	MRI Manufacturer	
Siemens	6370	*65.3*
GE	2230	*22.9*
Phillips	1154	*11.8*

Table 6: Participant Demographics Chapter 3

Additionally, we slightly altered the partitioning schema used in this analysis. While we continue to utilize the train, validation, and global test splits, the generation of train and validation are obtained via 10-folds. Whereas the previous strategy used random sampling to obtain train and validation sets. This allows for the generation of entirely distinct “test” sets

during tuning of the DNNs. Ultimately, this is unlikely to make a large impact given the size of the dataset being used, but more thoroughly ensures that every subject is in the test set at least once.

Image Processing

The image processing pipeline used in these analyses is the same as described above in Chapter 2: Image Acquisition). T1 sMRI data are linearly aligned to the MNI 152 standard template and segmented using the Desikan-Killiany atlas and the derivatives were again obtained from the ABCD-BIDS Community Collection (ABCC; NDA Collection 3165).

Due to the immense memory and computational requirements of both standard machine learning (SML) methods, such as PCA and the deep neural networks (DNNs), we again must evaluate how substantially we can reduce the resolution of the image while also retaining the potential predictive utility of the imaging data. Each 3D image can be scaled in each dimension by a factor of z to cubically reduce the size of the image (see Figure 31). Applying an 80% (.8x) reduction to each dimension can cut the size of the image by 50% ($.8^3$), without visibly obscuring the anatomical boundaries for gray matter, white matter, and other subcortical structures. This reduction in image size greatly reduces the time and computational resources required to train these large models (Abrol et al., 2021; Emmert-Streib et al., 2020; Kuang & He, 2014).

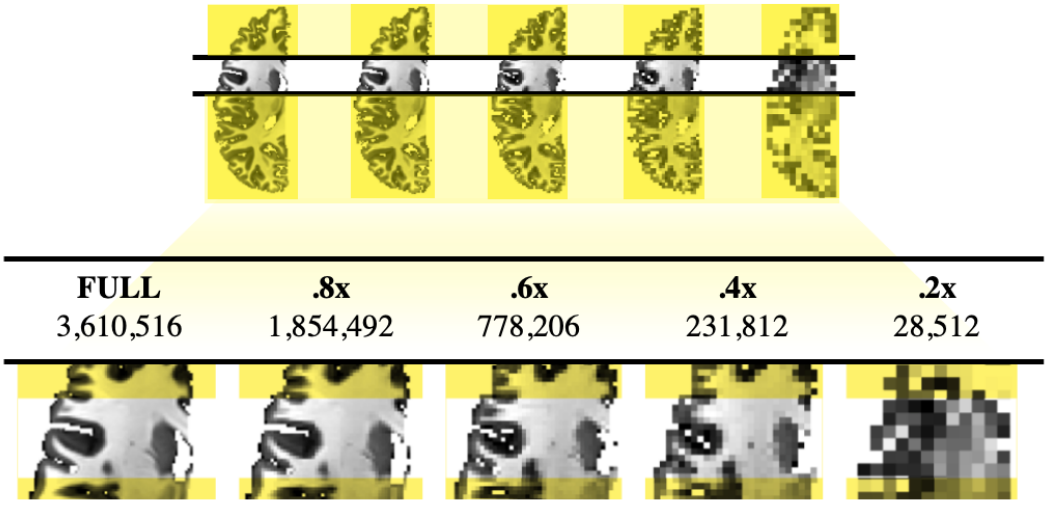


Figure 31: Downsampling

Most of those using sMRI with DL apply either patch-based methods, training using small 2- or 3-D patches of the input, or downsampling to make the models more computationally manageable.

Minimum Bounding Cube

While there exists substantial variability in structural morphology, some subjects that pass quality control (QC) still have segmented data that exists markedly outside of a "reasonable" area of the given anatomical boundary, this is illustrated clearly Figure 32, with heatmaps corresponding to the percentage of subjects with segmentation data over each given structure (cortical, white matter, or nucleus accumbens). Each additional row applies a percentage threshold to only show regions which have at least that number of subjects with corresponding segmentation data. Perhaps most egregiously, the nucleus accumbens specifically (first row on the right), illustrates how segmentation outliers within the accumbens erroneously extend all the way through the corpus collosum. If the group mask is constrained to include boundaries existing in at least 20 (row 2) or 50 (row 3) subjects, we preserve authentic

variability, while greatly reducing inaccurate segmentations from a few potential outliers that have data existing substantially outside of a given region due to potential segmentation issues. Thus, we chose to apply a threshold of 50 subjects to all segmentation data, and generate a minimum bounding cube, similar to that of Li et al., (2022). This method crops the original image to an area that only encompasses each corresponding structure over all given subjects, ensuring that we include 99.5% of subjects with labeled segmentation data while greatly reducing the size of the input for each model.

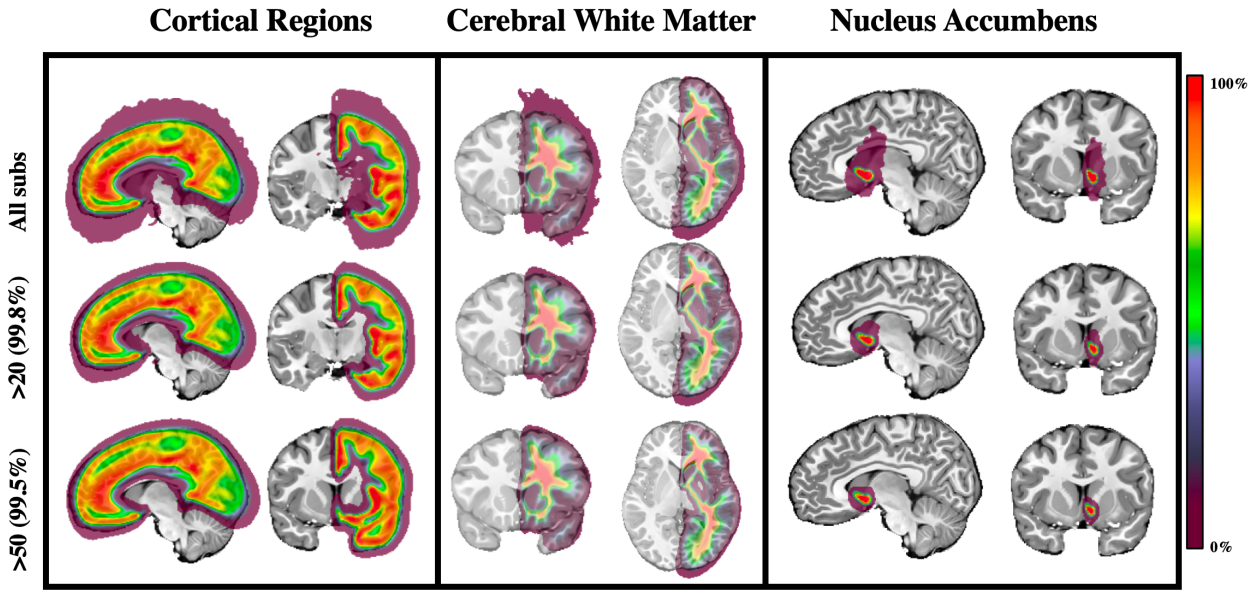


Figure 32: Minimum Bounding Cube

Dimensionality Reduction:

Principal Components Analysis

In these analyses, we will be using PCA to identify latent representations of the 3D structures. In these models, the segmented 3D imaging structures are flattened into a 1D vector

before minibatch PCA (see Figure 33) computes the projections. Unlike traditional PCA, this method allows for the handling of very large datasets that would not be able to fit into memory in a single PCA model. Data are processed in batches of a specified size, where an initial projection is estimated, and the estimations are updated iteratively using batches of the data. The algorithm, implemented using *sklearn*, continues iterating through the data until either the model stops improving over a specified number of iterations or it reaches a maximum number of iterations. Because there is inherent variability in the data being used in each batch, there is an element of stochasticity and possible reduced accuracy over traditional PCA.

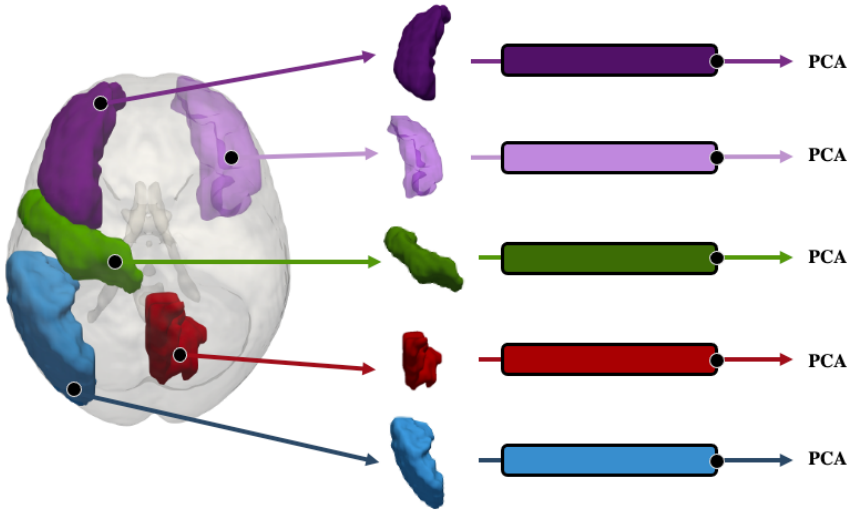


Figure 33: Multiple 1D PCA Models

The Autoencoder

Due to the previous success of dimensionality reduction (PCA) prior to prediction, we seek to evaluate a DL method for unsupervised dimensionality reduction, namely with the autoencoder. In this architecture (see Figure 34), the input is compressed through a series of layers within the encoder into a, typically, much smaller dimensionality described as the

bottleneck, or *latent* layer, and finally reconstructed through layers in the decoder. The encoder and decoder can be constructed using any configuration of number of layers, number of neurons, activation functions, and operations such as convolutions. It is important to note that the primary objective of this architecture is to reconstruct the original input (in this example a 3D T1 image) with the lowest error, typically mean squared error (MSE). In the vanilla example, there are no constraints on the bottleneck layer such as uncorrelated features as in the case with PCA; these are embellishments that can be added in various ways. Furthermore, to reduce the complexity of having models for each hemisphere, the right hemisphere is flipped to match the orientation of the left, and we include an additional binary variable that is attached to the latent layer indicating 0 (left) and 1 (right), in addition to including the previously established covariates of sex assigned at birth and scanner manufacturer (see Covariates and Corrections in Chapter 2).

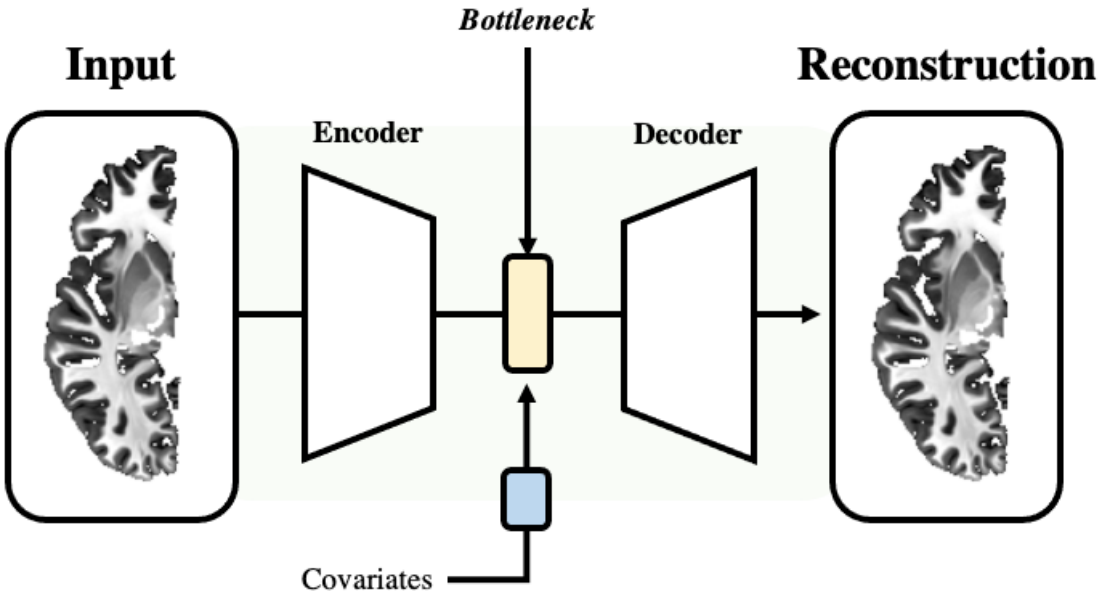


Figure 34: The Autoencoder

The specific architecture for this model is specified in Figure 35. While the input size of each brain structure varies, the structure of the autoencoder remains the same for simplicity and computational limitations. Reconstruction loss is measured using MSE, with the ADAM optimizer, an adaptive learning rate, and weight decay. The Encoder is comprised of three layers each containing 3D convolutions 3x3x3, 3D batch-normalization, the ELU activation function, and finally 3D MaxPooling with index storage for later UnPooling. In the bottleneck layer, the activations from the encoder are flattened and fed into three linear layers including a fully connected dense layer, batch-normalization, ELU activation function, and dropout. Covariates of sex, scanner, and corresponding structure hemisphere are included prior to the middle linear layer. The most compressed representation of the data is in this second layer in the bottleneck layer, and through tuning, contains only twenty neurons. The number of bottleneck neurons was tuned such that we could compress the imaging data into the smallest size without significantly hindering the reconstruction error. Additionally, this optimal number of latent neurons was used to select the number of components for PCA to match the reduced representation and subsequent performance comparison. These activations are stored and utilized later as components similar to that of PCA in further supervised analyses. The decoder is structured the same as the encoder but reversed. The data are reshaped from the last linear layer into the bottleneck layer, which includes an UnPooling layer (similar to upsampling), 3D transpositional convolution, and finally 3D batch-normalization.

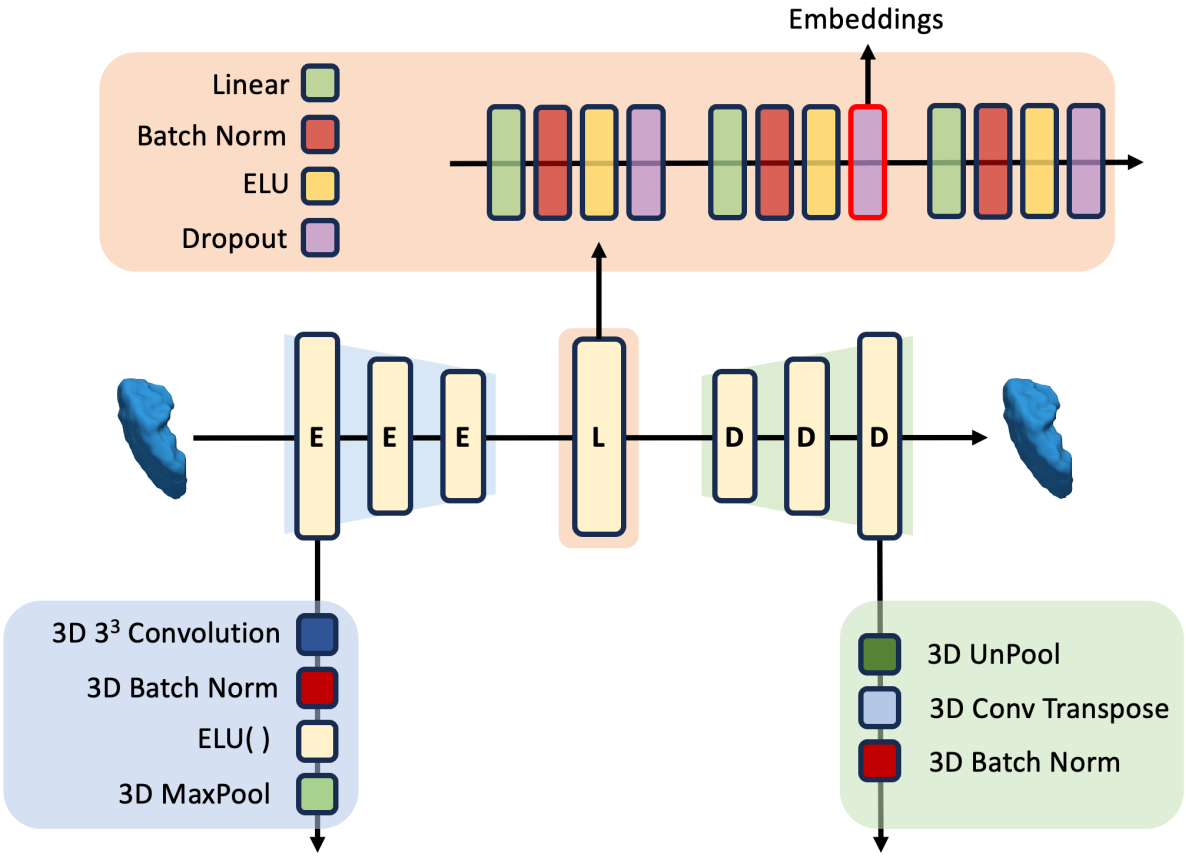


Figure 35: Architecture of the Autoencoder

The Variational Autoencoder

In addition to the traditional AE, there exists a probabilistic version that allows for the incorporation of stochasticity by learning parameters for, and sampling from a distribution in a method called the variational autoencoder (VAE). In the VAE, the *bottleneck* layer of the network learns parameters for distributions and the network samples from these distributions when passing input through the network, instead of generating weights for deterministic representations of the input. Latent distributions most frequently take the form of multivariate Gaussian but can be specified otherwise. In this architecture, an additional penalty term Kullback-Leibler divergence (KL divergence or D_{KL}, (see *Equation 7*) assesses how closely the

approximated *bottleneck* distribution $q(x)$ matches the *true* posterior distribution $p(x)$. Furthermore, we can modify our loss function to decide *how much* we want this penalty to factor into the generation of our loss term (see *Equation 8*). If the specified true posterior distribution is multivariate Gaussian with zero mean and unit variance $\mathcal{N}(\mu = 0, \sigma = 1)$ and the off-diagonal of the covariance matrix Σ is forced to zero, then the latent distributions encourage the representations in the bottleneck layer to be uncorrelated. A fact that allows for the emulation of uncorrelated projects in other techniques such as PCA.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx$$

Equation 7: Kullback-Leibler Divergence

The KL divergence term for a continuous distribution, ***P*** and ***Q*** are probability distributions and ***p(x)*** and ***q(x)*** are the probability density functions of ***P*** and ***Q*** respectively. In essence, this function identifies how closely ***P*** matches ***Q***.

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D}\sum_{j=1}^{D}\left(x_{i,j} - \hat{x}_{i,j}\right)^2 + \beta \cdot D_{KL}(P(z_i|x_i) \,||\, Q(z_i|x_i))$$

Equation 8: VAE Loss

In this loss function $\boldsymbol{\mathcal{L}(\theta)}$, ***N*** is the number of observations, $\boldsymbol{x_{i,j}}$ is the i^{th} observation of feature j. $\boldsymbol{D_{KL}}$ is the KL divergence term noted above in *Equation 7*, and β is a tunable parameter that controls *how much* we want to enforce the latent space regularization in that our approximated

distribution $\boldsymbol{Q(z_i|x_i)}$ matches the true posterior distribution $\boldsymbol{P(z_i|x_i)}$ when generating loss for the model.

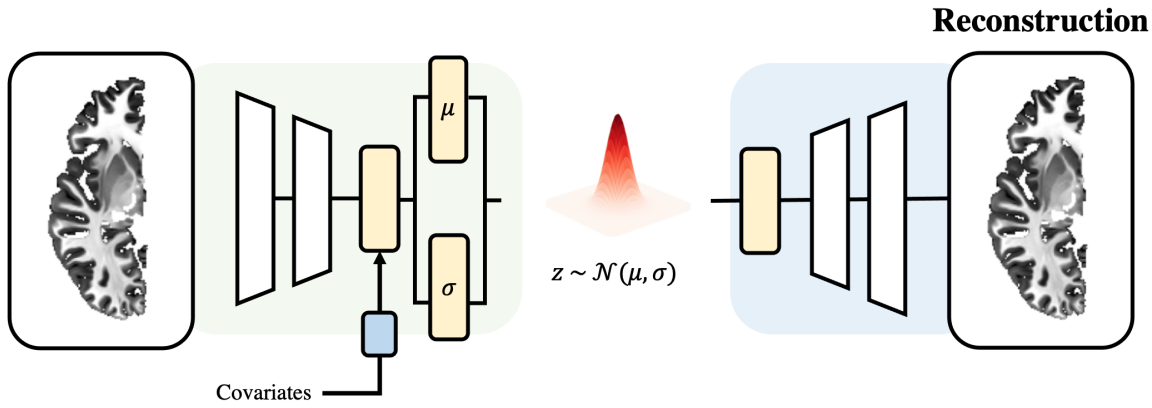


Figure 36: Architecture of the VAE

The proposed architecture of the VAE matches that of the traditional AE, save for the two linear layers estimating parameters of the multivariate Gaussian $\mathcal{N}(\mu = 0, \sigma = 1)$. These layers learn distribution parameters as weights in the model, and the latent representations of the input are sampled from this distribution before being fed through the decoder for reconstruction with the MSE and D_{KL} term. It is important to note that the sampling of latent distributions is stochastic in nature, and the reconstruction of an original image will vary slightly through multiple passes within the same network.

Region Specific Autoencoders

Additionally, we are evaluating a series of region specific autoencoders, similar to that of the multiple mini-batch PCA models discussed above. In this strategy, we are extracting the previously segmented (FreeSurfer, obtained via *recon-all*) structures from the T1 image using

the Desikan-Killiany atlas for the cortical, and Aseg (automatic subcortical segmentation) atlas for the subcortical structures, applying the previously mentioned minimum bounding cube, and feeding each autoencoder the segmented portion from the original T1 image. These regions are standalone, meaning no neighboring structures are included in that model, but preserve the original intensity of the voxels in that image (i.e., they are not binary segmentation masks). This strategy provides a few added benefits. First, this greatly reduces the dimensionality of each input image and complexity of the model. Instead of trying to learn representations of the entire brain within a single model, these region-specific models identify representations specific to each location, no convolutions or layers are shared between models. Additionally, this provides the benefit of being inherently more interpretable by way of creating distinct sets of representations for each brain region. While we may not readily interpret *what* exact aspects of intensity (tissue type) or morphology these latent representations characterize, we will know exactly which latent features represent which specific region of the brain.

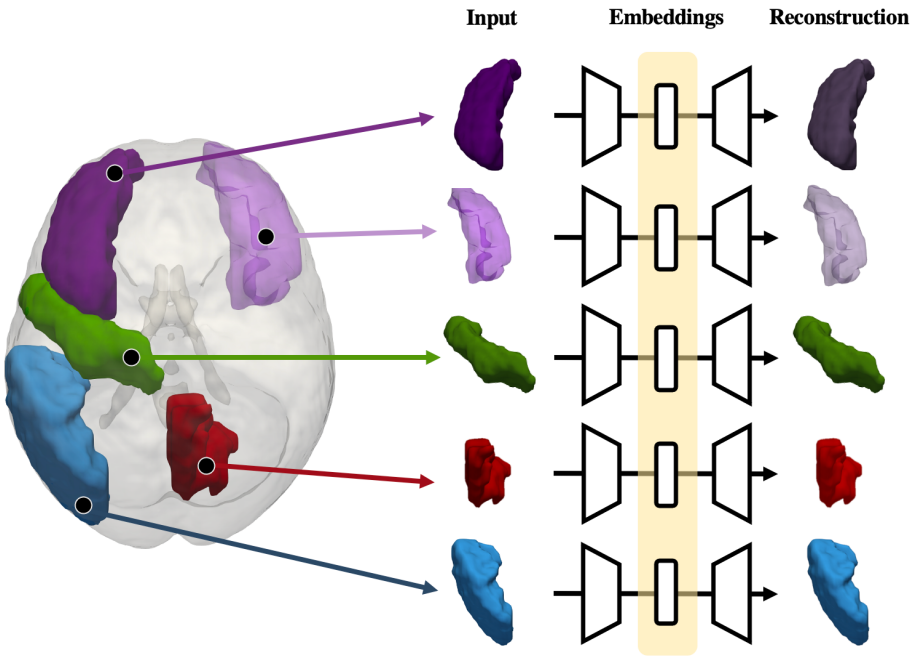


Figure 37: Region Specific Autoencoders

This method is also being used for the variational autoencoder. Additionally, it is important to note that the hyperparameter optimization will be done on a single randomly selected structure (superior frontal gyrus) in the AE, validated in a subsequent structure and these parameters will be used for all subsequent AEs. While this strategy inherently makes large assumptions about model architecture and ideal hyperparameters combinations over a series of unique structures, it is not computationally tractable to evaluate the full band of learning rates, number of layers, neurons, kernel sizes, activation functions, weight decay, and dropout for each of the 41 individual region specific networks. However, while we will ensure that all of the individual models converge and can accurately reconstruct the original structure for that given network, we note that these might not be the idealized parameters of each individual structure.

Supervised Models and Transfer Learning

We seek to examine the utility of the unsupervised latent embeddings in subsequent traditional SML (linear) models, we will also determine if supervised deep neural networks (including region specific supervised models) better predict working memory or externalizing disorders than the features from the unsupervised dimensionality reduction methods. Additionally, similar to Chapter 2, we will be evaluating these networks under two scenarios, the first in which we randomly initialize the weights in these networks, as well as transferring the weights from the AE encoder and pre-bottleneck linear layers into the supervised network and updating the weights in the model for improved prediction.

Results

Reconstruction error

The foremost critical aspect in evaluation is how well each model can capture aspects of morphology and intensity to 1) accurately reconstruct the original structure and 2) leverage the latent representations to subsequently predict WM and EXT. Both the AE and VAE were able to reduce the size of every original structure into a set of 20 neurons, or distribution parameters for the VAE, and are able to reconstruct the original image with impressively low error (average MSE: .0025 +/- .001). Furthermore, there was no difference in reconstruction error between the AE and VAE's. However, applying the inverse transformation of the PCA model to the projected components was not able to capture more nuanced aspects of morphology. In Figure 38, we can see the original image, reconstruction, and the squared absolute difference between the two. Furthermore, the evaluation of downsampling from the full image size to .8x reduction in dimension did not alter reconstruction error or predictive performance in the superior frontal

gyrus or amygdala (a randomly selected cortical and subcortical structure), but it is important to note that this was not evaluated in all structures due to the immense training time required per model.

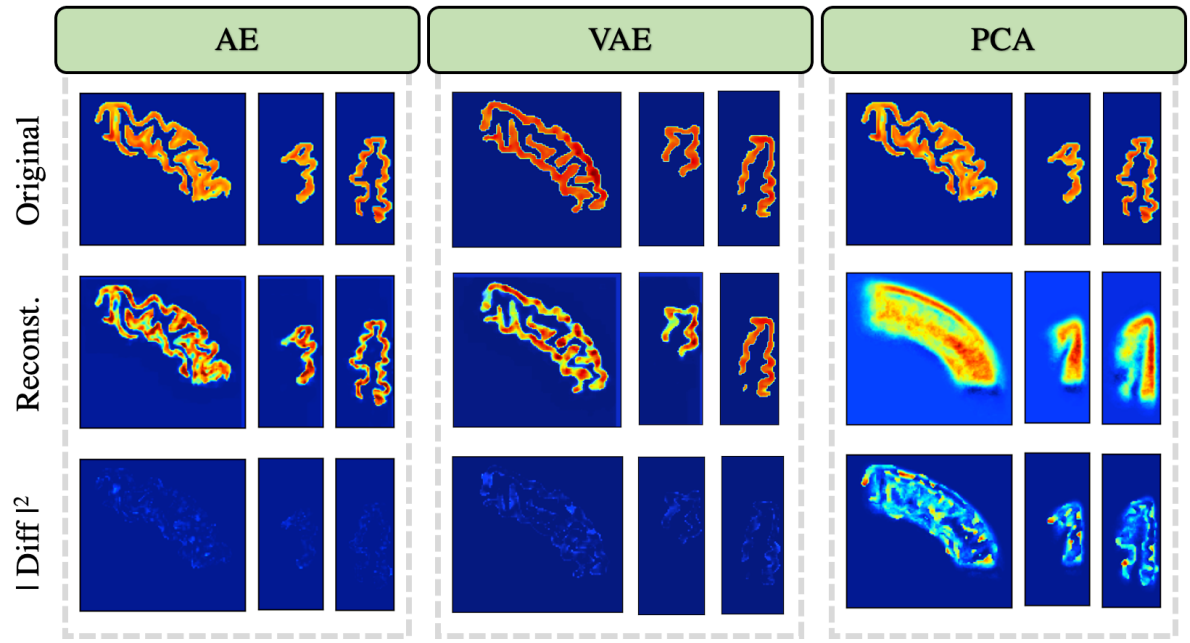


Figure 38: AE, VAE, and PCA Reconstruction

Stability and Network Stochasticity

In Chapter 2, we discussed the issues of inconsistent network convergence across multiple folds of training. Thus, in addition to the evaluation of reconstruction error, we want to assert that this error is consistent across both multiple instances of network and weight initializations and unique training folds. In the case of both the AE and VAE, the reconstruction error of each fold of each structure was consistent (an illustrated example of this in the superior frontal gyrus, first column Figure 39, mean MSE = .0020, standard deviation = .0005), in addition to the reconstruction between folds (MSE ~ .0010) also being properly consistent.

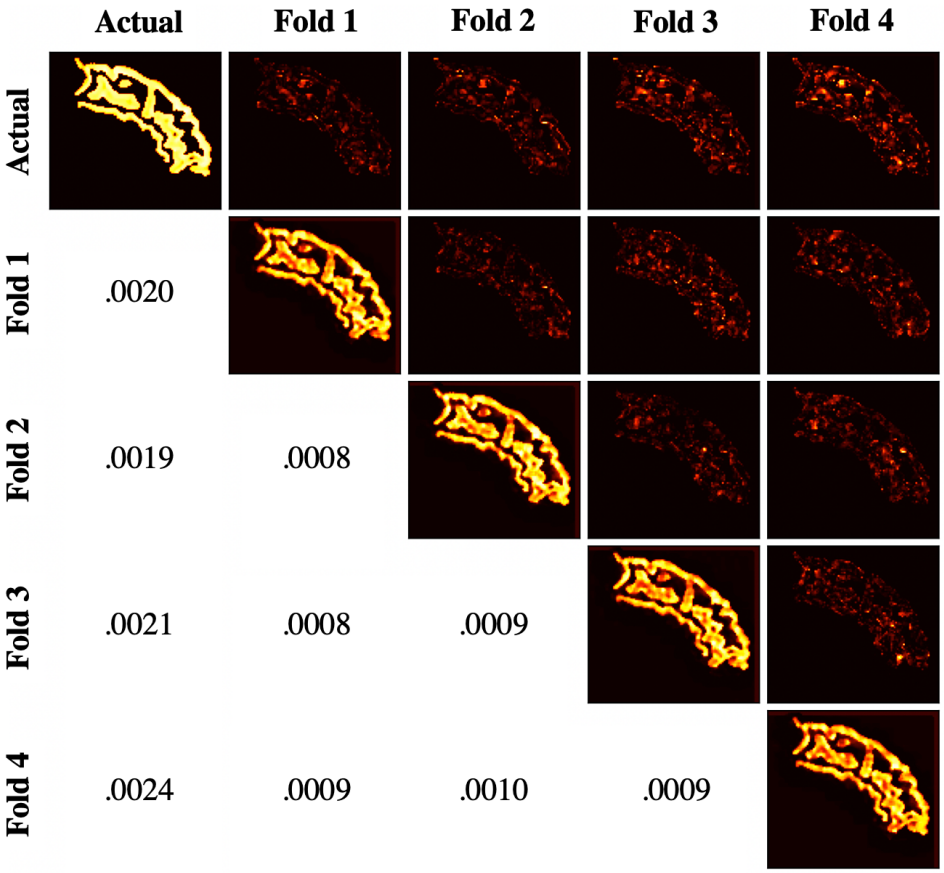


Figure 39: Stability of Autoencoder MSE

Actual image and reconstructions of the superior frontal cortex over several folds of autoencoder (AE) training. Diagonal displays actual (first item) and reconstructed images, upper triangle illustrates the squared difference between the actual and reconstructed images. Lower triangle, first column shows mean squared error (MSE) between the reconstructed and actual images, other cells show the MSE between the unique folds of AE training for the same observation.

Supervised Deep Neural Networks

As was the case in Chapter 2, we were unable to predict either working memory (WM) or externalizing psychopathology (EXT) better than random chance when using the supervised DNNs, using any structure or the entire brain. Furthermore, we found that the supervised DNNs achieved predictive performance that was, at best, equal to the linear models using embeddings from each region-specific AE, and only when using the weights from the encoder portion of the AE as initializations for the supervised networks. This was the case for both the prediction of working memory and externalizing t-scores and was the same in all structures. Due to both the supervised DNN results obtained from the previous chapter and these analyses, we will not be presenting results for the supervised models moving forward, as they performed worse than, or equal to models using unsupervised embeddings.

Predicting Working Memory

In addition to using the entire brain or segmented cerebral white matter we showcase the performance of predicting WM within each sub-structure (Figure 40). A few aspects of these predictions stand out; first, within ***all*** subcortical structures, FreeSurfer statistics are themselves not significantly predictive of WM (indicated by gray fill color). Critically, only when utilizing the imaging data directly with either PCA or the AE embeddings do we find significant predictive utility from these structures.

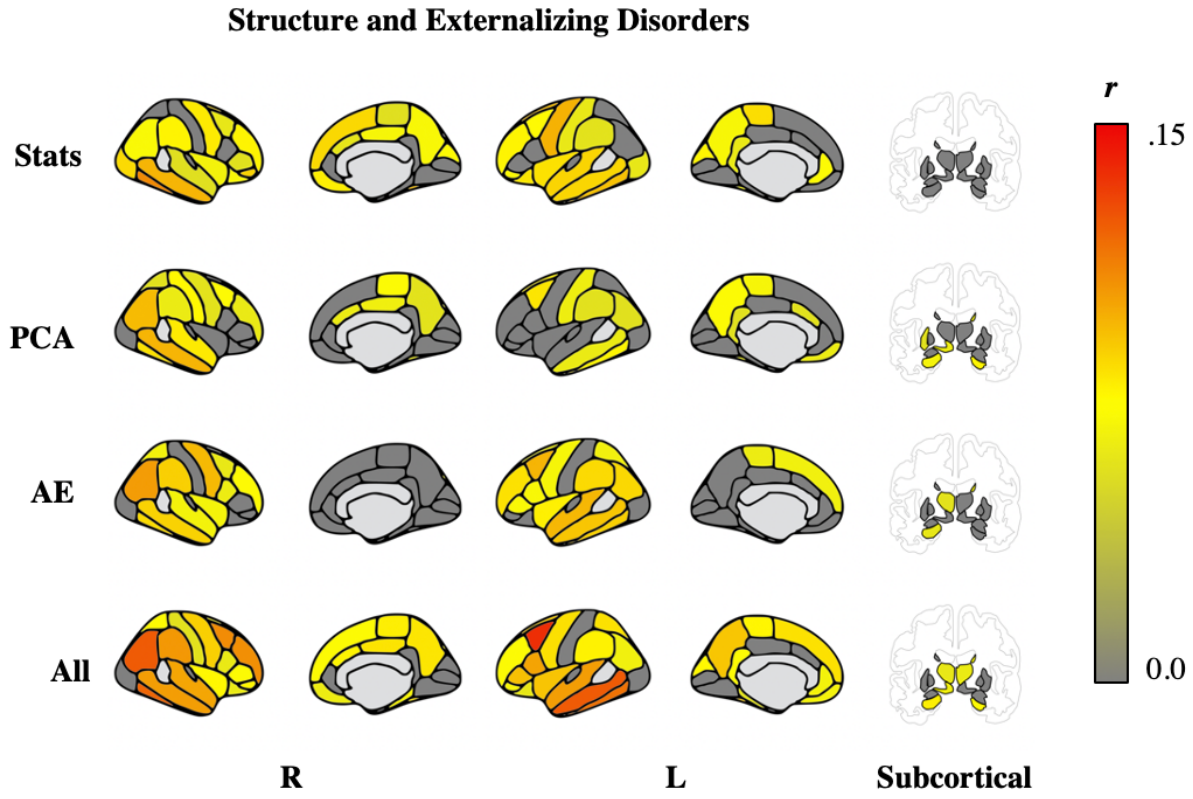


Figure 40: Predicting Working Memory

Linear regression working memory model performance, via Pearson correlation coefficient (r) of actual vs predicted working memory, using FreeSurfer statistics (stats), PCA components (PCA), autoencoder embeddings (AE), or a combined set of each of these strategies together. Views of the right (column 1 and 2) and left (column 3 and 4) hemispheres in addition to the subcortical view (column 5). Permutation testing was used to indicate region significance, and regions in gray were not statistically ($p > .001$) better than performance from 100 permuted models.

Interestingly, there are hemispheric differences in predictive utility across all feature methods, with the left hemisphere often, but not always, outperforming the right. Additionally, PCA outperforms that of the AE in all structures save the isthmus cingulate. Furthermore, there were several medial structures that performed significantly (Tukey HSD $\alpha < .001$) better using

only the FreeSurfer statistics including the cuneus, precuneus, paracentral, and posterior cingulate in both hemispheres (see Figure 41).

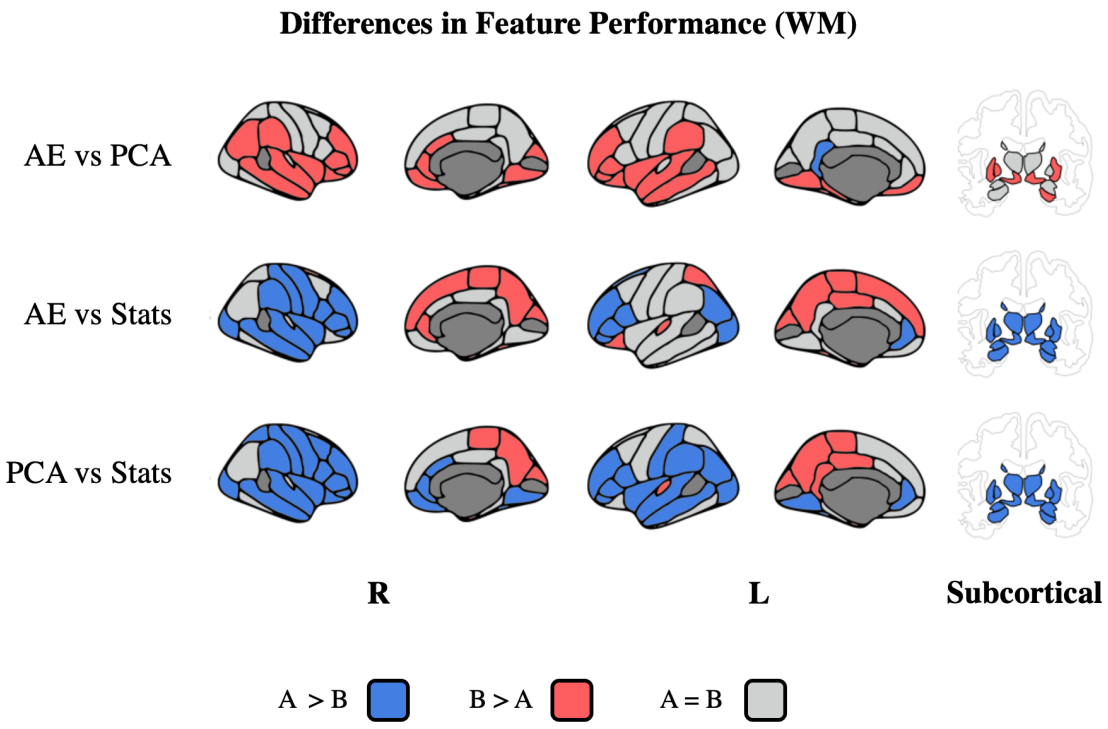


Figure 41: Differences in Predictions of WM features

Rows indicate a statistical comparison of structure performance of "A and B". Blue regions indicate that "A" performed significantly (Tukey HSD alpha < .001) better than "B", thus in the top row "all" is A and "AE" is B. Red regions indicate the opposite, B performed significantly better than A.

Perhaps one of the more interesting aspects is that the performance obtained from using both the left and right PCA components of the middle temporal gyrus obtained a performance ($r = .27$) that is not significantly different than the analysis in Chapter 2 leveraging ***all*** of the FreeSurfer statistics or from the use of all FreeSurfer statistics ***and*** the resting state data.

Additionally, the inferior temporal gyrus and precentral gyrus were also within the top performing structures predicting WM. The inferior temporal gyrus performed notably better over both hemispheres using all of the features, including the PCA components, AE embeddings, and FreeSurfer statistics. Specifically, all features performed better than imaging-derived AE or PCA features. However, we found no significant difference in predictive performance within the precentral gyrus between the AE embeddings and PCA components. These findings speak heavily to the inconsistent benefits of features obtained by each of the imaging derived feature extraction methods (AE vs PCA), with none of the top structures achieving best results using the FreeSurfer statistics alone.

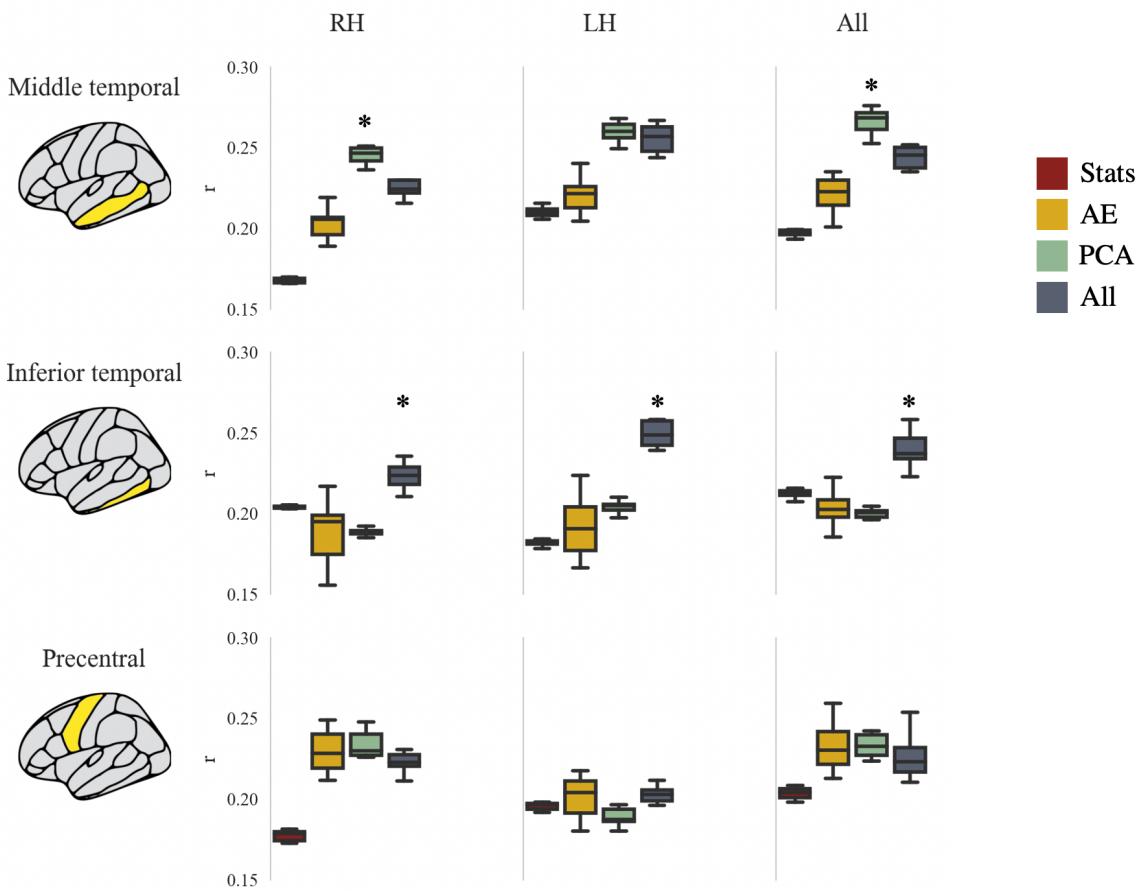


Figure 42: Top structures predicting WM

The top three structures that predict working memory (WM) by hemisphere. Asterisks (*) indicate a feature set that does significantly ($p < .001$) better than the other available feature methods.

Of the more striking improvements in a single structure's prediction was that of the cerebral white matter segmentations using the PCA components. These PCA models retained, on average, ~20% of the total variance from the original white matter segmentation data, with the first principal component containing ~7% of said variance. This fact alone is staggering, that an image containing over 1 million pixels can be compressed into only 20 features, roughly .002% of the original image size and retain such a large portion of the variance. Additionally, there were no significant (p's > .01) differences between the variance explained across hemispheres. While the comparison may be inequitable, as only a single feature (volume) represents this entire structure from the FreeSurfer statistics, the imaging data significantly outperformed the traditional extracted feature. The right hemisphere performed significantly better than the left and there was a significant improvement in PCA ($r = .18$) over the AE but not in the left hemisphere.

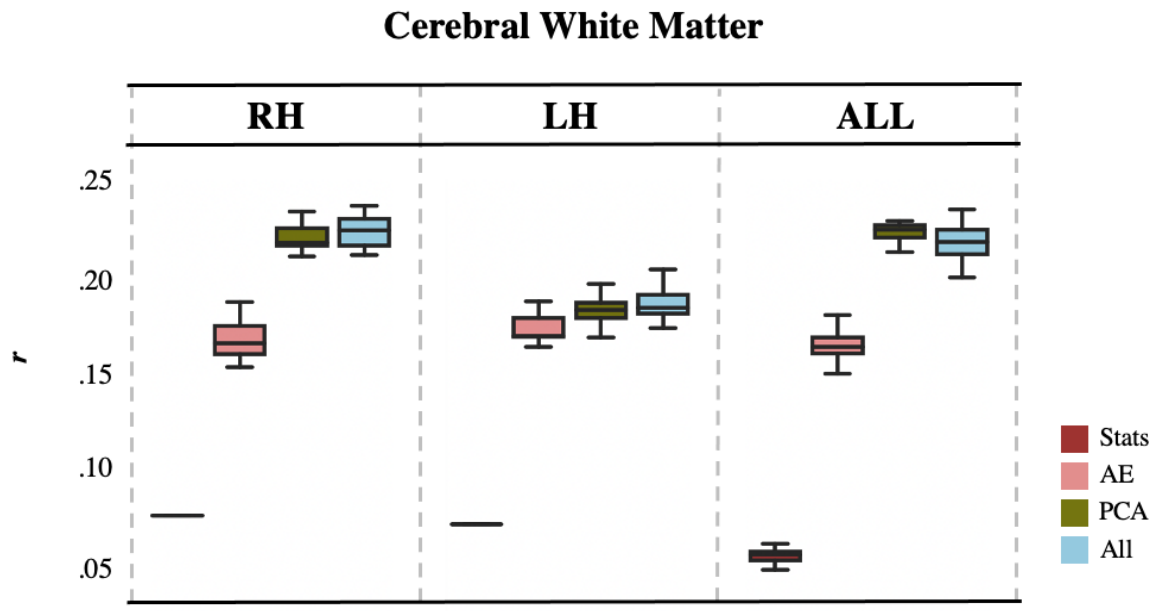


Figure 43: Cerebral White Matter predicting WM

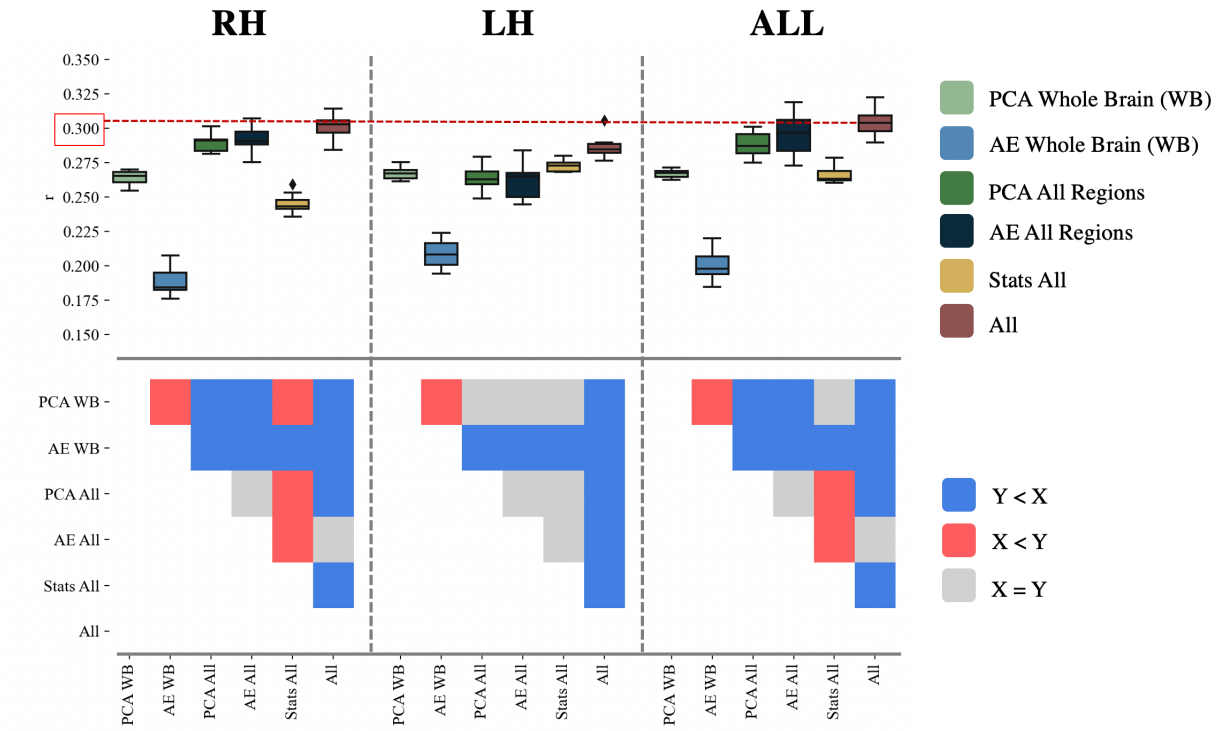


Figure 44: Using all Features in WM Prediction

Illustrated above is the final performance in predicting working memory (WM) using all of the available information. "Whole brain" refers to dimensionality reduction on the full T1 image (by hemisphere if specified), that is no segmentation information (other than hemisphere) is used in the generation of the latent data. "All regions" denotes that the embeddings or components from all of the individual AE or PCA models were concatenated and used in prediction. Stats all means that FreeSurfer statistics were used, and "all" is the combination of every method just described. Bottom row: Pairwise Tukey HSD for each of the mentioned methods, red indicates group in the Y axis performed significantly ($p < .01$ FWER: .05) than the corresponding X axis cell, the opposite is true for blue, with grey denoting no significant differences.

Ultimately, in predicting WM, the highest performance obtained over all strategies was achieved by using features from all methods (PCA components, AE embeddings, and FreeSurfer statistics) over both hemispheres ($r = .30$). No differences in predictive performance was obtained when using either all of the PCA components versus all region AE embeddings. Interestingly, in the left hemisphere, the FreeSurfer statistics performed equivalent to the AE embeddings and PCA components but not the right or both hemispheres. This is contrast to the left hemispheric findings of improvements with PCA and the AE in the top two structures (middle temporal gyrus and inferior temporal gyrus). Altogether these findings highlight the added predictive utility of features obtained from the T1 imaging data using unsupervised SML and DNN strategies, in addition to the use of FreeSurfer statistics best predict WM.

Predicting Externalizing Disorders

As mentioned above, we did not see improved performance when using the supervised DNNs for predicting externalizing disorders (EXT) (see Figure 45). However, it is interesting that we obtained equivalent or even improved performance using the information from single structure unsupervised features compared to the analyses in Chapter 2 using all features from the T1 sMRI and rs-fMRI data. Furthermore, as seen for WM, we again find that subcortical structures were useful for predicting EXT only when using additional information from the imaging data directly, either via the PCA components or AE embeddings.

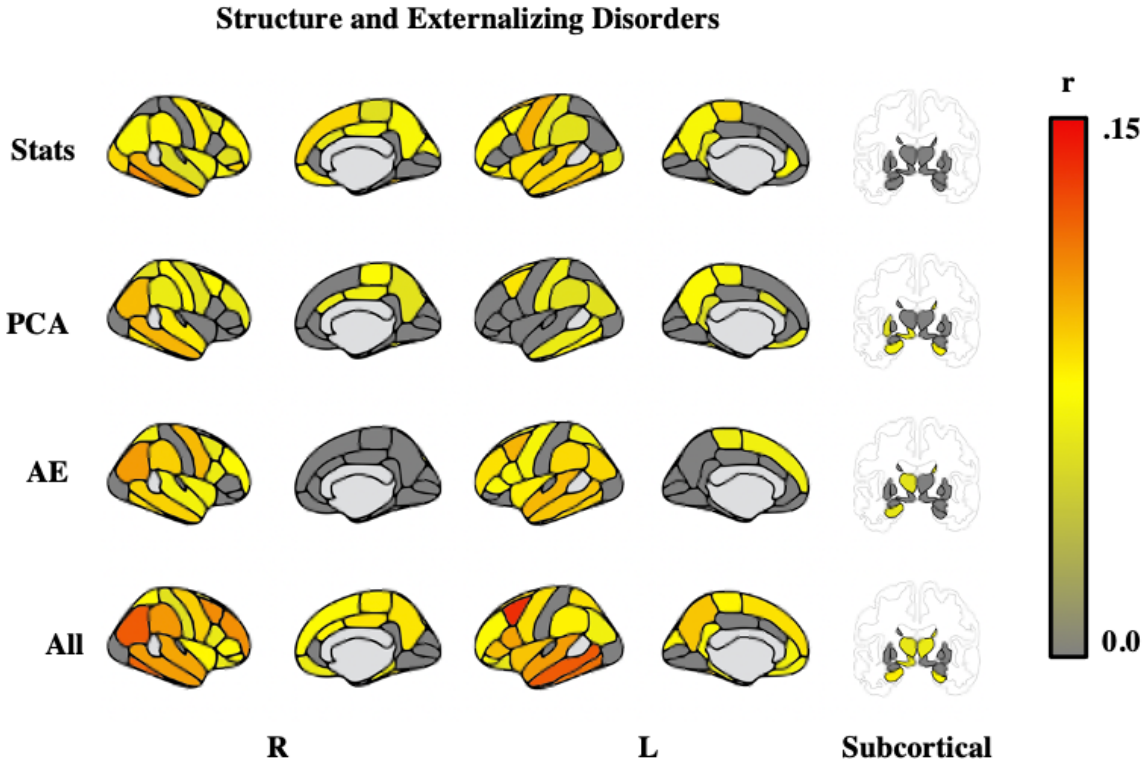


Figure 45: Predicting EXT

Dissimilar to the prediction of WM, the AE embeddings ability to predict EXT is either better than or equivalent to that of the PCA components (see Figure 46). Furthermore, it is critical to reiterate that these are not features obtained from supervised models. The features from, for example, the middle temporal gyrus are the same used to predict WM as they are to predict EXT. Thus, each predictive model is utilizing the same latent information from the original structure in a different manner for each targeted outcome. As noted with WM prediction, there are also structures that perform better when using only the FreeSurfer statistics alone, for example, the paracentral gyrus in the left hemisphere.

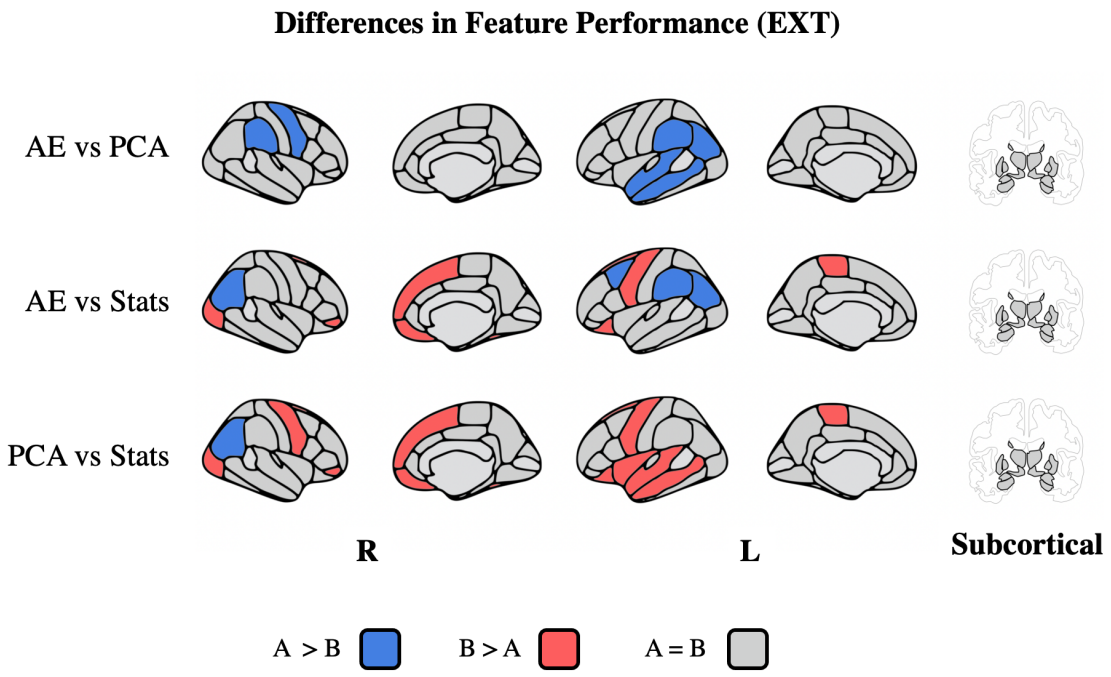


Figure 46: Differences in Predictions of EXT Features

Additionally, two of the top performing structures that predict EXT are also within the top structures that predict WM, specifically the middle temporal gyrus and the inferior temporal

gyrus. Similarly, the left hemisphere in the middle temporal gyrus significantly outperforms the best in both situations.

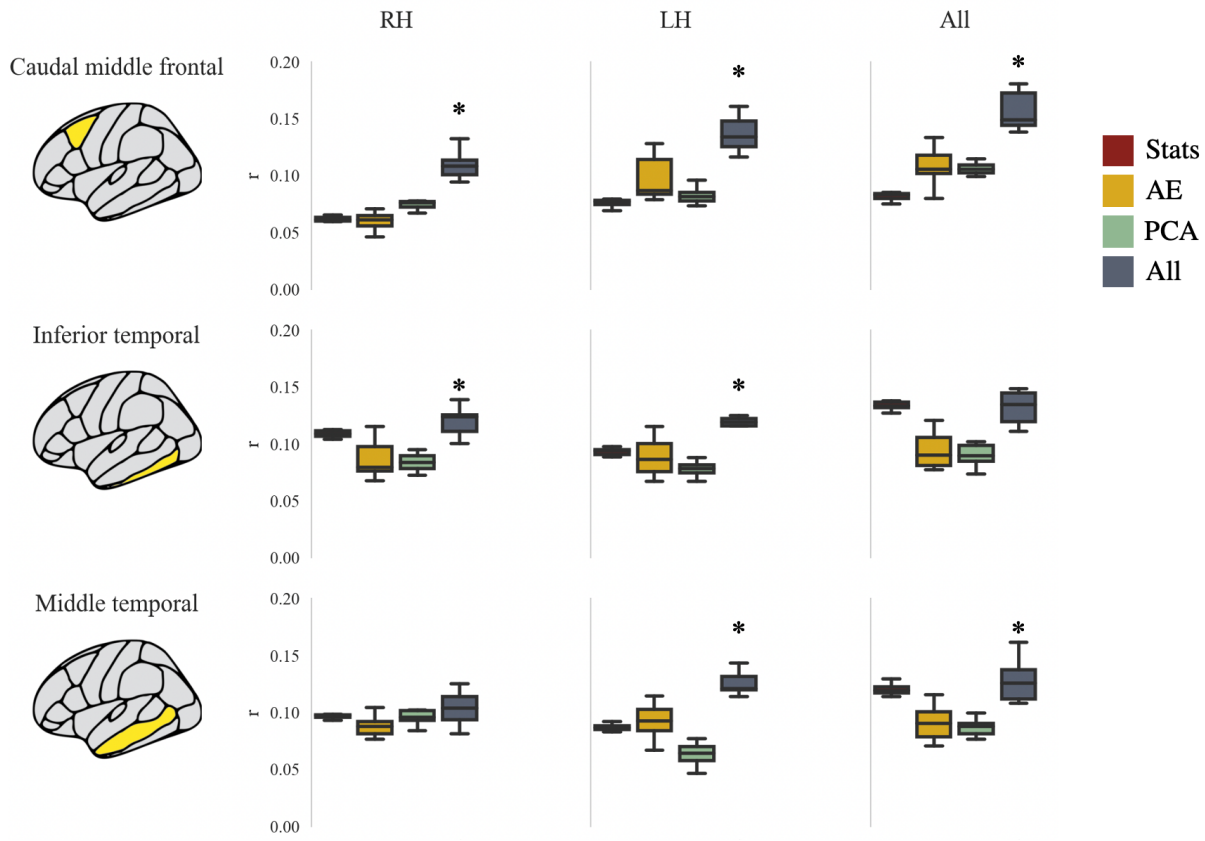


Figure 47: Top Predictive Structures for EXT

Features from the FreeSurfer statistics (stats: red), autoencoder embeddings (AE: yellow), PCA components (PCA: green) and all features together (All: blue).

In most situations, we see an improvement when utilizing features from all available strategies, save for in the case of both hemispheres in the inferior temporal gyrus and the right hemisphere of the middle temporal gyrus. Ultimately the top performing single structure

predicting EXT was the caudal middle frontal gyrus using both hemispheres and all feature subsets (*r = .15*). This is a significant improvement over the performance seen in Chapter 2 and again, highly interesting that it is from a single region of the brain.

Unlike the utility of imaging derived cerebral white matter features, neither PCA, the AE embeddings, or FreeSurfer statistics were able to predict EXT using this structure any better than permuted models (p's > .01).

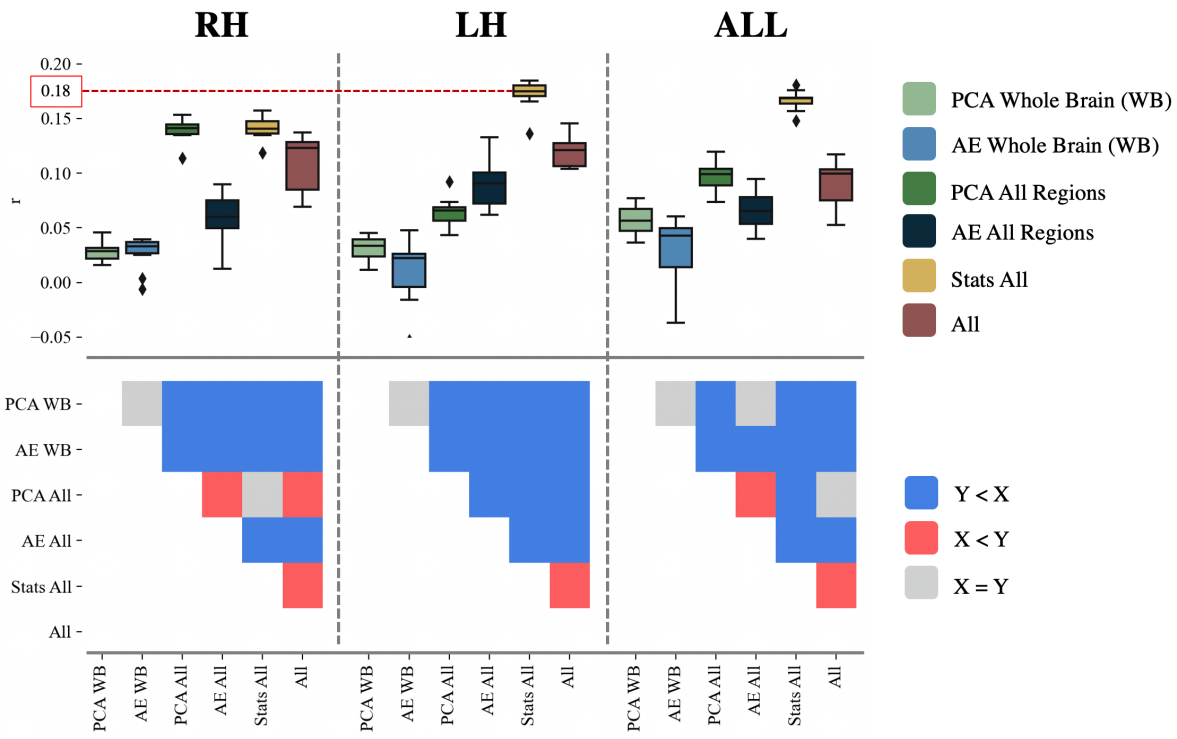


Figure 48: Using all Features in EXT Prediction

This figure follows the same format as Figure 44

The top performing method predicting EXT utilized only the FreeSurfer statistics within the left hemisphere (*r = .18*); however, this is not significantly different than using both

hemisphere statistics together. Interestingly, this is a performance substantially better than that obtained in Chapter 2. However, it must be noted that the sample changed between the two analyses. Without the need for exclusion due to missing rs-fMRI, we were able to include several thousand additional subjects. We also did not decompose predictive models by hemisphere in the analyses in Chapter 2. While we did see improvements utilizing additional information from the AE and PCA models in single structure models, they did not ultimately improve EXT prediction when leveraged with the information from the other structures.

Discussion

Altogether, these results paint an interesting, yet complicated picture toward the accessibility and predictive utility of residual information within in T1 sMRI data beyond the typically leveraged FreeSurfer summary statistics. Ultimately, these results assert that utilizing information from all methods of extracting information including latent representations maximizing variance (PCA), DNNs compressing and nearly perfectly reconstructing input (AE), in addition to the typically used FreeSurfer statistics produced the best predictive performance for WM. However, this was not true for the prediction of EXT. While there were interesting region-specific improvements predicting EXT using features from the AE, ultimately the highest performing composite included only information from the FreeSurfer statistics with no additional PCA or AE derived imaging features. Below, we take a deeper look into some of the considerations, issues, and interesting aspects that arose from the analyses present above.

The Top Predictive Single Structure Models

Interestingly, two of the top three performing structures (middle temporal and inferior temporal gyrus) were seen in both the prediction of WM and EXT. A large meta-analysis of 120 fMRI studies implicated the middle temporal gyrus' role in semantic memory, or "*the knowledge, about people objects, and actions, and culture learned through experience"*, notably in a left-lateralized network along with several other structures within the prefrontal cortex (Binder et al., 2009). While semantic memory, a type of accumulated and accessible experiential knowledge retrieval process, differs from the ability to hold, manipulate, and utilize information in memory to complete tasks, some researchers posit that these systems are not as distinct as previously believed. Evidence from PET, fMRI, and event-related potential (ERP) studies may link WM and prefrontal areas to the retrieval of long-term memory from posterior areas and maintain this past information in an active state (Marklund & Nyberg, 2007). Furthermore, atrophy within the left inferior temporal gyrus, a structure commonly implicated in visual processing, was found in both populations of patients with Alzheimer's Disease and semantic dementia, both disorders marked by large dysfunction in semantic and WM processes (Chan et al., 2001).

While not within the top three structures, the rostral middle frontal ($r = .24$, PCA), fusiform gyrus ($r = .23$, PCA), and the superior temporal gyrus ($r = .23$, PCA) were also important in the prediction of WM. These regions have been implicated in many processes but include WM and higher order cognitive process (Friedman & Robbins, 2021), facial recognition (Fur et al., 2011), and auditory short term memory processes respectively (Leff et al., 2009).

The middle temporal gyrus and inferior temporal gyrus were also important in models predicting EXT disorders. Additional regions not found within top features predicting WM

included the caudal middle frontal and inferior parietal gyrus. Interestingly, this is in contrasts to research from within the ABCD study which found cortical thickness from the left caudal middle frontal gyrus and the right inferior parietal gyrus to be significantly altered in patients with *internalizing* disorders, but not among those with externalizing disorders (Yu et al., 2023). However, it is also important to note the high degree of overlap in internalizing and externalizing disorders and symptomatology.

In speaking toward model interpretability, the linear decomposition (and subsequent inverse transformation) of PCA is natively more interpretable than DNNs. That is, we can examine the coefficients from each projection to identify specifically *where* (what pixel) contributes to the projection of each component. Interestingly, we find similar elements in both the cerebral white matter components and whole brain PCA models. The first component of the cerebral white matter (CWM) PCA model captures the majority of the variance and represents the largest portions of CWM in the brain, with the subsequent three components capturing smaller regional portions of white matter. Interestingly, the fifth component also captures the majority of tracts with less extension into the gray matter boundaries. Furthermore, the first component in the whole brain PCA models also reflect these regions of CWM, with the addition of the entire brain segmentation boundary. Components two, four, and five seem to reflect the outline boundary of the segmented T1 images. While difficult to speak to definitively, these elements could reflect both whole brain estimates and regions of high variability, such as the borders of the segmentation boundary, as these regions would be highly variable between subjects. Finally, the third component again attributes CWM with greater extension into the gray matter boundaries (see Figure 49).

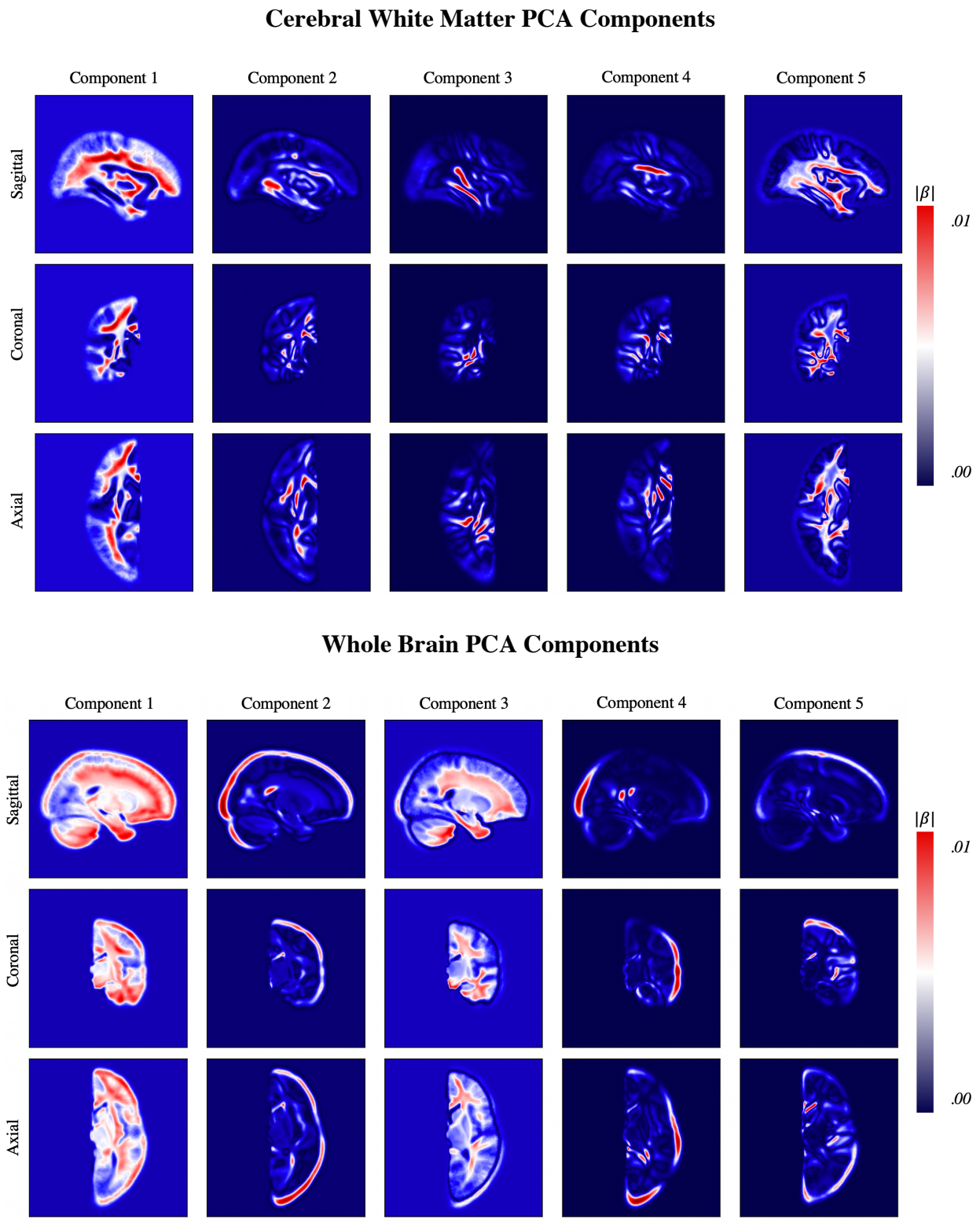


Figure 49: PCA Components of Cerebral White Matter and Whole Brain Imaging Models

The absolute value of the coefficients for the top five components from the PCA models using either the segmented cerebral white matter or entire T1 sMRI hemisphere (shown here left-hemisphere).

Non-Additive Region-Specific Utility

Interestingly, the effects of including information from multiple successful structures was not additive. That is, while the top three structures, from the region-specific models, obtained good performance (r = .25) in predicting WM, the best predictive performance obtained when using all structures (r = .30) was significantly, but only slightly (.05) better. The effects were non additive, meaning that the performance of using structure 1 (s1) and structure 2 (s1) in single model did not obtain the individual performance of $r_{s1} + r_{s2}$. These results may speak to the relationship of regional contribution to outcome performance, with whole brain PCA and AE models underperforming that of even the top individual structures. This point is highly interesting as it speaks to locality in prediction. It seems that the hypothesis of *more information is better* is not always inherently true when speaking to the identification of latent representations of whole brain vs localized regions and subsequent brain-behavior modeling.

Poor Predictive Utility of VAE embeddings

While the trained VAE's obtained nearly perfect reconstruction error in the holdout set and created uncorrelated latent representations, the predictive utility of these features for either WM or EXT was never better than random chance. It is again critical to consider that the objective of each of these DNN and PCA dimensionality reduction methods evaluated have unique goals. It is likely the additional constraint (D_{KL}) forcing the learned parameters to follow a

multivariate Gaussian distribution alters *what* is being learned in these latent representations. Furthermore, it is possible that the stochastic nature of the sampling perturbs the consistency of these learned representations.

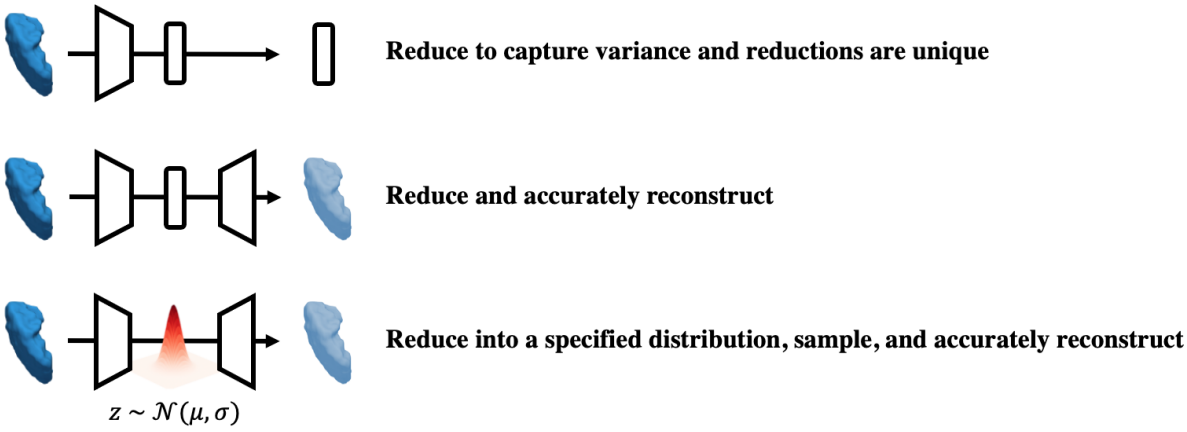


Figure 50: Objectives of Dimensionality Reduction

Identifying an uncorrelated set of embeddings was the primary goal behind the use of the VAE as we discuss issues of multicollinearity within the AE embeddings in the following section. However, traditionally, variational autoencoders are often used for the generation of synthetic data along with an additional architecture arm (the discriminator) that assess the validity of said synthetic examples. In the same realm, these networks are also used in the application of data augmentation to make networks more robust to noise and introduce differential properties into the original images in the goal of improving model generalizability (Kebaili et al., 2023). Additionally, it is possible that the inclusion of the D_{KL} term in the loss function may under-exploit the capacity of the network and create an undesirably noise latent space in prioritizing this additional error term to match the *a priori* latent multivariate Gaussian

distribution (Asperti et al., 2021). While we attempted to balance this distribution matching term and the minimizing of reconstruction error (MSE), it is difficult to tune both terms simultaneously while understanding the effects of the subsequent brain-behavior predictive utility (Asperti & Trentin, 2020). Altogether, though the VAE did learn uncorrelated low-dimensional representations of the original image capable of accurate reconstruction, we did not find the predictive utility of these features warranted the added complexity and training time.

Limited Utility of Evaluated Supervised DNNs

In Chapter 2, we discussed several aspects that may have made the DNNs using direct imaging data challenging. And likely, each of these aspects apply in the same sense for the task of predicting WM and EXT. While we were able to substantially reduce the size of model inputs by evaluating networks using single segmented region-specific structures, sometimes as small as 33x34x33 pixels, these supervised models did not improve the ability to predict either WM or EXT. As mentioned in the results, at best we were able to obtain equal performance if we used the pretrained AE network's encoder and froze the early layers of the network, which inherently emulates the method of using the latent embeddings from the AE in a linear model, only in a more complicated procedure. However, these results must always be taken in the understanding that we evaluated a small set of hypothesized architectures and hyperparameters to evaluate in a supervised setting. There are again, an infinite number of combinations, hyperparameters, and network configurations that prevent any such blanket statement of "*supervised deep neural networks cannot predict working memory or externalizing disorders*." Still, it is worth restating that the top performing models from the *ABCD Neuro-Cognition Prediction Challenge* (discussed at length in Chapter 2) were not deep neural networks, but algorithms falling into the

category of SML methods (Mihalik et al., 2019). As noted in Chapter 2, it is possible that the data required to learn highly complex traits of both cognition and psychopathology in the context of these large neuroimaging datasets is beyond what we had available within these analyses.

Computational Challenges and Optimization

One of the key aspects to discuss from these analyses is in regard to the challenges associated with optimizing region specific models. As highlighted in the methods section of this chapter, it is computationally infeasible to optimize each of the 41 region specific model architectures individually. While we did obtain favorable and consistent reconstruction errors (MSE) across every structure, we cannot comment as to how further region specific optimization might benefit some of these individual structures. However, the aspect of training so many DNNs creates a unique set of, unfortunately common, challenges. This is despite using several methods utilized to speed up model training on the available RTX 3060 GPU. Most notably, the use of automatic mixed precision (AMP) training, which significantly reduced model training time, sometimes up to half of the duration required using traditional training methods. This strategy dynamically sets the precision of matrices during training (e.g., 16- or 32-bit floating point integers) in order to maximize computational efficiency when possible without altering the numerical stability (Narang et al., 2017).

Additionally, we sought to create the simplest models (i.e., pruning layers and neurons within each layer, including the bottleneck layer) such that we could maintain stability of the reconstruction error (MSE) in the validation set and also minimize training time. However, the size and depth of these evaluated architectures creates a monumentally vast search space for model optimization that is often unfeasible and computationally prohibitive (Bergstra et al.,

2012). Individual region specific models provides latent regional specificity but exacerbates this challenge by introducing additional hyperparameters related to the selection and aggregation of individual model predictions. This further complication of the optimization process does not guarantee a *universally* applicable optimization configuration across the diverse set of unique structures and image properties (Huang et al., 2017).

Despite the use of much smaller individual region specific structures, the decomposition of large matrices into the bottleneck layer's latent space of 20 neurons is an extraordinary task. While we evaluated bottleneck configurations from 10 neurons up to 500 neurons, we saw only marginal (< 2%) improvements in reconstruction error (MSE), compared to the decided 20 neurons, with enormous added complexity and, at times, diminished generalizability. Again, this optimization was generalized across structures and the entire brain and may not be the ideal region-specific configuration for every structure, but is again, a noteworthy limitation of this modeling strategy and set of analyses. However, it must be said that any model leveraging imaging data this large will suffer similar computational limitations. This is made apparent in the required use of a special implementation of PCA able to train models in batches and iteratively update the projections. The majority of SML methods would also struggle or simply be unable to work with these data due to memory constraints. Furthermore, while we were able to leverage the mini-batch version of PCA, these models also took hours to train and were difficult to optimize as they were influenced by the size of the batches used to update the models.

Moreover, we must again consider that levels of abstraction are occurring in the generation of the AE embeddings. It is important to consider them within the context of the decoder, which reconstructs the original image. As such, the decoder architecture and reconstruction loss play a pivotal role in shaping the characteristics of the latent space. It is not

the same as PCA where a linear projection maps the original image into a set of features and the subsequent transposition of this transformation matrix allows for ready reconstruction. There are an entire series of learned weights that non-linearly map these latent embeddings to best reconstruct the original image. This again is not to imply or disregard their utility, it is simply a reminder of the complexity and context involved in their creation. Furthermore, they are not readily interpretable. However, creating region specific embeddings provides a spatially (segmentation region) constrained means of interpretability.

Multicollinearity

As mentioned previously, there exists inherent collinearity or relationships between features being created within each modeling strategy. Multicollinearity among features presents issues for traditional linear modeling strategies and can reduce accuracy and yield unreliable coefficient estimates (J. H. Kim, 2019). Perhaps most important in the case of the AE is that of redundant information. If the primary objective is to learn latent representations of the data, but several of these representations embody similar aspects of morphology or image intensity then we may not be fully utilizing the subtle and unique bits of information from the original image in subsequent modeling. For example, if we represent the middle temporal gyrus with 10 embeddings, but three of those embeddings are highly correlated, then we theoretical only have eight unique embeddings. It is precisely this aspect that again makes PCA such an attractive strategy for compressed representation as we ensure that each of the components represents entirely unique pieces of information residing within the original image. This fact is critical not only for traditional linear modeling strategies used here, but also to minimize information redundancy.

It is important to understand how these latent representation capture both characteristics of morphology and image intensity and subsequently, how this may affect model performance. One interesting portrayal of this scenario is captured in Figure 51. Here, we see the relationships between both PCA components and the AE embeddings in addition to their relationship with the FreeSurfer statistics of the precentral gyrus and middle temporal gyrus. The size of each node represents the coefficient from the linear model of that set of features (PCA, AE, or FreeSurfer statistics) in predicting WM. Edge color and width represents the correlation (Pearson Correlation Coefficient, *r*) between that latent representation (i.e., the PCA component, AE embedding, or the FreeSurfer statistics of either morphology or intensity). Features associated with morphology include mean thickness (thick), volume, and surface area (area). Other features represent average intensity within a structure such as GM/WM (a measure of the contrast between the gray and white matter in the cortex), GM (average gray matter intensity), and average white matter intensity.

Noted immediately is the larger size, and thus higher correlation, between the aspects of morphology within the middle temporal gyrus compared to that of the precentral gyrus. More interestingly, the AE embeddings from the middle temporal gyrus reflect substantially less (noted by the thinner and lighter colored edges) facets of morphology than that of the precentral gyrus. Thus, not only are the FreeSurfer statistic measures, that reflect morphology, more predictive of WM within the middle temporal gyrus, but the AE representations are less associated with these aspects of said morphology. Additionally, this phenomena is not the case for the PCA components, as they represent approximately equal aspects of morphology across both structures. While the association of the unsupervised latent representations and aspects of morphology and intensity does not inherently speak to subsequent brain-behavior modeling

utility, the fact that PCA significantly outperforms the AE in predicting WM within the middle temporal, but ***not*** the precentral gyrus is noteworthy. While the generation of all pairwise evaluations of this property is neither visually nor computationally practical, it is interesting to highlight this specific scenario that played out within the top three informative features.

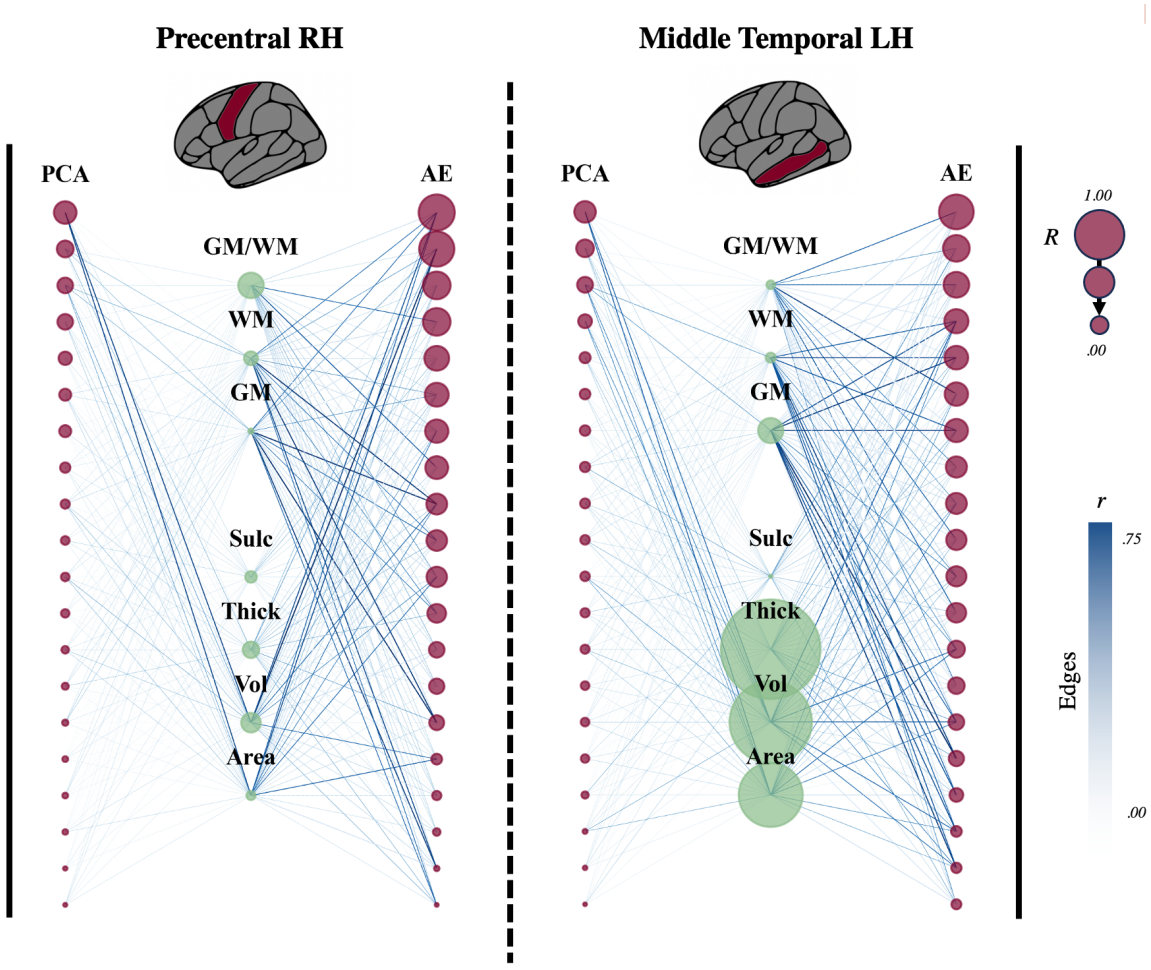


Figure 51: PCA and Autoencoder (AE) Feature Associations

Relationships between latent representations of the imaging data and FreeSurfer statistics from two different brain structures (precentral and middle temporal gyrus).

In addition to how these embeddings relate to high level FreeSurfer statistics, we can also examine how structure specific collinearity may alter model performance. Continuing with the examination of the middle temporal and precentral gyrus, we see that not only do the latent embeddings within the lower performing middle temporal gyrus not reflect aspects of morphology, but that they are also significantly ($p < .001$, ANOVA) more collinear than the precentral gyrus (see Figure 52). Therefore, we have two critical components to consider.

1. *High-level aspects of morphology are more correlated with WM, but the AE embeddings of the middle temporal gyrus are less correlated with morphology.*
2. *The underperforming AE embeddings of the middle temporal gyrus exhibit significantly higher collinearity, and thus have inherently greater information redundancy.*

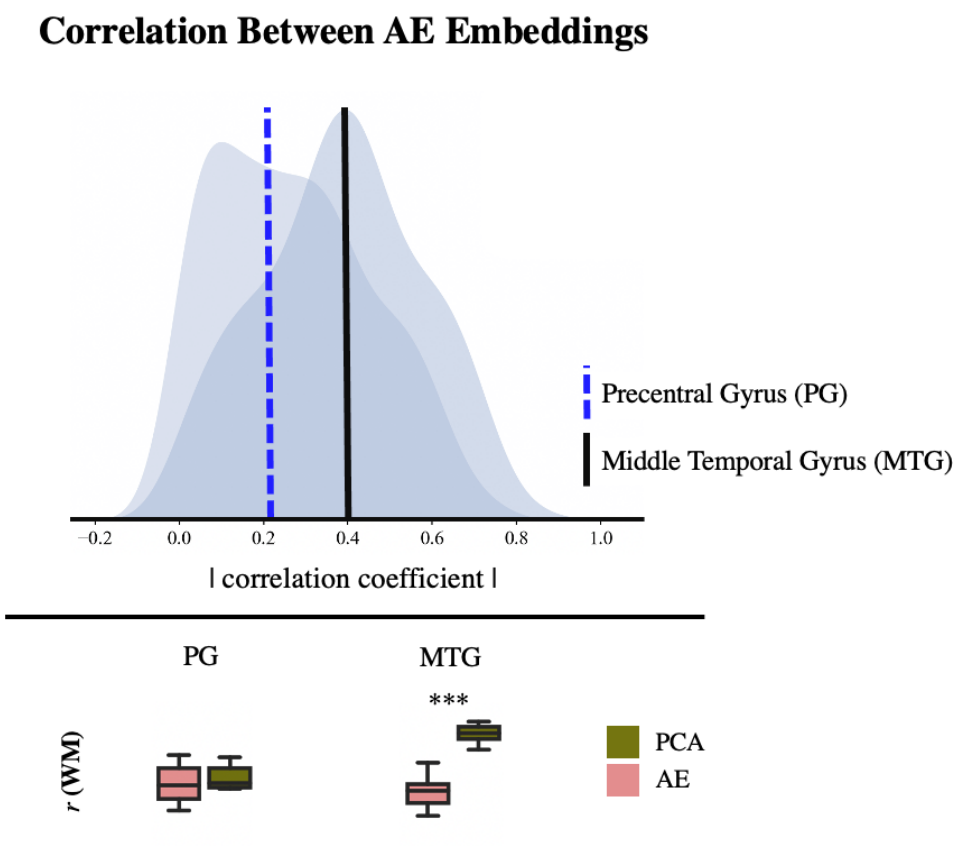


Figure 52: PCA vs AE Multicollinearity

Distributions of correlation coefficients within the autoencoder embeddings from the precentral and middle temporal gyrus (top). Dashed line indicates distribution mean for each structure. Bottom shows the relative correlation between the two feature sets (PCA and AE) in predicting WM, asterisks (***) denote p < .001 ANOVA.

Additionally, we can examine this property of multicollinearity over each feature set from the feature correlation matrices in Figure 53. The two strategies designed to explicitly handle multicollinearity do just this, PCA and the VAE have constraints such that they learn representations that are entirely uncorrelated. PCA achieves this through the generation of orthogonal projections and the VAE by forcing the latent distributions to adhere to a multivariate

Gaussian distribution with a covariance matrix with off-diagonal values of 0. However, within the FreeSurfer statistics, representations of gray and white matter intensity, in addition to surface area and volume are highly correlated. We also see relationships, both positive and negative, among the AE embeddings, again because there is no constraint on the objective function to penalize a correlated latent space. Theoretically, an AE could learn a set of latent representations that were nearly perfectly correlated, so long as reconstruction error was low. It is critical to contextualize that every iteration of *learning* that occurs within the AE is completed only in order to minimize the reconstruction error (MSE).

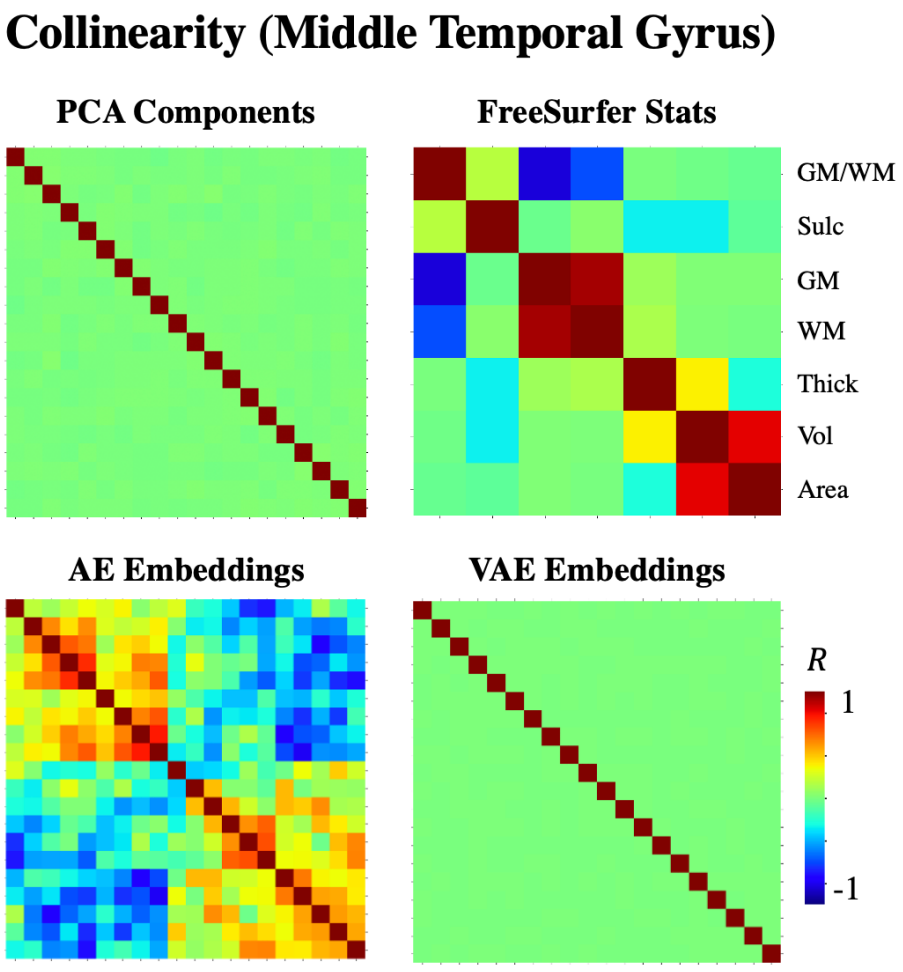


Figure 53: Latent Collinearity

Correlation between features sets for the PCA components, FreeSurfer statistics, AE embeddings, and VAE embeddings. Sulcal depth (sulc), thickness (thick), volume (vol), and area represent aspects of morphology and mean gray matter intensity (GM), mean white matter intensity (WM) and the contrast of the two (GM/WM) reflect image intensity measures.

While Figure 53 above does indeed illustrate that there is redundancy within latent features of the AE, there remains intrinsic utility within these features. It is important to clarify that we are not disparaging their utility but reiterating that they do suffer from attributes that are considered less than ideal from a modeling perspective. Using information from the AE, PCA, and FreeSurfer statistics improves our ability to predict WM within the inferior temporal and middle temporal gyrus. Additionally, AE embeddings within the left hemisphere structures of the supramarginal, superior temporal, middle temporal, and inferior parietal gyri outperform PCA in predicting EXT. Thus, while it is critical to note the potential shortcomings associated with existing multicollinearity among AE embeddings, it is clear that this is but a single piece of their puzzle and not a consistent barrier to improving brain-behavior model performance.

Clinical Utility

Perhaps the most important aspect of this analysis lies within the clinical utility of the EXT predictions, and unfortunately, each of these modeling strategies suffers from the same issues discussed in Chapter 2. Namely, predictions far from the mean of the data have a higher absolute error. This is, unfortunately, an inevitable limitation of predictive tasks using datasets sampling from a normative population not enriched for psychopathology. It is not uncommon for models to struggle in predicting observations furthest from the mean within in the tails of the

distribution (De Backer et al., 2023). Of the sample utilized in these analyses, 631 subjects have EXT t-scores above the clinically diagnostic threshold of 63. An additional 351 of these subjects fall into the sub-clinical threshold of $60 \geq x_i < 63$. Thus, we again point to the potential utility of examining additional strategies such as normative modeling strategies that seek to evaluate large deviations from the sample mean over raw outcome values. However, as discussed in Chapter 2 there are limitations with this approach as well (Marquand et al., 2019).

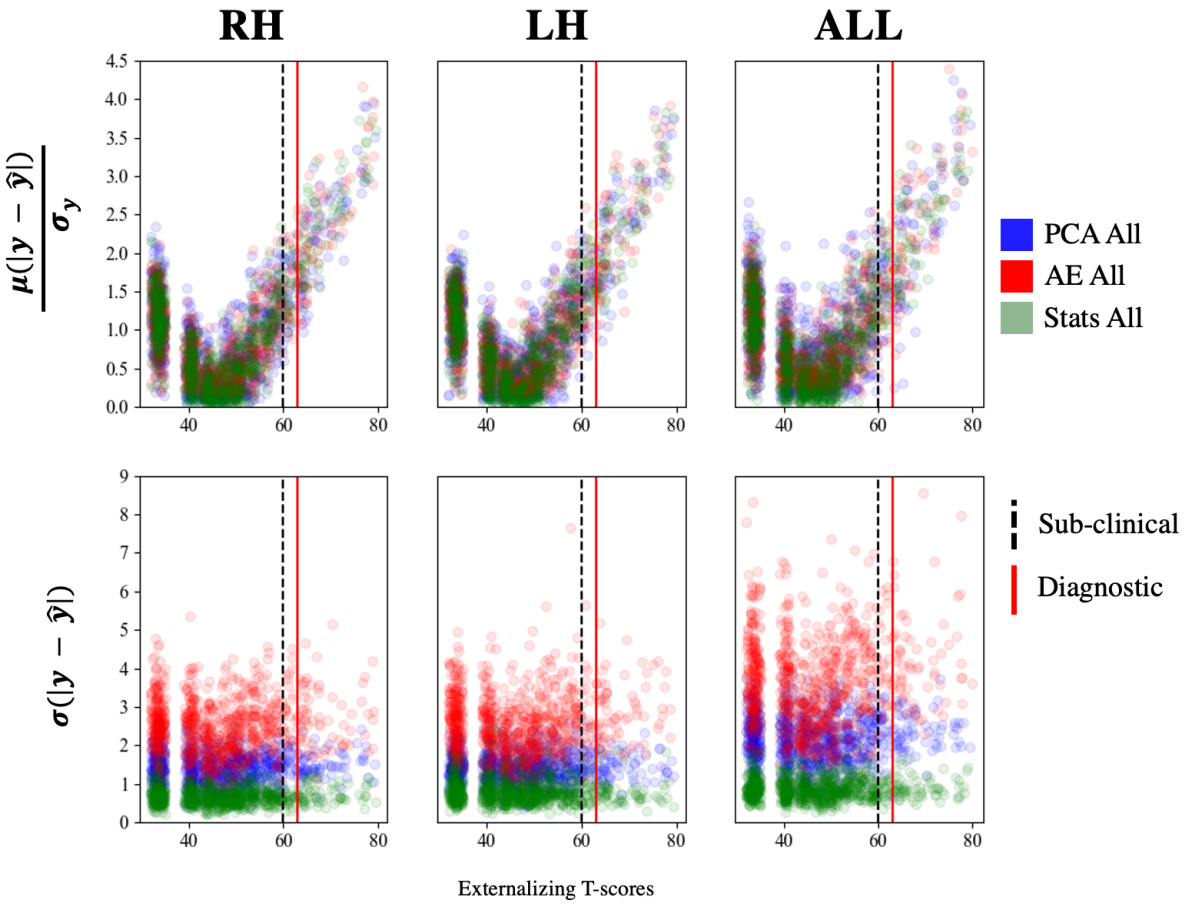


Figure 54: Externalizing T-score Error vs Externalizing T-scores

Top row: Mean absolute value of test prediction error (averaged over the 10 training splits) vs endorsed Externalizing t-scores via three modeling feature sets of PCA all region components, AE all region embeddings, and all FreeSurfer statistics. Red solid line indicates the threshold of score for diagnostic criteria (63) and the dashed black line represents the sub-clinical threshold (60). Bottom row: Variance of the error over the 10 folds of training vs actual t-scores.

Ultimately, we must contextualize the enormity of this problem. Modeling a heterogeneous complex aggregate of various clinical symptoms using data of considerably high-dimensionality that contains noise is monumentally difficult. While there are undoubtedly regionally specific benefits within the imaging data obtained using large complex and computationally expensive modeling strategies, we ultimately obtained the best performance in predicting psychopathology with the FreeSurfer statistics and a linear model. We repeat ad-nauseam, that this is not to say there are not new and alternative DL based modeling strategies, preprocessing streams, and alternative methods that may yield information within the imaging data that may outperform the reliable FreeSurfer summary statistics, however these analyses presented here do not support that theory.

Conclusion and Looking Ahead

It is important to reiterate the original goal of this project. We present a set of analyses and results seeking to evaluate emerging modeling strategies, namely deep learning, to determine if residual information from within the imaging data would improve the ability to predict both elements of executive function and psychopathology. Though we did not see improvements in predicting EXT using deep learning, this analysis provides a critical view into the possible

inclusion of additional information within the neuroimaging data which may improve the prediction of psychopathology. Furthermore, we have illustrated that using either deep learning methods, such as the unsupervised autoencoder (AE) or SML methods (PCA) to extract additional information beyond the FreeSurfer statistics, significantly improves the ability to understand critical cognitive processes during a time of marked neurodevelopment. Additionally, this performance improves upon existing literature modeling WM with the neuroimaging data in this population of 9- and 10-year old children. These results provide insight into the potential utility of a relatively new and advancing branch of machine learning methods. New algorithms, loss functions, and optimization methods emerge constantly with potential to substantially reshape the potential of these deep learning methods. Furthermore, computational advancements over the last 20 years have been, to say lightly, astronomical. While GPU acceleration and highly optimized software packages have greatly improved the capability of computer vision model performance and efficiency, the sheer size of the neuroimaging data is still, perhaps, the most substantial limitation that we encountered within our analyses. Optimizing hundreds of networks, with thousands of hyperparameter combinations, remains a staggering computational task.

Yet, despite these already challenging computational limitations, it must also be repeated that although these analyses utilized data from one of the largest neuroimaging studies in the world, it is still comparatively small given the complexity of both the targeted outcome and model input. As discussed in detail in Chapter 2, state of the art large-scale computer vision models typically are trained using datasets containing millions of samples which contain images that are typically not as large or as noisy as the T1 sMRI data we used here. These analyses provide additional insight into the prospective utility of complex modeling strategies using a complex data type. As

the field moves forward additional analyses may provide further evidence to the necessary exploration and utility of a large body of methods referred to as deep neural networks.

References

2022 National Healthcare Quality and Disparities Report. (2022). *2022 National Healthcare Quality and Disparities Report*. https://www.ncbi.nlm.nih.gov/books/NBK587182/

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., & Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications*, *12*(1), 353. https://doi.org/10.1038/s41467-020-20655-6

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and Related Forms for Assessing Behavioral/Emotional Problems and Competencies. *Pediatrics In Review*, *21*(8), 265–271. https://doi.org/10.1542/PIR.21-8-265

Albajara Sáenz, A., Villemonteix, T., & Massat, I. (2019). Structural and functional neuroimaging in attention-deficit/hyperactivity disorder. *Developmental Medicine and Child Neurology*, *61*(4), 399–405. https://doi.org/10.1111/dmcn.14050

Andreotti, C., Thigpen, J. E., Dunn, M. J., Watson, K., Potts, J., Reising, M. M., Robinson, K. E., Rodriguez, E. M., Roubinov, D., Luecken, L., & Compas, B. E. (2013). Cognitive reappraisal and secondary control coping: associations with working memory, positive and negative affect, and symptoms of anxiety/depression. *Anxiety, Stress, and Coping*, *26*(1), 20–35. https://doi.org/10.1080/10615806.2011.631526

Asperti, A., Evangelista, D., & Loli Piccolomini, E. (2021). A Survey on Variational Autoencoders from a Green AI Perspective. *SN Computer Science*, *2*(4), 1–23. https://doi.org/10.1007/S42979-021-00702-9/FIGURES/20

Asperti, A., & Trentin, M. (2020). Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*, *8*, 199440–199448. https://doi.org/10.1109/access.2020.3034828

Auchter, A. M., Hernandez Mejia, M., Heyser, C. J., Shilling, P. D., Jernigan, T. L., Brown, S. A., Tapert, S. F., & Dowling, G. J. (2018). A description of the ABCD organizational structure and communication framework. *Developmental Cognitive Neuroscience*, *32*, 8. https://doi.org/10.1016/J.DCN.2018.04.003

Aurich, N. K., Filho, J. O. A., da Silva, A. M. M., & Franco, A. R. (2015). Evaluating the reliability of different preprocessing steps to estimate graph theoretical measures in resting state fMRI data. *Frontiers in Neuroscience*, *9*(FEB). https://doi.org/10.3389/FNINS.2015.00048/ABSTRACT

Ayyar, M. P., Benois-Pineau, J., Zemmari, A., & Catheline, G. (2021). Explaining 3D CNNs for Alzheimer's Disease Classification on sMRI Images with Multiple ROIs. *2021 IEEE International Conference on Image Processing (ICIP)*, 284–288. https://doi.org/10.1109/ICIP42928.2021.9506472

Babaeeghazvini, P., Rueda-Delgado, L. M., Gooijers, J., Swinnen, S. P., & Daffertshofer, A. (2021). Brain Structural and Functional Connectivity: A Review of Combined Works of Diffusion Magnetic Resonance Imaging and Electro-Encephalography. *Frontiers in Human Neuroscience*, *15*, 721206. https://doi.org/10.3389/FNHUM.2021.721206/BIBTEX

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *ACM International Conference Proceeding Series*, *382*. https://doi.org/10.1145/1553374.1553380

Bennett, I. J., & Rypma, B. (2013). Advances in functional neuroanatomy: a review of combined DTI and fMRI studies in healthy younger and older adults. *Neuroscience and Biobehavioral Reviews*, *37*(7), 1201–1210. https://doi.org/10.1016/J.NEUBIOREV.2013.04.008

Bento, M., Fantini, I., Park, J., Rittner, L., & Frayne, R. (2022). Deep Learning in Large and Multi-Site Structural Brain MR Imaging Datasets. *Frontiers in Neuroinformatics*, *15*, 805669. https://doi.org/10.3389/FNINF.2021.805669/BIBTEX

Bergstra, J., Ca, J. B., & Ca, Y. B. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research*, *13*, 281–305. http://scikit-learn.sourceforge.net.

Biden-Harris Administration Announces Millions of Dollars in New Funds for States to Tackle Mental Health Crisis | SAMHSA. (n.d.). Retrieved March 15, 2024, from https://www.samhsa.gov/newsroom/press-announcements/20221018/biden-harris-administration-announces-funding-states-tackle-mental-health-crisis

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex (New York, NY)*, *19*(12), 2767. https://doi.org/10.1093/CERCOR/BHP055

Brückl, T. M., Spoormaker, V. I., Sämann, P. G., Brem, A. K., Henco, L., Czamara, D., Elbau, I., Grandi, N. C., Jollans, L., Kühnel, A., Leuchs, L., Pöhlchen, D., Schneider, M., Tontsch, A., Keck, M. E., Schilbach, L., Czisch, M., Lucae, S., Erhardt, A., & Binder, E. B. (2020). The biological classification of mental disorders (BeCOME) study: a protocol for an observational deep-phenotyping study for the identification of biological subtypes. *BMC Psychiatry 2020 20:1*, *20*(1), 1–25. https://doi.org/10.1186/S12888-020-02541-Z

Cai, W., Griffiths, K., Korgaonkar, M. S., Williams, L. M., & Menon, V. (2019). Inhibition-related modulation of salience and frontoparietal networks predicts cognitive control ability and inattention symptoms in children with ADHD. *Molecular Psychiatry 2019 26:8*, *26*(8), 4016–4025. https://doi.org/10.1038/s41380-019-0564-4

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., … Dale, A. M. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54. https://doi.org/10.1016/J.DCN.2018.03.001

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders? *Clinical Psychological Science : A Journal of the Association for Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Caspi, A., & Moffitt, T. E. (2018). All for One and One for All: Mental Disorders in One Dimension. *American Journal of Psychiatry*, *175*(9), 831–844. https://doi.org/10.1176/appi.ajp.2018.17121383

Cervin, M., Norris, L. A., Ginsburg, G., Gosch, E. A., Compton, S. N., Piacentini, J., Albano, A. M., Sakolsky, D., Birmaher, B., Keeton, C., Storch, E. A., & Kendall, P. C. (2021). The *p* Factor Consistently Predicts Long-Term Psychiatric and Functional Outcomes in Anxiety-Disordered Youth. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(7), 902-912.e5. https://doi.org/10.1016/j.jaac.2020.08.440

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., Rossor, A. M., Stevens, J. M., Cipolotti, L., & Rossor, M. N. (2001). Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Annals of Neurology*, *49*(4), 433–442. https://doi.org/10.1002/ANA.92

Chen, J., Tam, A., Kebets, V., Orban, C., Ooi, L. Q. R., Asplund, C. L., Marek, S., Dosenbach, N. U. F., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. T. (2022). Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nature Communications 2022 13:1*, *13*(1), 1–17. https://doi.org/10.1038/s41467-022-29766-8

Clark, D. A., Hicks, B. M., Angstadt, M., Rutherford, S., Taxali, A., Hyde, L., Weigard, A. S., Heitzeg, M. M., & Sripada, C. (2021). The General Factor of Psychopathology in the Adolescent Brain Cognitive Development (ABCD) Study: A Comparison of Alternative Modeling Approaches. *Https://Doi.Org/10.1177/2167702620959317*, *9*(2), 169–182. https://doi.org/10.1177/2167702620959317

Cohen, S. E., Zantvoord, J. B., Wezenberg, B. N., Bockting, C. L. H., & van Wingen, G. A. (2021). Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis. *Translational Psychiatry 2021 11:1*, *11*(1), 1–10. https://doi.org/10.1038/s41398-021-01286-x

Cordova, M., Shada, K., Demeter, D. V, Doyle, O., Miranda-Dominguez, O., Perrone, A., Schifsky, E., Graham, A., Fombonne, E., Langhorst, B., Nigg, J., Fair, D. A., & Feczko, E. (2020). Heterogeneity of executive function revealed by a functional random forest approach across ADHD and ASD. *NeuroImage: Clinical*, *26*, 102245. https://doi.org/10.1016/j.nicl.2020.102245

Crowe, S. F., Matthews, C., & Walkenhorst, E. (2007). Relationship between worry, anxiety and thought suppression and the components of working memory in a non-clinical sample. *Australian Psychologist*, *42*(3), 170–177. https://doi.org/10.1080/00050060601089462

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/NIMG.1998.0395

Dartora, C., Marseglia, A., Mårtensson, G., Rukh, G., Dang, J., Muehlboeck, J. S., Wahlund, L. O., Moreno, R., Barroso, J., Ferreira, D., Schiöth, H. B., & Westman, E. (2023). A deep learning model for brain age prediction using minimally preprocessed T1w images as input. *Frontiers in Aging Neuroscience*, *15*. https://doi.org/10.3389/FNAGI.2023.1303036/FULL

De Backer, M., Legrand, C., Péron, J., Lambert, A., & Buyse, M. (2023). On the use of extreme value tail modeling for generalized pairwise comparisons with censored outcomes. *Pharmaceutical Statistics*, *22*(2), 284–299. https://doi.org/10.1002/PST.2271

De Lissnyder, E., Koster, E. H. W., Goubert, L., Onraedt, T., Vanderhasselt, M.-A., & De Raedt, R. (2012). Cognitive control moderates the association between stress and rumination. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(1), 519–525. https://doi.org/10.1016/j.jbtep.2011.07.004

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/J.NEUROIMAGE.2006.01.021

Dhamala, E., Ooi, L. Q. R., Chen, J., Kong, R., Anderson, K. M., Chin, R., Yeo, B. T. T., & Holmes, A. J. J. (2022). Proportional intracranial volume correction differentially biases behavioral predictions across neuroanatomical features, sexes, and development. *NeuroImage*, *260*, 119485. https://doi.org/10.1016/J.NEUROIMAGE.2022.119485

Dimitriadis, A., Trivizakis, E., Papanikolaou, N., Tsiknakis, M., & Marias, K. (2022). Enhancing cancer differentiation with synthetic MRI examinations via generative models: a systematic review. *Insights into Imaging*, *13*(1), 1–27. https://doi.org/10.1186/S13244-022-01315-3/TABLES/6

Dockès, J., Poldrack, R. A., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., Thirion, B., & Varoquaux, G. (2021). NeuroQuery, comprehensive meta-analysis of human brain mapping. *ELife*, *9*, e53385. https://doi.org/10.7554/eLife.53385

Dosenbach, N. U. F., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., Snyder, A. Z., Nagel, B. J., Nigg, J. T., Nguyen, A. L., Wesevich, V., Greene, D. J., & Fair, D. A. (2017). Real-time motion analytics during brain MRI improve data quality and reduce costs. *NeuroImage*, *161*, 80–93. https://doi.org/10.1016/J.NEUROIMAGE.2017.08.025

Dular, L., Pernuš, F., & Špiclin, Ž. (2023a). Extensive T1-weighted MRI Preprocessing Improves Generalizability of Deep Brain Age Prediction Models. *BioRxiv*. https://doi.org/10.1101/2023.05.10.540134

Dular, L., Pernuš, F., & Špiclin, Ž. (2023b). Extensive T1-weighted MRI Preprocessing Improves Generalizability of Deep Brain Age Prediction Models. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.05.10.540134

Eisenberg, N., Valiente, C., Spinrad, T. L., Liew, J., Zhou, Q., Losoya, S. H., Reiser, M., & Cumberland, A. (2009). Longitudinal relations of children's effortful control, impulsivity, and negative emotionality to their externalizing, internalizing, and co-occurring behavior problems. *Developmental Psychology*, *45*(4), 988–1008. https://doi.org/10.1037/a0016213

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/FRAI.2020.00004

Espy, K. A., Sheffield, T. D., Wiebe, S. A., Clark, C. A. C., & Moehr, M. (2011a). Executive Control and Dimensions of Problem Behaviors in Preschool Children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *52*(1), 33–46. https://doi.org/10.1111/j.1469-7610.2010.02265.x

Espy, K. A., Sheffield, T. D., Wiebe, S. A., Clark, C. A. C., & Moehr, M. (2011b). Executive Control and Dimensions of Problem Behaviors in Preschool Children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *52*(1), 33–46. https://doi.org/10.1111/j.1469-7610.2010.02265.x

Faghiri, A., Stephen, J. M., Wang, Y. P., Wilson, T. W., & Calhoun, V. D. (2019). Brain Development Includes Linear and Multiple Nonlinear Trajectories: A Cross-Sectional Resting-State Functional Magnetic Resonance Imaging Study. *Brain Connectivity*, *9*(10), 777. https://doi.org/10.1089/BRAIN.2018.0641

Fair, D. A., Miranda-Dominguez, O., Snyder, A. Z., Perrone, A., Earl, E. A., Van, A. N., Koller, J. M., Feczko, E., Tisdall, M. D., van der Kouwe, A., Klein, R. L., Mirro, A. E., Hampton, J. M., Adeyemo, B., Laumann, T. O., Gratton, C., Greene, D. J., Schlaggar, B. L., Hagler, D. J., … Dosenbach, N. U. F. (2020). Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage*, *208*, 116400. https://doi.org/10.1016/J.NEUROIMAGE.2019.116400

Farzi, S., Kianian, S., & Rastkhadive, I. (2017). Diagnosis of attention deficit hyperactivity disorder using deep belief network based on greedy approach. *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, 96–99. https://doi.org/10.1109/ISCBI.2017.8053552

Feczko, E., Conan, G., Marek, S., Tervo-Clemmens, B., Cordova, M., Doyle, O., Earl, E., Perrone, A., Sturgeon, D., Klein, R., Harman, G., Kilamovich, D., Hermosillo, R., Miranda-Dominguez, O., Adebimpe, A., Bertolero, M., Cieslak, M., Covitz, S., Hendrickson, T., … Fair, D. A. (2021). *Adolescent Brain Cognitive Development (ABCD) Community MRI Collection and Utilities*. Neuroscience. http://biorxiv.org/lookup/doi/10.1101/2021.07.09.451638

Filippi, M., Agosta, F., Barkhof, F., Dubois, B., Fox, N. C., Frisoni, G. B., Jack, C. R., Johannsen, P., Miller, B. L., Nestor, P. J., Scheltens, P., Sorbi, S., Teipel, S., Thompson, P. M., & Wahlund, L. O. (2012). EFNS task force: the use of neuroimaging in the diagnosis of dementia. *European Journal of Neurology*, *19*(12), 1487–1501. https://doi.org/10.1111/J.1468-1331.2012.03859.X

Flores, R. E. U., Sánchez, R. D., Peña, F. R. de la, Sciutto, M. F. R., Cruz, L. P., & Villa, P. M. (2022). Executive Functioning in Children and Adolescents with ADHD and Disruptive

Behavior Disorders. *Innovations in Clinical Neuroscience*, *19*(10–12), 16. /pmc/articles/PMC9776777/

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/J.NEUROIMAGE.2010.07.033

Fraza, C., Zabihi, M., Beckmann, C. F., & Marquand, A. F. (2022). The Extremes of Normative Modelling. *BioRxiv*, 2022.08.23.505049. https://doi.org/10.1101/2022.08.23.505049

Friedman, N. P., & Robbins, T. W. (2021). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology 2021 47:1*, *47*(1), 72–89. https://doi.org/10.1038/s41386-021-01132-0

Fur, N., Garrido, L., Dolan, R. J., Driver, J., & Duchaine, B. (2011). Fusiform Gyrus Face Selectivity Relates to Individual Differences in Facial Recognition Ability. *Journal of Cognitive Neuroscience*, *23*(7), 1723–1740. https://doi.org/10.1162/JOCN.2010.21545

Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R. Z., Heeringa, S., Jernigan, T., Potter, A., Thompson, W., & Zahs, D. (2018). Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience*, *32*, 16–22. https://doi.org/10.1016/j.dcn.2018.04.004

Garcia, M., Dosenbach, N., & Kelly, C. (2023). BrainQCNet: a Deep Learning attention-based model for the automated detection of artifacts in brain structural MRI scans. *BioRxiv*, 2022.03.11.483983. https://doi.org/10.1101/2022.03.11.483983

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013).

The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/J.NEUROIMAGE.2013.04.127

Goceri, E. (2023). Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, *56*(11), 1. https://doi.org/10.1007/S10462-023-10453-Z

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., & Petersen, S. E. (2016). Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral Cortex (New York, NY)*, *26*(1), 288. https://doi.org/10.1093/CERCOR/BHU239

Greene, A. S., Shen, X., Noble, S., Horien, C., Hahn, C. A., Arora, J., Tokoglu, F., Spann, M. N., Carrión, C. I., Barron, D. S., Sanacora, G., Srihari, V. H., Woods, S. W., Scheinost, D., & Constable, R. T. (2022). Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature 2022 609:7925*, *609*(7925), 109–118. https://doi.org/10.1038/s41586-022-05118-w

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., & Reuter, M. (2020). FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, *219*, 117012. https://doi.org/10.1016/J.NEUROIMAGE.2020.117012

Honey, C. J., Thivierge, J. P., & Sporns, O. (2010). Can structure predict function in the human brain? *NeuroImage*, *52*(3), 766–776. https://doi.org/10.1016/J.NEUROIMAGE.2010.01.071

Hu, F., Chen, A. A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A., Li, M., Shou, H., Satterthwaite, T. D., Yu, M., & Shinohara, R. T. (2023). Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation

metrics for effective harmonization. *NeuroImage*, *274*, 120125. https://doi.org/10.1016/J.NEUROIMAGE.2023.120125

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot Ensembles: Train 1, get M for free. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. https://arxiv.org/abs/1704.00109v1

Huseyn, E. (n.d.). *Deep Learning Based Early Diagnostics of Parkinson's Disease*. 14. https://arxiv.org/ftp/arxiv/papers/2008/2008.01792.pdf

Hyatt, C. S., Owens, M. M., Crowe, M. L., Carter, N. T., Lynam, D. R., & Miller, J. D. (2020). The quandary of covarying: A brief review and empirical examination of covariate use in structural neuroimaging studies on psychological variables. *NeuroImage*, *205*, 116225. https://doi.org/10.1016/J.NEUROIMAGE.2019.116225

Jiang, S., Huang, H., Zhou, J., Li, H., Duan, M., Yao, D., & Luo, C. (2023). Progressive trajectories of schizophrenia across symptoms, genes, and the brain. *BMC Medicine*, *21*(1), 1–16. https://doi.org/10.1186/S12916-023-02935-2/FIGURES/5

Kang, J., Caffo, B., & Liu, H. (2016). Editorial: Recent Advances and Challenges on Big Data Analysis in Neuroimaging. *Frontiers in Neuroscience*, *10*, 505. https://doi.org/10.3389/fnins.2016.00505

Kaur, S., Aggarwal, H., & Rani, R. (2021). Diagnosis of Parkinson's disease using deep CNN with transfer learning and data augmentation. *Multimedia Tools and Applications*, *80*(7), 10113–10139. https://doi.org/10.1007/s11042-020-10114-1

Kebaili, A., Lapuyade-Lahorgue, J., & Ruan, S. (2023). Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. *Journal of Imaging*, *9*(4). https://doi.org/10.3390/JIMAGING9040081

Keshavan, A., Yeatman, J. D., & Rokem, A. (2019). Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. *Frontiers in Neuroinformatics*, *13*. https://doi.org/10.3389/FNINF.2019.00029

Kim, J., Calhoun, V. D., Shim, E., & Lee, J. H. (2016). 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. *NeuroImage*, *124*(0 0), 127. https://doi.org/10.1016/J.NEUROIMAGE.2015.05.018

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558–569. https://doi.org/10.4097/KJA.19087

Kjelkenes, R., Wolfers, T., Alnæs, D., Norbom, L. B., Voldsbekk, I., Holm, M., Dahl, A., Berthet, P., Tamnes, C. K., Marquand, A. F., & Westlye, L. T. (2022). Deviations from normative brain white and gray matter structure are associated with psychopathology in youth. *Developmental Cognitive Neuroscience*, *58*, 101173. https://doi.org/10.1016/J.DCN.2022.101173

Kliamovich, D., Jones, S. A., Chiapuzio, A. M., Baker, F. C., Clark, D. B., & Nagel, B. J. (2021). Sex-specific patterns of white matter microstructure are associated with emerging depression during adolescence. *Psychiatry Research. Neuroimaging*, *315*, 111324. https://doi.org/10.1016/J.PSCYCHRESNS.2021.111324

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (n.d.). *ImageNet Classification with Deep Convolutional Neural Networks*. Retrieved March 19, 2024, from http://code.google.com/p/cuda-convnet/

Kuang, D., & He, L. (2014). Classification on ADHD with Deep Learning. *2014 International Conference on Cloud Computing and Big Data*, 27–32. https://doi.org/10.1109/CCBD.2014.42

Kushol, R., Parnianpour, P., Wilman, A. H., Kalra, S., & Yang, Y. H. (2023). Effects of MRI scanner manufacturers in classification tasks with deep learning models. *Scientific Reports 2023 13:1*, *13*(1), 1–13. https://doi.org/10.1038/s41598-023-43715-5

Lee, J., Burkett, B., Min, H.-K., Senjem, M., Lundt, E., Botha, H., Graff-Radford, J., Barnard, L., Gunter, J., Schwarz, C., Kantarci, K., Knopman, D., Boeve, B., Lowe, V., Petersen, R., Jack, C., & Jones, D. (2021). *Deep learning-based brain age prediction in normal aging and dementia*. In Review. https://www.researchsquare.com/article/rs-804454/v1

Leff, A. P., Schofield, T. M., Crinion, J. T., Seghier, M. L., Grogan, A., Green, D. W., & Price, C. J. (2009). The left superior temporal gyrus is a shared substrate for auditory short-term memory and speech comprehension: evidence from 210 patients with stroke. *Brain*, *132*(12), 3401. https://doi.org/10.1093/BRAIN/AWP273

Leming, M., & Suckling, J. (2021). Deep learning for sex classification in resting-state and task functional brain networks from the UK Biobank. *NeuroImage*, *241*, 118409. https://doi.org/10.1016/j.neuroimage.2021.118409

Li, A., Li, H., & Yuan, G. (2024). Continual Learning with Deep Neural Networks in Physiological Signal Data: A Survey. *Healthcare*, *12*(2). https://doi.org/10.3390/HEALTHCARE12020155

Li, M., Jiang, M., Zhang, G., Liu, Y., & Zhou, X. (2022). Prediction of fluid intelligence from T1-w MRI images: A precise two-step deep learning framework. *PLOS ONE*, *17*(8), e0268707. https://doi.org/10.1371/JOURNAL.PONE.0268707

Lian, C., Liu, M., Zhang, J., & Shen, D. (2020). Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 880–893. https://doi.org/10.1109/TPAMI.2018.2889096

Liu, X., Li, L., Li, M., Ren, Z., & Ma, P. (2021). Characterizing the subtype of anhedonia in major depressive disorder: A symptom-specific multimodal MRI study. *Psychiatry Research. Neuroimaging*, *308*. https://doi.org/10.1016/J.PSCYCHRESNS.2020.111239

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., Weiner, M., Aisen, P., Petersen, R., Jack, C., Jagust, W., Trojanowki, J., Toga, A., Beckett, L., Green, R., Saykin, A., Morris, J. C., Shaw, L., Kaye, J., Quinn, J., … Fargher, K. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Scientific Reports 2018 8:1*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-22871-z

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290. https://doi.org/10.1038/NMETH.4627

Magyar, C. I., & Pandolfi, V. (2017). Utility of the CBCL DSM-Oriented Scales in Assessing Emotional Disorders in Youth with Autism. *Research in Autism Spectrum Disorders*, *37*, 11–20. https://doi.org/10.1016/j.rasd.2017.01.009

Malhi, G. S., & Lagopoulos, J. (2008). Making sense of neuroimaging in psychiatry. *Acta Psychiatrica Scandinavica*, *117*(2), 100–117. https://doi.org/10.1111/J.1600-0447.2007.01111.X

Malla, A., Shah, J., Iyer, S., Boksa, P., Joober, R., Andersson, N., Lal, S., & Fuhrer, R. (2018). Youth Mental Health Should Be a Top Priority for Health Care in Canada. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *63*(4), 216–222. https://doi.org/10.1177/0706743718758968

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature 2022 603:7902*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Marklund, P., & Nyberg, L. (2007). Intersecting the divide between working memory and episodic memory: Evidence from sustained and transient brain activity patterns. *The Cognitive Neuroscience of Working Memory*. https://doi.org/10.1093/ACPROF:OSO/9780198570394.003.0018

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., & Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, *24*(10), 1415–1424. https://doi.org/10.1038/s41380-019-0441-1

Martel, M. M., & Nigg, J. T. (2006). Child ADHD and personality/temperament traits of reactive and effortful control, resiliency, and emotionality. *Journal of Child Psychology and*

Psychiatry, and Allied Disciplines, *47*(11), 1175–1183. https://doi.org/10.1111/j.1469-7610.2006.01629.x

Mathias, S. R., Knowles, E. E. M., Mollon, J., Rodrigue, A., Koenis, M. M. C., Alexander-Bloch, A. F., Winkler, A. M., Olvera, R. L., Duggirala, R., Göring, H. H. H., Curran, J. E., Fox, P. T., Almasy, L., Blangero, J., & Glahn, D. C. (2020). Minimal Relationship between Local Gyrification and General Cognitive Ability in Humans. *Cerebral Cortex (New York, NY)*, *30*(6), 3439. https://doi.org/10.1093/CERCOR/BHZ319

Matsubara, T., Tashiro, T., & Uehara, K. (2019). Deep Neural Generative Model of Functional MRI Images for Psychiatric Disorder Diagnosis. *IEEE Transactions on Bio-Medical Engineering*, *66*(10), 2768–2779. https://doi.org/10.1109/TBME.2019.2895663

McGrath, J. J., Lim, C. C. W., Plana-Ripoll, O., Holtz, Y., Agerbo, E., Momen, N. C., Mortensen, P. B., Pedersen, C. B., Abdulmalik, J., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Bromet, E. J., Bruffaerts, R., Bunting, B., de Almeida, J. M. C., de Girolamo, G., De Vries, Y. A., Florescu, S., … de Jonge, P. (2020). Comorbidity within mental disorders: a comprehensive analysis based on 145 990 survey respondents from 27 countries. *Epidemiology and Psychiatric Sciences*, *29*, e153. https://doi.org/10.1017/S2045796020000633

Mihalik, A., Brudfors, M., Robu, M., Ferreira, F. S., Lin, H., Rau, A., Wu, T., Blumberg, S. B., Kanber, B., Tariq, M., Garcia, M. E., Zor, C., Nikitichev, D. I., Mourão-Miranda, J., & Oxtoby, N. P. (2019). ABCD Neurocognitive Prediction Challenge 2019: Predicting individual fluid intelligence scores from structural MRI using probabilistic segmentation and kernel ridge regression. *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *11791 LNCS*, 133–142. https://doi.org/10.1007/978-3-030-31901-4_16

Mozhdehfarahbakhsh, A., Chitsazian, S., Chakrabarti, P., Chakrabarti, T., Kateb, B., & Nami, M. (2021). *An MRI-based Deep Learning Model to Predict Parkinson's Disease Stages*. https://www.medrxiv.org/content/10.1101/2021.02.19.21252081v1

Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., & Castiglioni, I. (2020). Comparison of Transfer Learning and Conventional Machine Learning Applied to Structural Brain MRI for the Early Diagnosis and Prognosis of Alzheimer's Disease. *Frontiers in Neurology*, *11*, 576194. https://doi.org/10.3389/fneur.2020.576194

Narang, S., Diamos, G., Elsen, E., Micikevicius, P., Alben, J., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2017). Mixed Precision Training. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. https://arxiv.org/abs/1710.03740v3

Neelakantan, A., Vilnis, L., Le, Q. V, Sutskever, I., Kaiser, L., Kurach, K., Brain, G., & Martens, J. (2015). *Adding Gradient Noise Improves Learning for Very Deep Networks*. https://arxiv.org/abs/1511.06807v1

Nigg, J. T. (2017a). Annual Research Review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(4), 361–383. https://doi.org/10.1111/jcpp.12675

Nigg, J. T. (2017b). Annual Research Review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking,

and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(4), 361–383. https://doi.org/10.1111/jcpp.12675

Nygaard, V., Rødland, E. A., & Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, *17*(1), 29–39. https://doi.org/10.1093/BIOSTATISTICS/KXV027

Osuch, E., & Williamson, P. (2006). Brain imaging in psychiatry: from a technique of exclusion to a technique for diagnosis. *Acta Psychiatrica Scandinavica*, *114*(2), 73–74. https://doi.org/10.1111/J.1600-0447.2006.00860.X

Oxtoby, N. P., Ferreira, F. S., Mihalik, A., Wu, T., Brudfors, M., Lin, H., Rau, A., Blumberg, S. B., Robu, M., Zor, C., Tariq, M., Garcia, M. E., Kanber, B., Nikitichev, D. I., & Mourão-Miranda, J. (2019). ABCD Neurocognitive Prediction Challenge 2019: Predicting individual residual fluid intelligence scores from cortical grey matter morphology. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11791 LNCS*, 114–123. https://doi.org/10.1007/978-3-030-31901-4_14

Parkes, L., Moore, T. M., Calkins, M. E., Cook, P. A., Cieslak, M., Roalf, D. R., Wolf, D. H., Gur, R. C., Gur, R. E., Satterthwaite, T. D., & Bassett, D. S. (2021). Transdiagnostic dimensions of psychopathology explain individuals' unique deviations from normative neurodevelopment in brain structure. *Translational Psychiatry 2021 11:1*, *11*(1), 1–13. https://doi.org/10.1038/s41398-021-01342-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in

Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://arxiv.org/abs/1201.0490v4

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., & Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, *68*, 101871. https://doi.org/10.1016/j.media.2020.101871

Pinaya, W. H. L., Gadelha, A., Doyle, O. M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A. P., Bressan, R. A., & Sato, J. R. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports*, *6*, 38897. https://doi.org/10.1038/srep38897

Pinaya, W. H. L., Mechelli, A., & Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human Brain Mapping*, *40*(3), 944. https://doi.org/10.1002/HBM.24423

Rao, A., Monteiro, J. M., & Mourao-Miranda, J. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage*, *150*, 23. https://doi.org/10.1016/J.NEUROIMAGE.2017.01.066

Ray, A., Bhardwaj, A., Malik, Y. K., Singh, S., & Gupta, R. (2022). Artificial intelligence and Psychiatry: An overview. *Asian Journal of Psychiatry*, *70*, 103021. https://doi.org/10.1016/J.AJP.2022.103021

Riglin, L., Leppert, B., Dardani, C., Thapar, A. K., Rice, F., O'Donovan, M. C., Davey Smith, G., Stergiakouli, E., Tilling, K., & Thapar, A. (2021). ADHD and depression: investigating a causal explanation. *Psychological Medicine*, *51*(11), 1890–1897. https://doi.org/10.1017/S0033291720000665

Romer, A. L., & Pizzagalli, D. A. (2021). Is executive dysfunction a risk marker or consequence of psychopathology? A test of executive function as a prospective predictor and outcome of general psychopathology in the adolescent brain cognitive development study®. *Developmental Cognitive Neuroscience*, *51*. https://doi.org/10.1016/J.DCN.2021.100994

Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., Docherty, A., Simms, L. J., Bagby, R. M., Krueger, R. F., Callahan, J., Chmielewski, M., Conway, C. C., De Clercq, B. J., Dornbach-Bender, A., … Zimmermann, J. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into Clinical Practice. *Journal of Consulting and Clinical Psychology*, *87*(12), 1069–1084. https://doi.org/10.1037/ccp0000452

Samani, Z. R., Alappatt, J. A., Parker, D., Ismail, A. A. O., & Verma, R. (2019). QC-Automator: Deep Learning-based Automated Quality Control for Diffusion MR Images. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.01456

Samani, Z. R., Alappatt, J. A., Parker, D., Ismail, A. A. O., & Verma, R. (2020). QC-Automator: Deep Learning-Based Automated Quality Control for Diffusion MR Images. *Frontiers in Neuroscience*, *13*, 469388. https://doi.org/10.3389/FNINS.2019.01456/BIBTEX

Saragosa-Harris, N. M., Chaku, N., MacSweeney, N., Guazzelli Williamson, V., Scheuplein, M., Feola, B., Cardenas-Iniguez, C., Demir-Lira, E., McNeilly, E. A., Huffman, L. G., Whitmore, L., Michalska, K. J., Damme, K. S., Rakesh, D., & Mills, K. L. (2022). A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. *Developmental Cognitive Neuroscience*, *55*. https://doi.org/10.1016/J.DCN.2022.101115

Scheiner, C., Grashoff, J., Kleindienst, N., & Buerger, A. (2022). Mental disorders at the beginning of adolescence: Prevalence estimates in a sample aged 11-14 years. *Public Health in Practice*, *4*, 100348. https://doi.org/10.1016/j.puhip.2022.100348

Seitzman, B. A., Gratton, C., Marek, S., Raut, R. V., Dosenbach, N. U. F., Schlaggar, B. L., Petersen, S. E., & Greene, D. J. (2020). A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. *NeuroImage*, *206*, 116290. https://doi.org/10.1016/J.NEUROIMAGE.2019.116290

Sen, B., Borle, N. C., Greiner, R., & Brown, M. R. G. (2018). A general prediction model for the detection of ADHD and Autism using structural and functional MRI. *PLOS ONE*, *13*(4), e0194856. https://doi.org/10.1371/journal.pone.0194856

Shahrokhi, H., Tehrani-Doost, M., Shahrivar, Z., Farhang, S., Amiri, S., Shahrokhi, H., Tehrani-Doost, M., Shahrivar, Z., Farhang, S., & Amiri, S. (2017). Deficits of Executive Functioning in Conduct Disorder and Attention Deficit/Hyperactivity Disorder. *Ann Psychiatry Treatm*, *2*(1), 13–20. https://doi.org/10.17352/apt.000006

Shakil, S., Lee, C. H., & Keilholz, S. D. (2016). Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *NeuroImage*, *133*, 111–128. https://doi.org/10.1016/J.NEUROIMAGE.2016.02.074

Sheikhan, N. Y., Henderson, J. L., Halsall, T., Daley, M., Brownell, S., Shah, J., Iyer, S. N., & Hawke, L. D. (2023). Stigma as a barrier to early intervention among youth seeking mental health services in Ontario, Canada: a qualitative study. *BMC Health Services Research*, *23*(1), 1–12. https://doi.org/10.1186/S12913-023-09075-6/TABLES/1

Smith, S. M., & Nichols, T. E. (2018). Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*, *97*(2), 263–268. https://doi.org/10.1016/j.neuron.2017.12.018

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*(MAR). https://doi.org/10.3389/FPSYG.2015.00328

Sorter, M., Stark, L. J., Glauser, T., McClure, J., Pestian, J., Junger, K., & Cheng, T. L. (2023). Addressing the Pediatric Mental Health Crisis: Moving from a Reactive to a Proactive System of Care. *The Journal of Pediatrics*, 113479. https://doi.org/10.1016/j.jpeds.2023.113479

Sprooten, E., Franke, B., & Greven, C. U. (2021). The P-factor and its genomic and neural equivalents: an integrated perspective. *Molecular Psychiatry 2021 27:1*, *27*(1), 38–48. https://doi.org/10.1038/s41380-021-01031-2

Sripada, C., Angstadt, M., Taxali, A., Kessler, D., Greathouse, T., Rutherford, S., Clark, D. A., Hyde, L. W., Weigard, A., Brislin, S. J., Hicks, B., & Heitzeg, M. (2021). Widespread attenuating changes in brain connectivity associated with the general factor of psychopathology in 9- and 10-year olds. *Translational Psychiatry*, *11*(1). https://doi.org/10.1038/S41398-021-01708-W

Strauman, T. J. (2017). Self-Regulation and Psychopathology: Toward an Integrative Translational Research Paradigm. *Annual Review of Clinical Psychology*, *13*, 497–523. https://doi.org/10.1146/ANNUREV-CLINPSY-032816-045012

Taylor, A. J., Salerno, M., Dharmakumar, R., & Jerosch-Herold, M. (2016). T1 Mapping: Basic Techniques and Clinical Applications. *JACC. Cardiovascular Imaging*, *9*(1), 67–81. https://doi.org/10.1016/J.JCMG.2015.11.005

Trombello, J. M., Pizzagalli, D. A., Weissman, M. M., Grannemann, B. D., Cooper, C. M., Greer, T. L., Malchow, A. L., Jha, M. K., Carmody, T. J., Kurian, B. T., Webb, C. A.,

Dillon, D. G., McGrath, P. J., Bruder, G., Fava, M., Parsey, R. V., McInnis, M. G., Adams, P., & Trivedi, M. H. (2018). Characterizing anxiety subtypes and the relationship to behavioral phenotyping in major depression: Results from the EMBARC Study. *Journal of Psychiatric Research*, *102*, 207. https://doi.org/10.1016/J.JPSYCHIRES.2018.04.003

Ulloa, A. E., Plis, S., Erhardt, E., & Calhoun, V. (2015). Synthetic structural magnetic resonance image generator improves deep learning prediction of schizophrenia. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. https://doi.org/10.1109/MLSP.2015.7324379

Van Dijk, K. R. A., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., & Buckner, R. L. (2010). Intrinsic Functional Connectivity As a Tool For Human Connectomics: Theory, Properties, and Optimization. *Journal of Neurophysiology*, *103*(1), 297. https://doi.org/10.1152/JN.00783.2009

Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports 2021 11:1*, *11*(1), 1–13. https://doi.org/10.1038/s41598-020-74399-w

Villatoro, A. P., DuPont-Reyes, M. J., Phelan, J. C., & Link, B. G. (2022). 'Me' vs. 'Them': How Mental Illness Stigma Influences Adolescent Help-Seeking Behaviors for Oneself and Recommendations for Peers. *Stigma and Health*, *7*(3), 300. https://doi.org/10.1037/SAH0000392

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in

Python. *Nature Methods 2020 17:3*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., Conway, K., Deeds, B. G., Dowling, G. J., Grant, S., Howlett, K. D., Matochik, J. A., Morgan, G. D., Murray, M. M., Noronha, A., Spong, C. Y., … Weiss, S. R. B. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4. https://doi.org/10.1016/J.DCN.2017.10.002

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., … Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl 3), S54–S64. https://doi.org/10.1212/WNL.0b013e3182872ded

Whitmer, A. J., & Banich, M. T. (2007). Inhibition versus switching deficits in different forms of rumination. *Psychological Science*, *18*(6), 546–553. https://doi.org/10.1111/j.1467-9280.2007.01936.x

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*(1), 37–48. https://doi.org/10.1016/J.JML.2013.12.003

Xiao, X., Hammond, C., Salmeron, B. J., Wang, D., Gu, H., Zhai, T., Nguyen, H., Lu, H., Ross, T. J., & Yang, Y. (2023). Brain Functional Connectome Defines a Transdiagnostic Dimension Shared by Cognitive Function and Psychopathology in Preadolescents. *Biological Psychiatry*, *0*(0). https://doi.org/10.1016/j.biopsych.2023.08.028

Xu, B., Dall'Aglio, L., Flournoy, J., Bortsova, G., Tervo-Clemmens, B., Collins, P., de Bruijne, M., Luciana, M., Marquand, A., Wang, H., Tiemeier, H., & Muetzel, R. L. (2024). Limited generalizability of multivariate brain-based dimensions of child psychiatric symptoms. *Communications Psychology 2024 2:1*, *2*(1), 1–14. https://doi.org/10.1038/s44271-024-00063-y

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Yin, X., Li, Y., Zhang, X., & Shin, B. S. (2019). Medical Image Augmentation Using Image Synthesis with Contextual Function. *Proceedings - 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2019*. https://doi.org/10.1109/CISP-BMEI48845.2019.8965817

Yu, G., Liu, Z., Wu, X., Becker, B., Zhang, K., Fan, H., Peng, S., Kuang, N., Kang, J., Dong, G., Zhao, X. M., Schumann, G., Feng, J., Sahakian, B. J., Robbins, T. W., Palaniyappan, L., & Zhang, J. (2023). Common and disorder-specific cortical thickness alterations in internalizing, externalizing and thought disorders during early adolescence: an Adolescent Brain and Cognitive Development study. *Journal of Psychiatry and Neuroscience*, *48*(5), E345–E356. https://doi.org/10.1503/JPN.220202/TAB-RELATED-CONTENT

Zetsche, U., D'Avanzato, C., & Joormann, J. (2012). Depression and rumination: relation to components of inhibition. *Cognition & Emotion*, *26*(4), 758–767. https://doi.org/10.1080/02699931.2011.613919

Zhao, J., Huang, J., Zhi, D., Yan, W., Ma, X., Yang, X., Li, X., Ke, Q., Jiang, T., Calhoun, V. D., & Sui, J. (2020). Functional network connectivity (FNC)-based generative adversarial

network (GAN) and its applications in classification of mental disorders. *Journal of Neuroscience Methods*, *341*, 108756. https://doi.org/10.1016/J.JNEUMETH.2020.108756

Zindler, T., Frieling, H., Neyazi, A., Bleich, S., & Friedel, E. (2020). Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/S12859-020-03559-6

Zou, L., Zheng, J., Miao, C., Mckeown, M. J., & Wang, Z. J. (2017). 3D CNN Based Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Functional and Structural MRI. *IEEE Access*, *5*, 23626–23636. https://doi.org/10.1109/ACCESS.2017.2762703