

Université de Montréal

Multi-Task Learning for Joint Diagnosis of CNVs and Psychiatric Conditions from rs-fMRI

Par

Annabelle Harvey

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présentée à la Faculté des arts et des sciences en vue de l'obtention du grade de Maître
ès sciences (M. Sc.) en informatique

Avril 2023

© Annabelle Harvey, 2023

Université de Montréal

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Ce mémoire intitulé:

Multi-Task Learning for Joint Diagnosis of CNVs and Psychiatric Conditions from rs-fMRI

présenté par:

Annabelle Harvey

a été évalué par un jury composé des personnes suivantes:

Aaron Courville

Président

Guy Wolf

Membre

Pierre-Louis Bellec

Directeur

Guillaume Dumas

Co-directeur

Résumé

L'imagerie par résonance magnétique fonctionnelle à l'état de repos (IRMf-R) s'est imposée comme une technologie diagnostique prometteuse. Toutefois, l'application dans la pratique clinique des biomarqueurs de l'IRMf-R visant à capturer les mécanismes biologiques sous-jacents aux troubles psychiatriques a été entravée par le manque de généralisation. Le diagnostic de ces troubles repose entièrement sur des évaluations comportementales et les taux élevés de comorbidités et de chevauchement génétique et symptomatique confirment l'existence de facteurs latents communs à toutes les pathologies. De grandes mutations génétiques rares, appelées variants du nombre de copies (CNV), ont été associées à une série de troubles psychiatriques et ont des effets beaucoup plus importants sur la structure et la fonction du cerveau, ce qui en fait une voie prometteuse pour démêler la génétique des catégories diagnostiques actuelles. L'apprentissage multitâche est une approche prometteuse pour extraire des représentations communes à des tâches connexes, qui permet de mieux utiliser les données en tirant parti des informations partagées et en améliorant la généralisabilité. Nous avons recueilli un ensemble de données sans précédent composé de 19 CNV et de troubles psychiatriques et nous avons cherché à évaluer systématiquement les avantages potentiels de l'apprentissage multitâche pour la précision de la prédiction, afin d'effectuer un diagnostic conjoint de ces conditions interdépendantes. Nous avons estimé les tailles d'effet pour chaque condition, comparé la précision du diagnostic en utilisant des méthodes courantes d'apprentissage automatique, puis en utilisant l'apprentissage multitâches. Nous avons tenté de contrôler les multiples facteurs confondants tout au long des analyses et discutons des différentes approches permettant de le faire dans le contexte de la modélisation prédictive. L'hypothèse selon laquelle les facteurs latents partagés entre les CNV et les troubles psychiatriques les rendraient suffisamment liés en tant que tâches de prédiction pour bénéficier d'un apprentissage conjoint n'a pas été confirmée. Cependant, nous avons également appliqué l'apprentissage multitâche entre les sites pour prédire une cible commune et nous avons montré que la prédiction peut être améliorée lorsque les tâches sont très étroitement liées. Nous avons mis en œuvre un modèle léger de partage des paramètres durs, mais nos résultats et la littérature montrent que ce cadre n'est pas bien adapté aux tâches hétérogènes ou, de manière contre-intuitive, aux échantillons de petite taille. Nous pensons qu'il est possible d'exploiter les similitudes entre les CNV et les troubles psychiatriques en utilisant des méthodes qui modélisent les relations entre les tâches, mais la petite taille des échantillons pour les CNV rares constitue une limitation majeure pour l'application de l'apprentissage multitâche.

Mots clés: Apprentissage multitâche, IRMf, CNVs, troubles psychiatriques, confonds

Abstract

Resting state functional magnetic resonance imaging (rs-fMRI) has emerged as a promising diagnostic technology, however translation into clinical practice of rs-fMRI biomarkers that aim to capture the biological mechanisms underlying psychiatric disorders has been hindered by lack of generalizability. The diagnosis of these disorders is completely based on behavioural assessments and high rates of comorbidities and genetic and symptom overlap supports the existence of latent factors shared across conditions. Rare large genetic mutations, called copy number variants (CNVs), have been associated with a range of psychiatric conditions and have much larger effect sizes on brain structure and function, which makes them a promising avenue for untangling the genetics of the current diagnostic categories. Multi-task learning is a promising approach to extract common representations across related tasks that makes better use of data by leveraging shared information and improves generalizability. We collected an unprecedented dataset consisting of 19 CNVs and psychiatric disorders and aimed to systematically assess the potential benefits for prediction accuracy of using multi-task learning to perform joint diagnosis of these interlinked conditions. We estimated effect sizes for each condition, benchmarked diagnostic accuracy using common machine learning methods, and then using multi-task learning. We attempted to control for multiple confounding factors throughout the analyses, and discuss different approaches to do so in the predictive modelling context. The hypothesis that latent factors shared between CNVs and psychiatric conditions would make them sufficiently related as prediction tasks to benefit from being learned jointly was not supported. However, we also applied multi-task learning across sites to predict a common target and showed that prediction can be improved when tasks are very tightly related. We implemented a lightweight hard parameter sharing model, but evidence from our results and the literature shows this framework is not well suited to heterogeneous tasks or, counterintuitively, to small sample sizes. While we believe there is potential to exploit the similarities between CNVs and psychiatric conditions using methods that model relationships between tasks, small sample sizes for rare CNVs are a major limitation for the application of multi-task learning.

Keywords: Multi-task learning, fMRI, CNVs, psychiatric conditions, confounds

Table of Contents

Résumé.....	3
Abstract.....	4
Table of Contents.....	5
List of Tables.....	9
List of Figures.....	10
List of acronyms and abbreviations.....	11
Acknowledgements.....	12
Chapter 1 - Introduction.....	13
1.1 - Psychiatric Conditions & CNVs.....	13
1.1.1 - Copy Number Variants.....	14
1.1.2 - Autism Spectrum Disorder.....	14
1.1.3 - Attention-Deficit/Hyperactivity Disorder.....	14
1.1.4 - Bipolar Disorder.....	15
1.1.5 - Schizophrenia.....	15
1.2 - fMRI.....	15
1.2.1 - fMRI Data.....	16
1.2.2 - Preprocessing.....	16
1.2.3 - rs-fMRI.....	17
1.2.4 - Multi-Site Data.....	17
1.3 - Machine Learning.....	19
1.3.1 ML in fMRI.....	19
1.3.2 ML Basics.....	19
1.3.3 Artificial Neural Networks.....	21
1.3.4 - Managing Confounds.....	23
1.4 - Multi-Task Learning.....	24
1.4.1 - Parameter Sharing.....	25
1.4.2 - Multi-Task Learning in fMRI.....	25
1.5 - In This Work.....	26
Chapter 2 - Datasets.....	27
2.1 - Intro.....	27
2.2 - Methods.....	28
2.2.1 - Cohorts.....	28

2.2.2 - CNV Calling.....	29
2.2.3 - Reducing class Imbalance.....	29
2.2.4 - rs-fMRI Preprocessing.....	30
2.2.5 - Computing Connectomes.....	30
Chapter 3 - Connectome Wide Association Studies.....	31
3.1 - Intro.....	31
3.2 - Methods.....	31
3.2.1 - Connectome-Wide Association Studies.....	31
3.2.2 - Estimating Effect Size.....	32
3.3 - Results.....	33
3.3.1 - Effect sizes of neurodevelopmental and psychiatric conditions follow a spectrum from small to large.....	33
3.3.2 - Effect sizes of neurodevelopmental and psychiatric conditions are robust to cross-validation.....	34
3.4 - Discussion.....	34
Chapter 4 - Benchmark Study.....	36
4.1 - Intro.....	36
4.2 - Methods.....	37
4.2.1 - Cross-validation.....	37
4.2.2 - Confound variables.....	37
4.2.3 - Classifiers.....	37
4.3 - Results.....	38
4.3.1 - Only large effect CNVs and psychiatric conditions can be predicted above chance level when controlling for site effects.....	38
4.3.2 - Datasets featuring a large number of sites (over 6) generalised to unseen sites...	39
4.4 - Discussion.....	40
Chapter 5 - Confound-Isolating Cross-Validation.....	42
5.1 - Intro.....	42
5.2 - Methods.....	43
5.2.1 - Automated balanced test set generation.....	43
5.2.2 - Iterative generation of test sets.....	43
5.2.3 - Evaluating test sets.....	44
5.2.4 - Connectomes alone benchmark.....	44
5.3 - Results.....	45

5.3.1 - Confounds models with balanced test datasets predict near chance level for most conditions.....	45
5.3.2 - Diagnosis from connectomes alone reaches similar performance on balanced test sets than traditional intra-site cross-validation.....	46
5.4 - Discussion.....	47
Chapter 6 - Multi-Task Prediction of Age and Sex.....	49
6.1 - Intro.....	49
6.2 - Methods.....	49
6.2.1 - Implementation.....	49
6.2.2 - Architectures.....	49
6.2.3 - Training.....	50
6.2.4 - Predicting Sex & Age.....	50
6.3 - Results.....	52
6.3.1 - Multi-task Learning improves prediction of sex for a majority of sites.....	52
6.3.2 - Multi-task Learning improves prediction of age for a majority of sites.....	53
6.4 - Discussion.....	54
Chapter 7 - Multi-task Learning for Joint Diagnosis of CNVs and Psychiatric Conditions.....	55
7.1 - Intro.....	55
7.2 - Methods.....	55
7.3 - Results.....	56
7.3.1 - Multi-task learning fails to improve automatic diagnosis across heterogeneous conditions.....	56
7.4 - Discussion.....	57
Chapter 8 - Negative Transfer Study.....	58
8.1 - Intro.....	58
8.2 - Methods.....	58
8.2.1 - Architectures.....	58
8.2.2 - Exploring Negative Transfer.....	59
8.3 - Results.....	60
8.3.1 - Negative transfer between conditions trained pairwise is stable across models and data settings.....	60
8.4 - Discussion.....	61
Chapter 9 - General Discussion.....	63
References.....	67
Appendix A - Evaluating Confound-Isolating Cross-Validation in MTL Setting.....	78

A.1 - Intro.....	78
A.2 - Methods.....	78
A.2.1 - Architectures.....	78
A.3 - Results.....	79
A.3.1 - Confounds models consistently predict at chance level for most conditions using a neural network on balanced test sets.....	79
A.3.2 - Confounds models predict at chance level using multi-task learning averaged across balanced test sets.....	80
A.4 - Discussion.....	80
Appendix B - PyNM.....	82
Appendix C - Supplementary Materials.....	87
C.1 - Demographics by Site.....	87
C.2 - Effect Size Table.....	89
C.3 - Confound-Isolating Cross-Validation.....	90
C.4 - Single vs Multi-task - Sex.....	90
C.5 - Single vs Multi-Task Learning - Age.....	91
C.6 - Single vs Multi-Task Learning - Conditions.....	92

List of Tables

Table 1	Demographics by condition.....	27
Table 2	Female subjects by scanning site.....	51
Table 3	Demographics by scanning site.....	89
Table 4	Effect size of conditions on FC.....	91
Table 5	Accuracy of age prediction for each site using MLPs in single and multi-task learning.....	92
Table 6	Performance of age prediction for each site using MLPs in single and multi-task learning.....	93
Table 7	Accuracy of automatic diagnosis for each condition using MLPs in single and multi-task learning.....	94

List of Figures

Figure 1	4D (3D + time) fMRI data.....	15
Figure 2	Common rs-fMRI pipeline for machine learning prediction.....	17
Figure 3	MLP architecture.....	22
Figure 4	Parameter sharing in artificial neural networks.....	24
Figure 5	The effect size and 95% CI of each condition on FC.....	33
Figure 6	Cross-validation of the effect size of each condition on FC.....	34
Figure 7	Intra- and inter-site cross-validation.....	37
Figure 8	Performance accuracy of automated diagnosis using intra-site cross-validation.....	38
Figure 9	Performance accuracy of automated diagnosis using inter-site cross-validation.....	39
Figure 10	Confound-isolating cross-validation.....	42
Figure 11	Performance accuracy of automated diagnosis from confounds using confound-isolating cross-validation.....	45
Figure 12	Performance accuracy of automated diagnosis from connectomes using confound-isolating cross-validation.....	46
Figure 13	Accuracy of sex prediction using single vs. multi-task learning.....	52
Figure 14	Performance of age prediction using single vs. multi-task learning.....	53
Figure 15	Accuracy of automated diagnosis using single vs multi-task learning.....	56
Figure 16	Difference in accuracy from single-task baseline of conditions trained pairwise using multi-task learning in different data and model settings.....	60
Figure 17	Accuracy during training of MLPconf in the single-task setting.....	81
Figure 18	Accuracy during training of MLPconf in the multi-task setting.....	82
Figure 19	DEL 22q11.2 - Distribution of confounds by class.....	92
Figure 20	SZ - Distribution of confounds by site.....	92

List of acronyms and abbreviations

ABIDE	Autism Brain Imaging Data Exchange
ADHD	Attention Deficit Hyperactivity Disorder
ASD	Autism Spectrum Disorder
BIP	Bipolar
BOLD	Blood Oxygenation Level Dependent
CI	Confidence Interval
CNN	Convolutional Neural Network
CNP	Consortium for Neuropsychiatric Phenomics
CNV	Copy Number Variant
CV	Cross-Validation
CWAS	Connectome Wide Association Study
DEL	Deletion
DUP	Duplication
ENIGMA	Enhancing Neuro Imaging Genetics through Meta Analysis
FDR	False Discovery Rate
FC	Functional Connectivity
GNB	Gaussian Naive Bayes
fMRI	Functional Magnetic Resonance Imaging
kNN	k-Nearest Neighbours
LR	Logistic Regression
ML	Machine Learning
MLP	Multi Layer Perceptron
MRG	Montreal Rare Genomic disorder family project
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NIAK	Neuroimaging Analysis Kit
ReLU	Rectified Linear Unit
Ridge	Ridge Regression
RF	Random Forest
rs-fMRI	Resting State Functional Magnetic Resonance Imaging
SVC	Support Vector Classifier
SVIP	Simons Variation in Individuals Project
SZ	Schizophrenia
UCLA	University of California Los Angeles
UKBB	UK Biobank

Acknowledgements

First and foremost I'd like to thank my research directors Pierre Bellec, Sebastien Jacquemont, and Guillaume Dumas, and my mentor Clara Moreau. Thanks to my classmate Nadine Younis for setting up an interview for me with Sebastien when I was floating around looking for a research director in the computational biology specialty. Coming from a pure mathematics background and having left the artificial intelligence specialty last minute, I was looking for something more tangible but did not dream of having the opportunity to be in neuroscience. I feel like I ended up where I was supposed to be and I am tremendously grateful for everything I've learned and for who I got to learn from - inspiring and passionate people who are changing the field for the better and who I simply had fun spending time with. I would also like to especially thank Julie Boyle for making the lab feel like a home and giving me such well-timed and insightful advice.

I suffered a concussion partway through this master's degree and spent a difficult year recovering. I want to express my deepest gratitude to Pierre, Sebastien, Guillaume, Clara and Julie for their patience, support and compassion, and to my physiotherapist Geneviève Ferland, who rescued me and taught me how to manage. I also want to thank the most wonderful administrative staff Celine Begin, Laura Peyras, Marine Lardennois, and Emilie Dessureault for all their help.

Thank you to my parents, sister, and friends for their love and care throughout this master's and always.

AH was supported by a donation from the Courtois foundation to PB. The project was also funded by an IVADO grant to PB and SJ. PB is a senior fellow (chercheur boursier) of the Quebec Research Funds - Health (FRQ-S). The computational infrastructure used in this project was made available by Digital Alliance Canada, through resource allocation grants to SJ and PB.

Chapter 1 - Introduction

Resting state functional magnetic resonance imaging (rs-fMRI) has emerged as a promising diagnostic biomarker technology that is sensitive to a wide range of conditions, including neurodevelopmental, genetic and psychiatric. Translation of rs-fMRI biomarkers into clinical practice has however been hindered by lack of generalizability, i.e. findings from research studies working on an isolated group did not translate to the massive heterogeneity seen in clinical practice. Clinical heterogeneity stems both from the varied technical characteristics of the images, as well as the biological and phenotypic presentation of individuals. The diagnosis of these disorders is completely based on behavioural assessments and high rates of comorbidities and genetic and symptom overlap supports the existence of latent factors shared across conditions. Rare large genetic mutations, called copy number variants (CNVs), have been associated with a range of psychiatric conditions and have much larger effect sizes on brain structure and function, which makes them a promising avenue for untangling the genetics of the current diagnostic categories. There is a growing trend towards the use of machine learning (ML) models trained on massive datasets aggregated across many sources, and multi-task learning in particular has the potential to extract common representations shared across diagnostic tasks, effectively generalising imaging signatures to novel acquisition sites and populations, at the level of individual participants. In this work, we collected an unprecedented dataset consisting of 19 CNVs and psychiatric disorders and aimed to systematically assess the potential benefits for prediction accuracy of using multi-task learning to perform joint diagnosis of these interlinked conditions.

1.1 - Psychiatric Conditions & CNVs

Neuroimaging biomarkers aim to capture the biological mechanisms underlying psychiatric disorders. Currently, the diagnosis of these disorders is completely based on behavioural assessments. Although these conditions are severe, they have a small effect size on brain structure and function, which implies large biological heterogeneity within the current diagnostic categories (Moreau, Raznahan, et al. 2021; Bernanke et al. 2022) and explains the relative lack of reproducibility of results across studies. High rates of comorbidities and evidence of genetic overlap across diagnostic categories (Romero et al. 2022) support the existence of latent factors shared across diagnostic boundaries. Large genetic mutations, called copy number variants (CNVs), have been associated with a range of neurodevelopmental and psychiatric conditions and can serve as criteria for delineating genetically-informed groups of patients. Studies focussing on brain alterations in CNV carriers have found much larger effect sizes than in psychiatric conditions (Modenato et al. 2021; Moreau, Ching, et al. 2021; Moreau, Raznahan, et al. 2021; Moreau et al. 2020; Sønnerby et al. 2022), which makes them a promising avenue for untangling the genetics of the current diagnostic categories.

1.1.1 - Copy Number Variants

A CNV is a large deletion (DEL) or duplication (DUP) of genetic material relative to a reference genome, and often comprises thousands to hundreds of thousands of base pairs spanning multiple genes. CNVs can be classified into three groups: pathogenic, uncertain effect, and benign. In most cases, benign CNVs are relatively common, over 1% of the population, and heritable, which means that the mutation is passed down from a parent. By contrast, pathogenic CNVs are often de novo, which means that the mutation occurs during reproduction (Kearney et al. 2011). Pathogenic CNVs have been associated with a variety of neurodevelopmental and psychiatric conditions, including autism spectrum disorder (ASD) associated with 16 different CNVs, and schizophrenia (SZ) associated with 14 different CNVs (Satterstrom et al. 2020; Sanders et al. 2019; Marshall et al. 2017; Rees and Kirov 2021). CNVs have also been associated with Bipolar (BIP) disorder and Attention-Deficit/Hyperactivity Disorder (ADHD), although such associations are less frequent (Rees and Kirov 2021). Carriers of the same pathogenic CNV can have large variability in the severity of clinical symptoms.

DEL 22q11.2 and DEL 16p11.2 are the most studied pathogenic CNVs, and rare examples of heritable (non de-novo) CNVs with severe clinical manifestations. Both variants have indeed been found to have large clinical effect sizes (Crawford et al. 2019; Jonas, Montojo, and Bearden 2014; Rees and Kirov 2021; Willsey et al. 2022; Moreau et al. 2023). DEL 22q11.2 is the biggest known risk factor for SZ: 30% of carriers will develop the condition in their lifetime (Marshall et al. 2017) and its diagnosis also carries an elevated risk for ASD (32 times higher than the general population). DEL 16p11.2 is associated with ASD, as well as with ADHD (Moreno-De-Luca et al. 2013; Niarchou et al. 2019; Sanders et al. 2015).

1.1.2 - Autism Spectrum Disorder

ASD is a neurodevelopmental condition estimated to have a worldwide prevalence of about 1 in 100 children (Zeidan et al. 2022). ASD is characterised by impaired social interaction and communication, atypical patterns of activities and behaviours, focus on details and sensitivity to sensory stimuli. The presentation of symptoms is very diverse among people with ASD, some people are independent and exceptionally intelligent, while others have severe disabilities and require life-long care and support. Comorbidities are common in ASD: in a cohort of 112 children Simonoff and colleagues (Simonoff et al. 2008) found 70% of individuals were diagnosed with another psychiatric condition, most commonly ADHD or social anxiety, and 41% were diagnosed with two or more.

1.1.3 - Attention-Deficit/Hyperactivity Disorder

ADHD is a neurodevelopmental condition, diagnosed in about 2.5% of the general population (Simon et al. 2009). ADHD is characterised in children by developmentally inappropriate levels of inattention, impulsivity, and hyperactivity, however the presentation of symptoms in adults

is more heterogeneous (Katzman et al. 2017). ADHD is associated with social, academic, occupational, and neuropsychiatric deficits and represents a considerable burden for diagnosed individuals and society (Barkley 2002; Biederman et al. 2008; Posner, Park, and Wang 2014). Up to 80% of adults with ADHD are diagnosed with at least one comorbid psychiatric disorder (Katzman et al. 2017), frequently mood and anxiety disorders (Rösler et al. 2010), and there are overlapping symptoms and neurobiological similarities with these conditions.

1.1.4 - Bipolar Disorder

BIP is a psychiatric condition estimated to impact 1% of people worldwide (Merikangas et al. 2011). BIP is characterised by periods of depression and mania (bipolar disorder type I) or hypomania (bipolar disorder type II) (Syan et al. 2018). Comorbidity is so common in BIP that McElroy (McElroy 2004) states that it is the rule rather than the exception. Frequently observed comorbidities include anxiety, substance abuse and conduct disorders, ADHD, and ASD (Sajatovic 2005).

1.1.5 - Schizophrenia

SZ is a psychiatric condition with relatively low prevalence, estimated to be 4.4 in 1000 in the general population (Moreno-Küstner, Martín, and Pastor 2018). Despite being relatively uncommon, SZ was ranked the 12th most disabling disorder among 310 diseases and injuries globally (Hay et al. 2017) and its economic burden globally was estimated to range from 0.02% (UK) to 1.65% (Sweden) of gross domestic product (Chong et al. 2016). SZ is characterised by hallucinations, disorganised communication or behaviour, impaired cognitive ability, and blunted affect (Saha et al. 2005; Patel et al. 2014). Comorbidities with other psychiatric disorders are common, including substance abuse, depression, panic disorders, post-traumatic stress disorder, and obsessive compulsive disorder (Tsai and Rosenheck 2013).

1.2 - fMRI

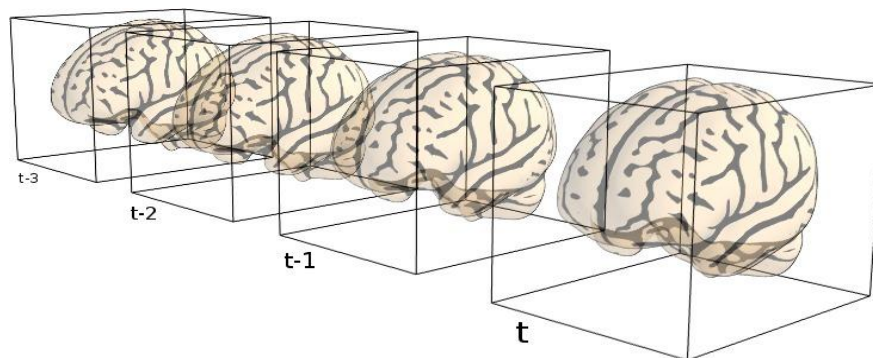


Figure 1 - 4D (3D + time) fMRI data. Image from (Nilearn n.d.).

1.2.1 - fMRI Data

Magnetic resonance imaging (MRI) uses the magnetic properties of tissues in the body to non-invasively create a detailed 3D image of an organ. An MRI machine produces a strong magnetic field that aligns the magnetic properties (spin) of atomic nuclei, and then excites these spins with magnetic pulses sent from a Radio-frequency coil. The excited nuclei then emit a signal in return, referred to as the echo pulse which is received by another coil. The speed of decay of the echo pulse at each location reflects various characteristics of the composition of biological tissues, and is mapped to a corresponding intensity value which results in a 3D image.

When neurons in the brain become active, the amount of blood flowing near them increases at a microscopic scale and causes a localised surplus in blood oxygen. The two major forms of haemoglobin in the blood (oxygenated vs deoxygenated) have markedly different magnetic properties. Deoxy-haemoglobin acts as a natural, endogenous contrast agent for MRI, we can measure the blood oxygenation level dependent (BOLD) signal. Functional MRI (fMRI) uses rapid acquisition of MRI images capturing changes in BOLD signal over time to study the dynamic activity of the brain.

fMRI data captures information across three spatial dimensions as well as time and results in a 4D tensor. The smallest unit of data is a voxel (3D pixels) at a given point in time. In general, a voxel includes a very large quantity of individual neurons and the response signal represents a summary of the neural activity in that region. Typical fMRI scanners acquire an image of the whole brain approximately every few seconds (2-3 s) with a voxel size of typically $3 \times 3 \times 3 \text{ mm}^3$ (Glover 2011), resulting in data with very high dimensionality. Note that the exact spatial and temporal resolution of an acquisition is a matter of trade-off between several factors, including emphasis on spatial vs temporal resolution, signal-to-noise ratio, acquisition time as well as the size of the field-of-view which may cover only parts of the brain for increased resolution. Modern acquisitions can reach less than 2 mm isotropic spatial resolution, and under 1 second repetition time for a full brain volume. Due to statistical issues with such large data, it is common to reduce dimensionality by using a parcellation of the brain and considering the average activity within each parcel - resulting in a 2D matrix (time x space) often referred to as time series. Parcellations can be defined in many ways, highlighting various functional or structural landmarks, depending on the aim of the study (Kong et al. 2021; Dadi et al. 2019).

1.2.2 - Preprocessing

Preprocessing of fMRI data is complex and designed to correct for issues with data acquisition, especially artefacts from the scanner caused by movement of the subject. In a typical pipeline, data is quality controlled visually, then images are corrected for spatial distortions and realigned across time to correct for motion. Individual scans are next aligned to a common

brain template, and finally spatially smoothed and filtered to remove low-frequency noise before being used for analyses. Detailed information about preprocessing can be found in the reference documentation for the common software packages fMRIPrep (Esteban et al. 2019) and Neuroimaging Analysis Kit (NIAK) (Bellec et al. 2012).

1.2.3 - rs-fMRI

There are two major types of fMRI data acquisition: task based, and resting state. In task based studies, the subject performs a task or responds to some stimuli in the scanner and the following analysis is focussed on the response in the fMRI signal. In resting state studies, there is no controlled experimental paradigm, subjects are often asked to think of nothing in particular and focus their eyes on a point for the duration of the scan. This approach is well suited to large multi-site studies, as it reduces the difficulty of harmonising experimental settings at different sites of data collection. Resting state fMRI (rs-fMRI) measures spontaneous fluctuations in the BOLD signal across the brain, which exhibit well-replicated networks of coordinated brain regions (Khosla et al. 2019).

Functional connectivity (FC) between any two regions is the degree to which their activity is coupled, usually measured as the correlation between the signals from each region. The connectome, a matrix representing the FC between all pairs of regions, has been the focus of neuroimaging research as a biomarker (Khosla et al. 2019), an individual's connectome has been shown to be unique and reliable (Finn et al. 2015). Studies using predictive models on connectomes have succeeded in automatic diagnosis of a wide range of neurodevelopmental and psychiatric conditions (Arbabshirani et al. 2017).

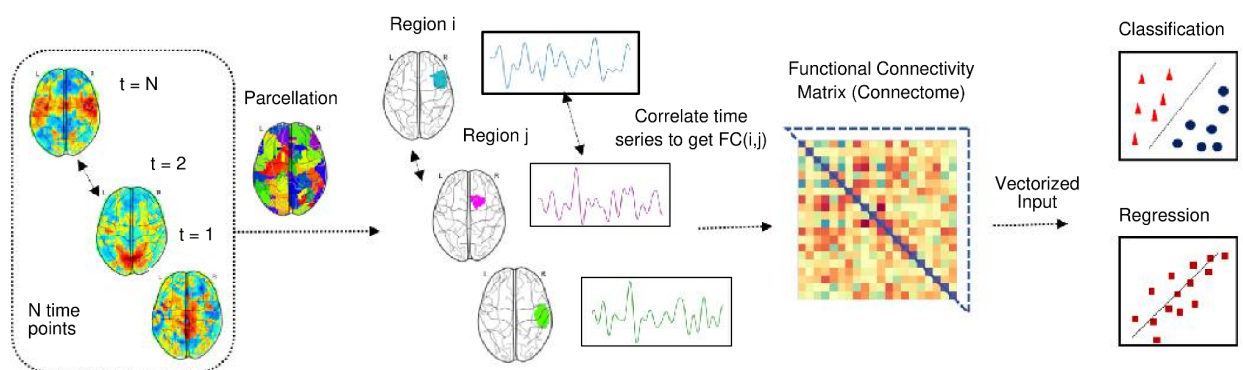


Figure 2 - Common rs-fMRI pipeline for machine learning prediction. Image from (Khosla et al. 2019).

1.2.4 - Multi-Site Data

Neuroimaging data is complex and expensive to collect. As a result most studies to date on single-site data have rarely provided more than a few hundred subjects. While in traditional analyses statistical power is limited by sample size, the problem is worse for ML studies in which larger datasets are needed to properly train and accurately estimate the generalisation of

complex models (Varoquaux 2018). Most classification works in mental illness are performed in the single site context in which performance is evaluated on a test sample of subjects with the same characteristics (and acquisition site) as the training sample. This approach is problematic as it does not test if the pattern learned by the model can generalise to different populations or sites of acquisition. A review by Arbabshirani et al 2017 (Arbabshirani et al. 2017) found that almost all studies that reported very high accuracies, had sample sizes smaller than 100. Most concerning, the reported overall accuracy decreased with sample size in most disorders such as SZ and ADHD. This trend may reflect overfitting following questionable research practices, which is easier to achieve on a small sample size, but likely reflect the difficulty to generalise to the large heterogeneity of phenotypic presentations found in clinical cohorts.

Larger and more diverse samples are crucial to properly evaluate a model's ability to generalise and be reliable in a clinical context. As a result, it is becoming more and more common in the field to pool datasets across studies and create large shared repositories of data collected at various sites. However, massive heterogeneity exists in data across different sites of collection and studies due to differences in scanner make, data acquisition, data processing and sampling. There is an established impact of scanner and sequence characteristics on the reproducibility of multisite fMRI (Badhwar et al. 2020). Individual acquisitions themselves are variable, even at a single site, and this variability depends critically on how much data is collected by subject (Gordon et al. 2017; Noble et al. 2017), see Noble and colleagues (Noble, Scheinost, and Constable 2019) for review. These sources of acquisition variance remain an obstacle for developing good prediction models, in addition to phenotypic variability. Using data from 191 people diagnosed with SZ and matched controls collected from six scanning sites, Orban (Orban et al. 2018) found that classification of sites could be performed with 84% accuracy, while classification of SZ didn't exceed 75%. While site heterogeneity can have a larger effect on fMRI than the condition being studied, Orban also found that increasing the heterogeneity of the training set improved the ability of the classifier to generalise to an unseen site.

There exist a variety of methods for treating data to reduce site effects. ComBat (Johnson, Li, and Rabinovic 2007), is a popular strategy originally designed to correct 'batch effects' in genomic studies and has been demonstrated to successfully remove site effects in neuroimaging studies (M. Yu et al. 2018). Normative modelling is another framework that can be applied to harmonise data across sites (Bayer et al. 2022). Although it is outside the scope of the current work, a package implementing normative modelling in python, called PyNM, was developed and published (Harvey and Dumas 2022), in addition to this master's project. The accompanying paper is included in the appendices.

1.3 - Machine Learning

1.3.1 ML in fMRI

The standard in rs-fMRI studies has traditionally been to apply mass-univariate brain mapping, in which statistical tests are applied independently at each region to assess group differences. However, these approaches are limited in that they don't make use of information contained jointly across regions. They also do not provide information at the individual level, but rather characterise differences in average between group distributions. Recently, the field has shifted towards flexible ML techniques that aim to predict a category or score for each subject. The goal of such studies is often both to find individualised scores for diagnosis as well as to identify patterns and mechanisms in the brain that characterise a condition (Linn et al. 2016).

Many studies have successfully applied predictive modelling to automatically diagnose psychiatric and neurodevelopmental conditions using biomarkers derived from rs-fMRI, including SZ (Venkataraman et al. 2012; Kim et al. 2016; Bassett et al. 2012), ASD (Traut et al. 2022; Nielsen et al. 2013; Abraham et al. 2017; Khosla et al. 2018), ADHD (Eloyan et al. 2012; Jian Li, Joshi, and Leahy 2020; Z. Wang et al. 2023), and BIP (Rashid et al. 2016; H. Wang et al. 2022). In spite of promising results, prediction accuracies reported in the literature should be interpreted with caution as the majority of prediction studies in mental illness to date have been performed on data collected from a single site with small sample size (Orban et al. 2018). Limited sample size and heterogeneity in data collection across studies has led to issues with statistical power and reproducibility (Abraham et al. 2017; Varoquaux 2018). Integrating large multi-site datasets is essential to train models that can detect subtle patterns and generalise to the diversity encountered in the clinical setting (Q. Ma et al. 2018).

1.3.2 ML Basics

The following two sections briefly introduce basic concepts in ML and neural networks. For a more in depth discussion of all the topics covered here and a thorough presentation see chapters 5 and 6 Deep Learning (Goodfellow, Bengio, and Courville 2016).

ML Algorithms

An ML algorithm is a model that is able to learn a task from a dataset and then execute this task reliably on new inputs. ML algorithms can be split into two main categories: supervised and unsupervised. Unsupervised algorithms are trained on a dataset $\{x^t\}$, where x^t can be considered an example of the random vector \mathbf{x} , with the task of implicitly or explicitly learning the probability distribution $p(\mathbf{x})$, or properties of that distribution, e.g. clustering and density estimation. Supervised algorithms are trained on a dataset $\{(x^t, y^t)\}$ of examples x^t and associated labels y^t with the task of predicting the label, usually by estimating $p(y|\mathbf{x})$. In

supervised learning training is achieved by optimising a cost function, often some measure of distance between the prediction and the target. Supervised learning algorithms can further be divided into classification or regression based on the target of the learning task y . In regression y is a real number such as age of a subject and in classification y is a category such as the diagnostic status of a subject.

Generalisation

The ultimate goal of ML is to be able to make predictions on new data using a model that has been trained on an available dataset. This idea of transferring to new data is called generalisation. When training a model, we try to estimate a model's ability to generalise by evaluating its predictions on a held out sample of the available data called a test set, while the remaining data is referred to as the training set. In the test set, we know the value of the target y and can compare these values with the predictions of the model using some performance measure, e.g. accuracy for classification studies or a mean-squared-error for regression studies. This performance of prediction on the held-out test set is what allows us to score our model.

Cross-validation

When the available dataset is very large (hundreds of thousands of examples or more), there is no practical issue with holding out a portion as a dedicated test set. However, with smaller datasets we can make better use of the data and provide a more reliable estimate of generalisation error by using cross-validation. In cross-validation, we split the dataset into a training and test set then estimate the performance of the model multiple times and look at the average performance across iterations. There are various schemes for how to implement cross-validation, the most common being to split the data into K non-overlapping sets and for K iterations use all but one as training data and the last as test (K -fold cross-validation). Additional requirements can be made of the splits such as having them be stratified with respect to a particular variable, such as gender of subjects in automated diagnosis or the target variable in classification when there are imbalances across class. This means for example that if 70% of the total dataset is composed of males, each training and test set needs to be generated such that roughly 70% of both training and test examples are males, while random choice of train/test set splits may not satisfy this condition.

Overfitting

When we train a model, we are 'fitting' its parameters such that the performance of the model is good on the training data. We want to strike a balance between under- and overfitting, as final performance of the model is based on generalisation with the test data. If we underfit the model, it has not learned sufficiently from the training data and the generalisation performance can be improved with a more flexible training. If the model is overfit, we have the inverse problem and the model has memorised aspects of the training set that do not generalise to the test set. We call the flexibility of a model to fit to data its capacity. Models with high

capacity can learn more complex patterns, but are more susceptible to overfitting. We can reduce overfitting either by training the model less, training it on more data, or reducing the model capacity through various means such as regularisation.

Regularisation

A common way to reduce model capacity is regularisation. In practice we have some prior knowledge about the data distribution we want the model to learn. We express this preference during training with regularisation, which acts as a pressure that pushes it to learn some functions over others. In supervised learning, this is often achieved by adding a penalty to the cost function during training, or a technique called drop-out for artificial neural networks (see section 1.3.3 below).

The curse of dimensionality

Each dimension of the data is a feature for the model to learn. As the dimension of the data increases, the number of distinct ways in which the features can be configured increases exponentially. When data is low-dimensional and the sample size is large, a model can easily learn to generalise well because the feature space is well covered by examples it has seen in training. In the inverse case, we can have a high-dimensional feature space that is only sparsely populated by examples. In this setting, it is very difficult for a model to generalise well. Unfortunately, this problem is very severe in rs-fMRI studies where the dimension of the raw data far exceeds the size of the dataset even in the largest datasets available.

1.3.3 Artificial Neural Networks

Artificial neural networks are a powerful and flexible class of models in ML inspired by biological neurons. The simple units of neurons (also called perceptrons) can be organised into highly complex architectures that reach state of the art performance in a wide variety of tasks. The simplest neural network architecture is a multi-layer perceptron (MLP).

Multi-Layer Perceptrons

In an MLP, units are organised into a series of layers, the number of units in each layer is its dimension and the number of layers in the network is its depth. MLPs are also referred to as fully connected networks, since each unit is connected to all those in the preceding and following layers. Each layer can be considered as a function that are composed to form the network, e.g. a three layer network is a mapping $y = f(x; \theta) = f_3(f_2(f_1(x; \theta_1); \theta_2); \theta_3)$, where f_i is the i th layer, θ_i are its parameters. Each layer computes $f_i(x; \theta_i) = a(W_i x + b_i)$, where a is an activation function, and the parameters θ_i are the matrix of weights W_i and vector of biases b_i . The first layer is called the input layer, intermediate layers are called hidden layers, and the last layer is called the output layer, whose dimension depends on the objective of the task the network is designed to accomplish.

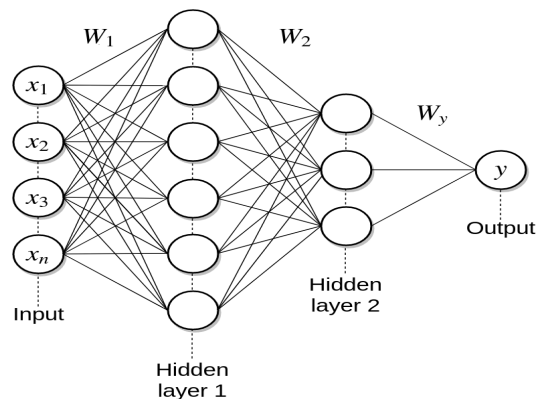


Figure 3 - MLP architecture. Image from (Grattarola 2017).

Output layer

In the regression setting, the output layer delivers the prediction of the network with a unit for each dimension of the target variable. In the classification setting, the dimension of the output layer corresponds to the number of categories in the prediction task and the value of each unit is normalised by the softmax function to represent the probability the network assigns to each.

Activation functions

The most commonly used activation is rectified linear units (ReLUs), which apply the mapping $a(x) = \max\{0, x\}$. Leaky ReLUs, which apply the mapping $a(x) = \max\{x/k, x\}$ where $k \in \mathbb{R}^+$, are another common choice.

1.3.3.4 - Optimisation

The parameters θ of the network f are initialised randomly and optimised during training to reduce the generalisation error of the network as measured by a loss function L , e.g. cross entropy for classification tasks or mean squared error for regression. Stochastic gradient descent is the most commonly used optimization algorithm. For each example $\{x^t, y^t\}$ we calculate the loss $L(f(x^t; \theta), y^t)$, and its gradient with respect to the parameters θ $\nabla_{\theta} L(f(x^t; \theta), y^t)$. We then update each parameter by taking a step, whose size is defined by the learning rate and the magnitude of the gradient, in the opposite direction of the gradient, moving towards a minimum of the loss. For efficiency, the model is often trained using batches of data, across which the loss function at the output is averaged before the gradient is computed. Each pass of training through the whole data is called an epoch. There are several algorithms that improve on classic gradient descent, Adam (Kingma and Ba 2014) is a popular choice that works by computing individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients.

Regularisation

Regularisation (described above in 1.3.2) is often achieved in neural networks by adding a term $\Omega(\theta)$ to the loss function to penalise the parameters i.e. we calculate $L(f(x^t; \theta), y^t) + \Omega(\theta)$ before computing the gradient. The regularisation term can be chosen to produce certain behaviours. For example, L_2 regularisation, also called weight decay, constrains the magnitude of the weights, while L_1 regularisation encourages weights to be sparse. Dropout (Srivastava et al. 2014) is a popular regularisation strategy, in which units (and their connections) are randomly dropped from the neural network during training, preventing them from overly co-adapting. Another strategy is gradient clipping, in which we limit the magnitude of the gradients to prevent them from exploding and causing large jumps in parameter values - potentially skipping an optimal region.

Artificial Neural Networks in fMRI

Artificial neural networks with fully connected dense layers and weights initialised through pre-training with autoencoders, have been successfully used for automatic diagnosis from connectomes (Kim et al. 2016; Heinsfeld et al. 2018). Convolutional neural networks (CNNs) are an efficient modification of MLPs for grid-like data, e.g. images, in which groups of parameters called filters share parameters and slide across the input to extract features which are pooled between layers. CNNs have been successfully used to perform a variety of connectome based classification and regression tasks (Khosla et al. 2018; Kawahara et al. 2017; Meszlényi, Buza, and Vidnyánszky 2017; Leming and Suckling 2021).

1.3.4 - Managing Confounds

There is no standard definition of confounding in ML, but the term is widely used to refer to variables that affect both the brain imaging data and the prediction target but are not considered relevant to the study. For example, ASD diagnosis is markedly more prevalent in males, and therefore many ASD studies recruit either only males, or a majority of male participants. However, gender is reflected in rs-fMRI data and can be predicted with high accuracy. Imagine there is a connection X that is overconnected in females vs males, and also overconnected in male ASD subjects vs typical controls. Say a model is trained on a sample of only male subjects to diagnose ASD and learns to heavily rely on feature X. If we apply that same model to a new population including female subjects, the model will be more likely to classify females as ASD due to the confounding effect of gender on X.

Confounding variables are inevitable in predictive modelling of neuroimaging studies. While some factors such as distributions in age and sex can be controlled with careful participant recruitment, others such as head motion cannot be avoided. Complicating the matter, the need for larger datasets to train ML models has led researchers to recruiting large population imaging cohorts without a balanced design as well as pooling data across heterogeneous studies

and sites of collection. These conditions imply that the data is nearly always confounded by a variety of factors.

Prediction of the target variable can be driven by confounding factors, so it is important to manage them correctly in order to correctly interpret the results of the study and make practical use of biomarkers. The literature on controlling confounding variables is well developed for traditional statistical analysis used in mass-univariate brain mapping. The standard approach is to ‘regress out’ confounds: adjusting input variables for confounds using linear regression before being used as input to an ML model (Dinga et al. 2020; A. Rao et al. 2017). Unfortunately, this approach is problematic for ML models, which are often non-linear and multivariate. Dinga shows that such models can learn information from the data that cannot be regressed out. Additionally, ‘regressing out’ confounds before splitting the dataset for cross-validation leaks information from the test set. We consider two alternative approaches available to ML practitioners. The first is to include confounds alongside the imaging data as predictors and compare the results with a baseline model consisting of only confounds (A. Rao et al. 2017; Orban et al. 2018). The second is to evaluate models on test sets that are balanced with respect to confounds (Leming, Górriz, and Suckling 2020; A. Rao et al. 2017; Chyzhyk et al. 2022), which we refer to as confound-isolating cross-validation following Chyzhyk. In the ideal case the prediction of the model can be considered free from confounding factors, however creating balanced test sets becomes more difficult as the number of confounds increases and may not always be possible.

1.4 - Multi-Task Learning

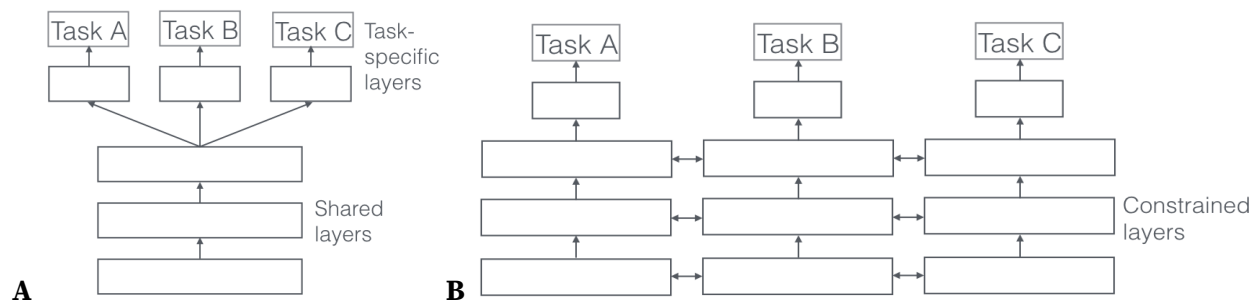


Figure 4 - Parameter sharing in artificial neural networks. A: hard parameter sharing, B: soft parameter sharing. Image from (Ruder 2017).

Multi-task learning is a framework in which rather than training an independent model on each task in parallel (single task learning), a shared model is trained on multiple related tasks concurrently. This idea is analogous to human learning in which useful information shared across related tasks helps both tasks. Ruder (Ruder 2017) provides a perfect example: in the movie *The Karate Kid*, sensei Mr Miyagi has the karate kid complete non-karate tasks such as sanding the floor and waxing a car, which end up giving him skills that are relevant for karate.

When the tasks are well grouped together, multi-task learning makes better use of data by implicitly augmenting the data from each task with the data from the others, and the shared latent representation acts as a form of regularisation across tasks preventing the model from overfitting to any specific task and can boost learning across tasks. In his review, Ruder also highlights attention focusing (one task emphasises a feature that is noisy in another) and eavesdropping (one task uses a feature another task has learned) as mechanisms that can give multi-task learning an advantage.

1.4.1 - Parameter Sharing

In artificial neural networks, there are two main ways in which parts of a model can be shared across tasks. In hard parameter sharing a portion of the model is accessed and modified directly by each task, whereas in soft parameter sharing parallel models for each task have a shared portion linked together. Generally, hidden layers are shared between all tasks and output layers are task-specific. Hard parameter is the most commonly used approach to multi-task learning, the reduction of parameters relative to parallel models makes it lightweight and greatly reduces the chances of overfitting. However, it runs the risk of having tasks that are not sufficiently similar undermine each other, a phenomenon called negative transfer. Soft parameter sharing offers much more flexibility in the design of the model, as information can be shared across tasks in many different configurations. However, in general it increases the overall number of parameters to train relative to training parallel models.

Outside of the artificial neural network setting, multi-task learning is usually achieved by regularising the matrix of model parameters jointly, which we consider a form of soft parameter sharing. In this form, task relationships can be learned by including a constraint e.g. that encourages clustering (Ruder 2017).

1.4.2 - Multi-Task Learning in fMRI

There has been a lot of recent interest in multi-task learning in the ML literature, as reviewed in (Pan and Yang 2010), and this interest has begun to spread to neuroimaging. Although there are still relatively few examples, multi-task learning has been applied across target clinical variables using rs-fMRI data (Rahim et al. 2017) and combined imaging modalities (D. Zhang, Shen, and Alzheimer's Disease Neuroimaging Initiative 2012), across timepoints to predict disease progression using cortical surface data (Zhou et al. 2013), across individuals to perform brain decoding using fMRI data (N. Rao et al. 2013; Marquand et al. 2014), across fMRI task conditions to predict intelligence quotient (IQ) (Xiao et al. 2020), and across sites (Q. Ma et al. 2018; Hu and Zeng 2019) (Q. Ma et al. 2018; Hu and Zeng 2019; Watanabe et al. 2014) and disease subtypes (X. Wang et al. 2015) to perform automatic diagnosis. Most of these studies explored regularisation schemes to jointly learn features across tasks (soft-parameter sharing) with a range of models. Various deep learning architectures have also been applied to neuroimaging data using multi-task learning (Liang et al. 2021; C. Yu et al. 2021; Tabarestani et al. 2022; He et

al. 2020; Dong et al. 2020; Ngo et al. 2020). There are only two previous studies applying multi-task learning across psychiatric conditions, the first examined ASD and ADHD and the second added SZ (Huang, Liu, and Tan 2020; Huang et al. 2022). In both studies they applied a Multigate Mixture of Experts model, which learns relationships between tasks (soft parameter sharing), using a pre-training feature selection step and reported improvements relative to single task learning and multi-task MLP models (hard parameter sharing).

1.5 - In This Work

In Chapter 2 of this thesis, we first introduce the 19 CNVs and psychiatric conditions in our dataset, and then apply traditional statistical approaches as a baseline to estimate effect sizes of all conditions in Chapter 3. We then benchmarked the accuracy of common prediction algorithms on the same conditions, using two different cross-validation strategies to account for site effects and evaluating accuracy relative to a confounds-only baseline model (Chapter 4). In the following chapters, we focussed our attention on the 9 CNVs and psychiatric conditions that could be well predicted in the benchmark. Next, we described confound-isolating cross-validation, found appropriate test sets for each condition, and compared the accuracy of prediction using this strategy to the results of the benchmark (Chapter 5). We introduced our multi-task learning framework and evaluated its performance on a simple, well controlled benchmark where we predicted age and sex using different sites of data collections as different tasks (Chapter 6). Finally, we applied multi-task learning to perform joint diagnosis of the 9 CNVs and psychiatric conditions using confound-isolating cross-validation (Chapter 7). In Chapter 8, we explored the phenomenon of tasks interfering with rather than aiding each other's learning that was observed in Chapter 7.

Chapter 2 - Datasets

Condition	N		Age		Global Signal		Motion		Sites
	Total	(F)	Mean	(SD)	Mean	(SD)	Mean	(SD)	
A DUP 15q13.3	191	(100)	64.34	(7.32)	0.42	(0.12)	0.19	(0.05)	4
DEL 2q13	183	(110)	63.06	(7.24)	0.43	(0.13)	0.19	(0.05)	3
DUP 15q11.2	136	(76)	63.71	(7.19)	0.41	(0.13)	0.19	(0.05)	3
DUP 2q13	88	(43)	64.68	(7.62)	0.42	(0.13)	0.19	(0.05)	3
DUP 16p13.11	41	(21)	63.99	(7.70)	0.38	(0.13)	0.19	(0.04)	4
DUP TAR	29	(14)	59.85	(7.54)	0.42	(0.11)	0.17	(0.05)	3
DUP 13q12.12	20	(10)	60.84	(7.34)	0.49	(0.13)	0.20	(0.06)	3
DEL 13q12.12	22	(12)	63.54	(5.86)	0.42	(0.14)	0.20	(0.07)	3
B DEL 15q11.2	103	(55)	64.29	(7.44)	0.43	(0.14)	0.19	(0.06)	3
DUP 16p11.2	35	(14)	34.15	(19.53)	0.31	(0.12)	0.21	(0.09)	6
DUP 22q11.2	22	(12)	39.43	(23.49)	0.44	(0.11)	0.19	(0.09)	5
DEL 1q21.1	25	(12)	44.40	(18.87)	0.38	(0.13)	0.18	(0.07)	6
DUP 1q21.1	19	(13)	50.86	(19.35)	0.45	(0.19)	0.21	(0.08)	7
DEL 16p11.2	32	(13)	21.74	(20.14)	0.37	(0.11)	0.22	(0.09)	5
DEL 22q11.2	43	(19)	16.86	(6.95)	0.35	(0.14)	0.18	(0.07)	1
C ADHD	223	(66)	14.71	(9.47)	0.41	(0.12)	0.15	(0.04)	7
ASD	472	(0)	14.71	(5.88)	0.37	(0.140)	0.17	(0.05)	28
SZ	283	(73)	33.90	(9.22)	0.37	(0.140)	0.17	(0.06)	12
BIP	44	(20)	35.02	(8.95)	0.40	(0.14)	0.17	(0.07)	2
CON	31425	(16590)	62.41	(11.47)	0.42	(0.13)	0.18	(0.05)	53

Table 1 - Demographics by condition. A: Non-Psychiatric CNVs, B: Psychiatric CNVs, C: Psychiatric Conditions. The first two columns are the number of total subjects, and of female subjects (in parentheses) for condition and the pooled control subjects (CON). The intermediate columns show the mean age, global signal, and head motion, with standard deviation (in parentheses). The final column shows the number of scanning sites contributing to the dataset.

2.1 - Intro

In this chapter we present the mega-dataset used in this thesis, which is a compilation from 9 rs-fMRI datasets. In chapter 3, traditional statistical analyses are applied for each condition separately on the entire mega-dataset. We then generate class-balanced sub-datasets for each

condition which are used for the application of ML techniques in the remaining chapters. All the data were preprocessed from raw rs-fMRI time series using a standardised pipeline. This pipeline ultimately generates connectomes, i.e. a measure of FC between brain regions over the course of the entire scan for each subject.

2.2 - Methods

2.2.1 - Cohorts

The 9 datasets included in this thesis are themselves typically compiled from different studies and sites of data collection. Specifically, the datasets included four clinical cohorts, five idiopathic neuropsychiatric datasets and one very large sample of unselected individuals based in the United Kingdom. In total, 33,436 individuals were included, who were either neurotypical control subjects or individuals diagnosed with with one of 7 CNVs associated with neurodevelopmental and psychiatric disorders (so-called psychiatric CNVs), 8 nonpsychiatric CNVs, or 4 idiopathic psychiatric disorders (ASD, SZ, BIP, ADHD). The research ethics review boards of each relevant institution approved the study of the corresponding dataset. The present secondary analysis project was approved by the research ethics review board at the Centre Hospitalier Universitaire Sainte-Justine.

Clinical Genetic Datasets

Participants in the four clinical genetic datasets were recruited based on the presence of psychiatric CNVs regardless of the presentation of symptoms, along with matched control subjects. These four clinical CNVs datasets included the Simons Variation in Individuals Project (SVIP) (Simons Vip Consortium 2012) and the following unpublished datasets: University of California, Los Angeles 22q11.2 CNV project (UCLA), the Montreal rare genomic disorder family project (MRG, Canada) and the Define Neuropsychiatric-CNVs Project (Cardiff, United Kingdom).

Idiopathic Psychiatric Conditions Cohorts

We included 5 psychiatric datasets: Autism Brain Imaging Data Exchange 1 (ABIDE1) (A. Di Martino et al. 2014), Autism Brain Imaging Data Exchange 2 (ABIDE2) (Adriana Di Martino et al. 2017), ADHD-200 (ADHD-200 Consortium 2012), Consortium for Neuropsychiatric Phenomics (CNP) (Poldrack et al. 2016), and aggregate dataset of 10 SZ studies (Orban et al. 2017; Moreau et al. 2020). These studies provided data for individuals with ASD, ADHD, SZ, BIP and matched control subjects.

Unselected Population

CNVs were identified in the UK Biobank (UKBB) (Sudlow et al. 2015), which included both CNVs associated with neurodevelopmental and psychiatric disorders and nonpsychiatric CNVs.

Nonpsychiatric CNVs were defined as variants without any previous association with a psychiatric condition in large case-control studies (Marshall et al. 2017; Sanders et al. 2015; Jönch et al. 2019; Moreno-De-Luca et al. 2013).

2.2.2 - CNV Calling

CNVs were identified in the UKBB using PennCNV (K. Wang et al. 2007) and QuantiSNP (Colella et al. 2007) following previously published methods (Huguet et al. 2018).

2.2.3 - Reducing class Imbalance

General Class Balancer

For the ML studies, we used the General Class Balancer to select subsets of each dataset which were balanced regarding diagnostic classes as well as the distribution of confound variables inside each diagnostic class. The General Class Balancer algorithm (Leming, Górriz, and Suckling 2020) exactly matches categorical variables such as diagnosis, sex and site, while continuous confounds such as age and head motion are quantized into discrete bins prior to matching. Smaller and smaller bins are created recursively until all subjects can be matched while the distributions of the confound between classes fails a Mann-Whitney U-test, which evaluates if two distributions are statistically different. As some of the datasets were highly unbalanced regarding some confounding factors, the final subsets were not always perfectly matched in terms of confound distributions across classes. In summary, this procedure perfectly balanced classes, and reduced the imbalance of confound distribution to the greatest extent possible. Some exceptions to this general procedure were also made, as detailed below.

CNVs

The CNV datasets have major class imbalance, with far more controls than case subjects. For most of the CNVs, we applied General Class Balancer with no modifications. When applied to DUP16p11.2, the General Class Balancer algorithm consistently failed to find a match for one specific subject when applied repeatedly with different random seeds. We had to hand-select the closest matching control for this subject. The DEL 22q11.2 dataset was collected entirely from a single site and participants were recruited in a balanced design, and in this case we used all the subjects available without applying General Class Balancer.

Psychiatric conditions

For the psychiatric conditions, the sample size was markedly larger than with CNVs, and class imbalance was also less severe. We thus used all the available cases and controls from each study, without application of General Class Balancer.

2.2.4 - rs-fMRI Preprocessing

All datasets were preprocessed using the same parameters of NIAK (Bellec et al. 2012). Preprocessed data were visually controlled for quality of the co-registration, head motion, and related artefacts.

2.2.5 - Computing Connectomes

We used the Multiresolution Intrinsic Segmentation Template (MIST) brain parcellation (Urchs 2017) to segment the brain into 64 regions. This functional brain parcellation was found to have excellent performance in several ML benchmarks on either functional or structural brain imaging (Hahn et al. 2022; Dadi et al. 2020; Mellema et al. 2022). We chose the 64 parcel atlas of the MIST parcellation because this range of network resolution was found to be sensitive to changes in FC in techniques neurodevelopmental and psychiatric disorders such as autism, both using ML (see previous references) as well as classical mass univariate regression (Bellec et al. 2015). Functional connectivity (FC) between any two regions was defined as the Fisher z-transformed Pearson's correlation between the average time series of each region, while within region connectivity is the Fisher z-transformed average of Pearson's correlation between any pair of distinct voxels within the region. Each connectome consisted of 2080 values: $(63*64)/2 = 2016$ region-to-region connections plus 64 within region connectivity values.

Chapter 3 - Connectome Wide Association Studies

3.1 - Intro

This chapter aims to measure how difficult each automatic diagnosis task is, across a wide range of different conditions. Traditional fMRI research often approaches group comparisons using traditional regression models applied independently on each feature (brain connection), a technique called connectome-wide association study (CWAS). In this context, the most classic measure of “task difficulty” is so-called Cohen’s d estimate, which is the difference in average between two groups, relative to the standard deviation of the feature within-group. Before diving into multivariate ML techniques, whether applied for single-task or multi-task diagnosis, we wanted to clarify the effect size associated with each condition using CWAS. We would like to emphasise that there is no reason for CWAS effect sizes to match accuracy with ML tools, as was previously shown (Bzdok and Ioannidis 2019; Shmueli 2010; Lo et al. 2015). However, CWAS effect sizes are a common metric and will provide intuitive guidance for interpretation in all following chapters. Specifically, we implemented 19 CWAS for the following conditions: 15 CNVs and 4 idiopathic psychiatric diagnoses. For each condition, we estimated an effect size (along with a 95% confidence interval (CI) bound and empirical p-value), and also investigated the impact of different cross-validation schemes on the effect size estimates.

The results presented in this chapter were published in two studies: (Moreau et al. 2023) and (Moreau et al. 2022). AH designed and implemented these analyses, and wrote the relevant parts of the two manuscripts.

3.2 - Methods

3.2.1 - Connectome-Wide Association Studies

We conducted a CWAS for each of the 19 CNVs and psychiatric conditions included in our study, contrasting cases and their respective controls. Control subjects refers to individuals without a CNV for analysis investigating the effect of CNVs, and individuals without a diagnosis in analyses investigating effects of psychiatric conditions. We applied linear regression independently for each of the 2080 values of the connectome: the FC values were first z-scored based on the variance of the relevant control subjects, so the regression estimates can also be interpreted as z-scores, and then used as the dependent variable with the genetic or diagnostic status as the explanatory variable. Models were adjusted for sex, scanning site, head motion, age and global signal. Global signal was defined as the mean of the connectome, and was included in the analysis as it has been shown that global signal-adjusted FC profiles show stronger correlations with cognition (Jingwei Li et al. 2019) and reduce confounding effects in multisite studies (Yan et al. 2013). FC profiles were defined as the 2080 beta values of 2080

connections from each CWAS. The significance of beta values corrected for multiple tests using the Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg 1995) at a threshold of $q < 0.05$.

3.2.2 - Estimating Effect Size

We defined effect size on connectivity as the mean of the top decile of the absolute value of the 2080 beta values in the FC profile. We then used a cross-validation approach to perform sensitivity analyses for these effect size estimates, a bootstrap approach to provide a 95% CI, and permutation testing to provide empirical p-values.

Estimating Effect Sizes Using Cross-validation

We generated effect size estimates for each sample using 2-, 5-, and 10- k-fold cross-validation. For each condition we split the sample into K folds, stratified to keep a consistent ratio of case and control subjects. For K iterations we used all but one fold as a training sample, generated an FC profile, and identified the connections with beta values in the top decile. Then on the remaining independent test fold, we generated another FC profile and computed the mean effect sizes of connections identified in the training sample. The resulting effect size is the mean of the estimates across the K folds.

Bootstrap Procedure to Estimate 95th Confidence Intervals of Effect Size

We identified the 95% confidence intervals for the effect sizes using a bootstrap procedure. For each condition we generated the actual FC profile, identified the top decile connections, and computed their mean as described above. Then, we generated a bootstrap distribution of 5000 pseudo-FC profiles by resampling the same number of case and control subjects with replacement and generating an FC profile. In each pseudo-FC profile, we took the mean of the identified connections to form a distribution of 5000 effect sizes, from which we identified the 5% and 95% interval.

Permutation Testing to Estimate Significance of Effect Size

We computed an empirical p value for each condition by conducting a permutation test. For 5000 iterations, we shuffled the genetic or clinical status labels of the individuals, performed a CWAS, and calculated the effect size. We then estimated the empirical p value by calculating the frequency of obtaining an effect size equal to or greater than the original observation (Phipson and Smyth 2010).

3.3 - Results

3.3.1 - Effect sizes of neurodevelopmental and psychiatric conditions follow a spectrum from small to large

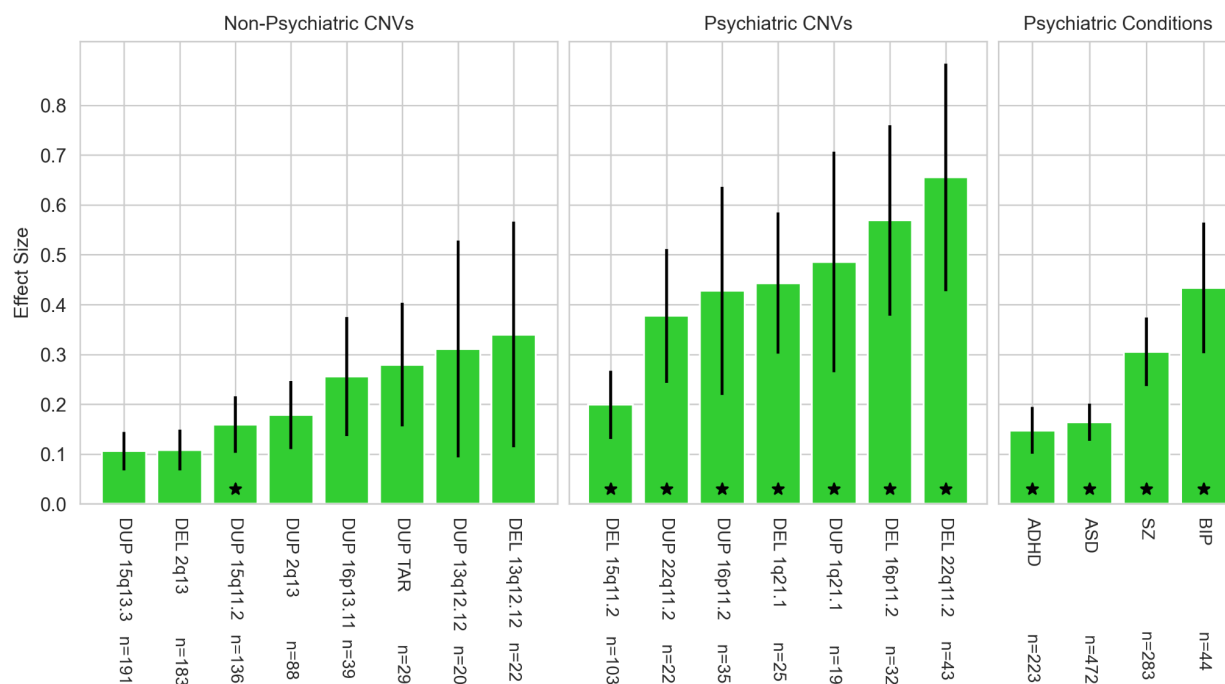


Figure 5 - The effect size and 95% CI of each condition on FC. Effect size is defined as the mean of the top decile (MTD) of the connections in the FC profile. The x axis represents the different conditions included in the study, and the y axis represents effect size. The MTD is plotted in green, while the 95% CI is indicated by a black error bar. Columns are marked with a star if the empirical p-value was found to be significant (< 0.05).

We generated a FC profile for each of the 19 CNVs and psychiatric conditions, identified the top decile connections and took the mean as the effect size, then calculated a 95% CI bound using a bootstrap procedure and an empirical p-value using permutation testing. The DEL 22q11.2 FC profile showed the largest effects (0.65), followed by DEL 16p11.2 (0.57). DEL 15q11.2 showed the mildest effects among CNVs associated with neurodevelopmental and psychiatric disorders. Effect sizes were largest for CNVs associated with neurodevelopmental and psychiatric disorders (mean 0.26), followed by psychiatric conditions (mean 0.26), and finally nonpsychiatric CNVs (mean 0.22). Each of the psychiatric conditions and CNVs, as well as DUP 15q11.2, were found to have significantly altered FC on the profile level using permutation testing. The psychiatric conditions and CNVs, except ADHD, had significantly altered FC on the individual connection level after FDR correction ($q < 0.05$) (see Table effect_size in supplementary materials). Overall, effect sizes of neurodevelopmental and psychiatric conditions followed a spectrum from small to large, with psychiatric CNVs, SZ and BIP having the largest effect sizes.

3.3.2 - Effect sizes of neurodevelopmental and psychiatric conditions are robust to cross-validation

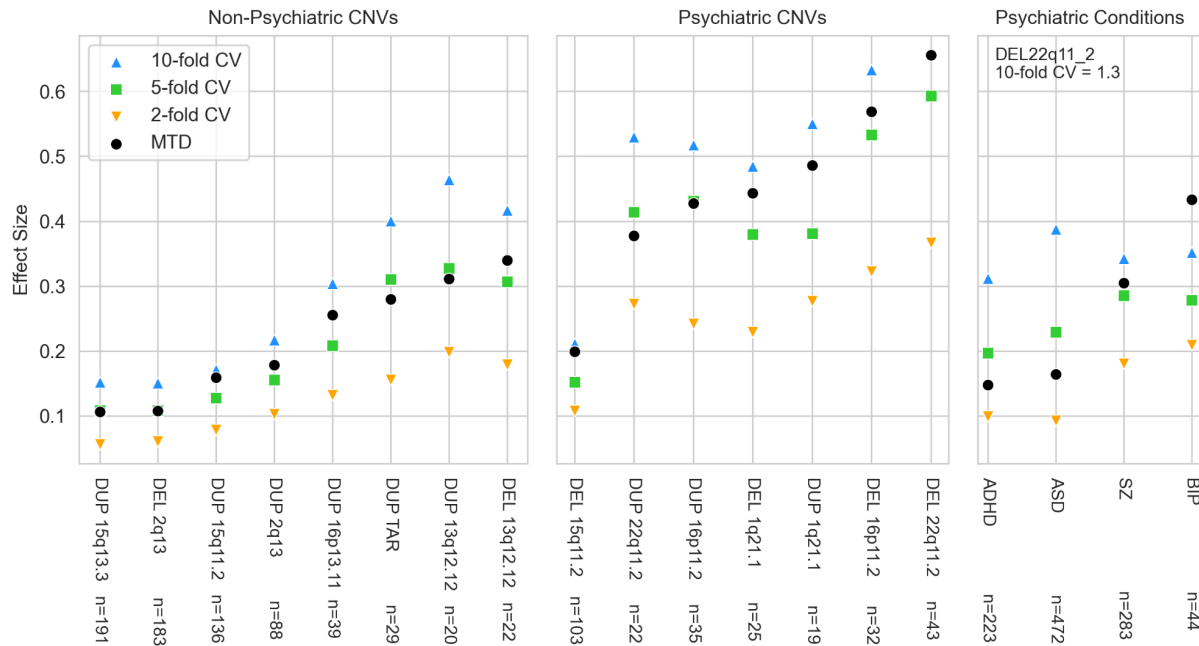


Figure 6 - Cross-validation of the effect size of each condition on FC. Effect size is defined as the mean of the top decile (MTD) of the connections in the FC profile. The x axis represents the different conditions included in the study, and the y axis represents effect size. The MTD is plotted in black, while the effect size under 2-, 5- and 10-fold cross-validation is plotted in orange, green and blue respectively. The effect size of DEL 22q11.2 under 10-fold cross-validation was far out of range and is noted above the plot.

We aimed to evaluate if the effect sizes for each of the 19 CNVs and idiopathic conditions were robust under cross-validation. We implemented 2-, 5-, and 10-fold cross-validation by splitting the case and control groups and generating a FC profile to identify the top decile connections and another independent FC profile in which to take the mean of the same connections. With the exception of BIP, non-cross validated effect size estimates were within the range of cross-validated estimates. The definition of effect size for CWAS we introduced thus appeared to be robust to overfitting, and meaningful estimates can be derived even in the absence of careful cross-validation.

3.4 - Discussion

Applying CWAS on a range of psychiatric and genetic conditions, we observed a wide range of effect sizes. The psychiatric CNVs, as well as some psychiatric conditions (BIP, SZ) had large effect sizes on brain connectivity. Non-psychiatric CNVs had a small effect size on brain connectivity. Finally, ASD and ADHD had small-to-moderate effect sizes.

Effect sizes on FC across psychiatric CNVs and psychiatric conditions are consistent with those reported for structural MRI measures (Assem et al. 2020; Moreau, Ching, et al. 2021). Previously reported effect sizes on cortical thickness for ASD (Cohen's $d = 0.21$) (Van Rooij, Anagnostou, and Arango 2018), BIP (Cohen's $d = -0.35$) (de Zwarte et al. 2019) and SZ (Cohen's $d = -0.5$) (van Erp et al. 2018), were also similar in magnitude to what we observed for FC. Regarding ADHD, our results are consistent with reported effect sizes from two ENIGMA (Enhancing Neuro Imaging Genetics through Meta Analysis) Consortium studies (Cohen's d between -0.21 and 0.19) examining global intracranial and subcortical grey matter volumes, total surface area, and cortical thickness (Hoogman et al. 2017, 2019).

Effect sizes of CNVs on structural and functional measures of the brain have been reported to be two to five times higher than those of psychiatric conditions (Modenato et al. 2021; Moreau, Ching, et al. 2021; Moreau, Raznahan, et al. 2021; Moreau et al. 2020; Sønderby et al. 2022), which is line with what we observed. While there is some controversy on the impact of small sample sizes on estimates of effect size (Schäfer and Schwarz 2019), our estimate for the large effect size of DEL 22q11.2 replicates findings in a much larger sample of 475 carriers (Cohen's $d = -1$ for surface area and $d = 0.6$ for cortical thickness) (Modenato et al. 2021). In our published study (Moreau et al. 2023), we additionally found a correlation between the effect size of CNVs on FC and their previously reported effect size on cognitive ability (Huguet et al. 2018) and general risk for neurodevelopmental and psychiatric disorders (Sanders et al. 2015; Moreno-De-Luca et al. 2013; Marshall et al. 2017), lending further support to the accuracy of the results.

Chapter 4 - Benchmark Study

4.1 - Intro

In this chapter we begin our investigation of the mega dataset using ML methods, which has never been done for any of the CNVs, starting with widely used simple classifiers before looking at neural networks and multi-task learning in later chapters. In this setting, we aim to classify each subject as either a CNV carrier or not, or as diagnosed with a psychiatric condition or not. The model is evaluated in an independent test sample after being exposed to the training sample. This is a more difficult task than performing group level comparisons, as the model must learn information that can distinguish individual subjects rather than estimating the significance of a difference in group averages.

We use class-balanced datasets sub-selected from the mega dataset for the CNVs to avoid incentivising a model to learn solely based on class sizes, since in the mega dataset a model could reach nearly 100% accuracy simply labelling each subject as a control due to the very large group of control subjects coming from the UKBB. This process of selection is described in detail in Chapter 2.

We take confounding factors into account in the estimate of the performance accuracy of each model for each condition, by comparing a model trained on confounds alone to another trained on both confounds and connectomes. This approach effectively provides an estimate of how much accuracy can be attributed to confounds. Differences across sites of data collection are the most troubling confounding factor in our large heterogeneous dataset, and generalising to a new site of data is an important task to measure the capacity of a model to be applicable in a clinical setting.

We first aimed to evaluate if different conditions included in the benchmark could be diagnosed using classic ML approaches on connectomes, when rigorously controlling for site effects. We next aimed to evaluate if the automatic diagnosis of conditions included in the benchmark could generalise to new sites of data collection. In this setting we expected cases with large sample sizes to generalise better than cases with little available data.

While there is no direct link between effect size and accuracy of automatic diagnosis (Shmueli 2010; Bzdok and Ioannidis 2019; Lo et al. 2015), we expected the accuracy to reflect effect size in general as it is a measure of task-difficulty. Specifically, we expected psychiatric conditions to be predicted within range of commonly reported accuracies in the field and for psychiatric CNVs to be well predicted. While we expected non-psychiatric CNVs to be difficult or impossible to distinguish.

4.2 - Methods

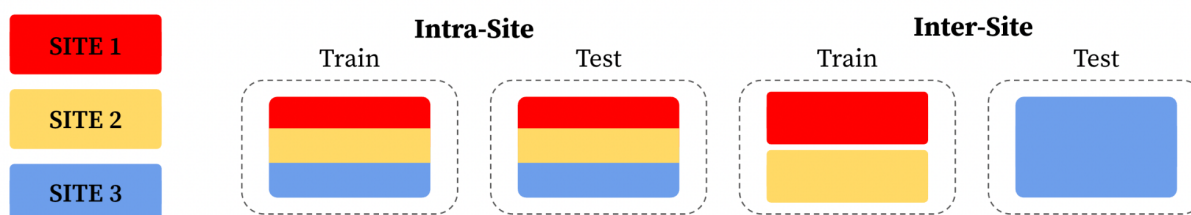


Figure 7 - Intra- and inter-site cross-validation.

4.2.1 - Cross-validation

We compared two cross-validation strategies that account for site effects (Orban 2018), as illustrated in Figure 7.

- In intra-site cross-validation, the model is exposed to identical sites of data collection during training and testing. Specifically, five random folds of training and test groups are built such that they have roughly the same proportion of cases and controls for each site. Both the training and test groups feature every available site at each fold. The reported accuracy is the average of the model performance across all folds.
- In inter-site cross-validation the model is trained on data from all but one site and tested on the left out site. In this approach, there are as many folds of cross-validation as there are sites of data collection. Since there is a large variation in the number of subjects across sites, the reported accuracy is the average across folds, weighted by the sample size of each site.

4.2.2 - Confound variables

In order to evaluate the added predictive value of connectomes for automated diagnosis, we constructed a baseline model for prediction composed solely of so-called confound variables. These variables included age, head motion, global signal, scanning site and sex. We assessed the diagnostic performance of the confound model alone, and then compared this baseline model to a full model including confound variables alongside connectomes.

4.2.3 - Classifiers

We used six classic ML techniques implemented in scikit-learn that performed well in the (Dadi et al. 2019) benchmarking paper and represented a range of underlying strategies. Support Vector Classifier (SVC) (linear kernel, $C=100$, and L_2 penalty), Logistic Regression (LR) (L_2 penalty), Ridge Regression (Ridge), Gaussian Naive Bayes (GNB), Random Forest (RF) and k-Nearest Neighbours (kNN) ($k=1$).

4.3 - Results

4.3.1 - Only large effect CNVs and psychiatric conditions can be predicted above chance level when controlling for site effects

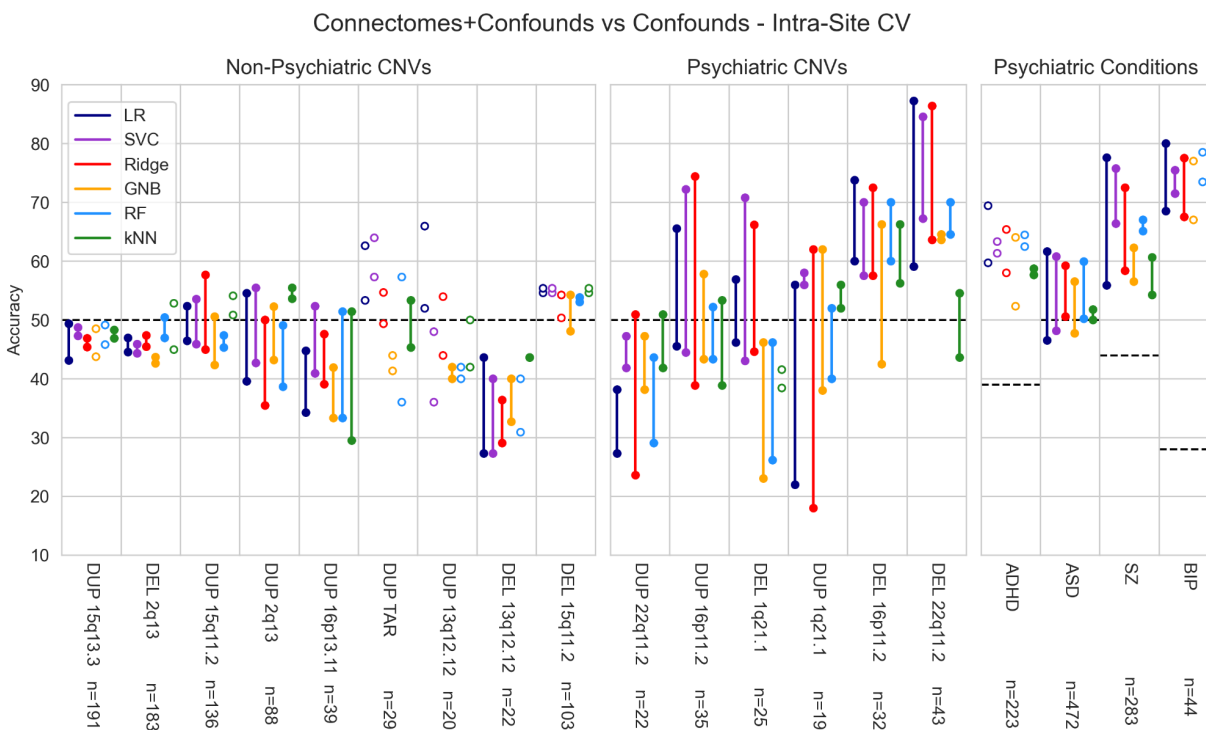


Figure 8 - Performance accuracy of automated diagnosis using intra-site cross-validation. The x axis represents the different datasets and diagnostic tasks included in this benchmark, along with total sample sizes (number of subjects). The y axis shows the accuracy of prediction using 6 different ML techniques. For each dataset and technique, the top point shows prediction using both connectomes and confounds, while the bottom point shows prediction for confounds only. If the connectome prediction outperformed the confounds, points are filled and connected by a line, and otherwise they are not. Chance level of prediction is indicated by a dashed line for each condition.

We first aimed to evaluate if different conditions included in the benchmark could be diagnosed using classic ML approaches on connectomes, when rigorously controlling for site effects. To this end, we trained six different ML algorithms for automated connectome-based diagnosis on a collection of 19 psychiatric or neurodevelopmental conditions, carefully stratifying for each site of data collection (intra-site cross-validation) and comparing the performance of each algorithm to a dummy model based solely on confounding effects. ML models were trained on each condition independently, as is common in the fMRI biomarker literature. We observed that only CNVs associated with psychiatric conditions and most idiopathic conditions could be predicted above chance levels, with the notable exception of DEL15q11.2, DUP22q11.2 and ADHD. For all these conditions, adding connectomes to the prediction model led to substantial increase in accuracy of the diagnosis, compared to using confounds only. DEL22q11.2 reached

the highest accuracy, close to 90% with LR and Ridge, while several other conditions reached over 70% accuracy (SZ, BIP, DEL 16p11.2, DEL 1q21.1, DUP 16p11.2). Overall, standard ML models seem capable of automatically diagnosing most psychiatric CNVs and idiopathic conditions, sometimes with fair accuracy, but not non-psychiatric CNVs.

4.3.2 - Datasets featuring a large number of sites (over 6) generalised to unseen sites

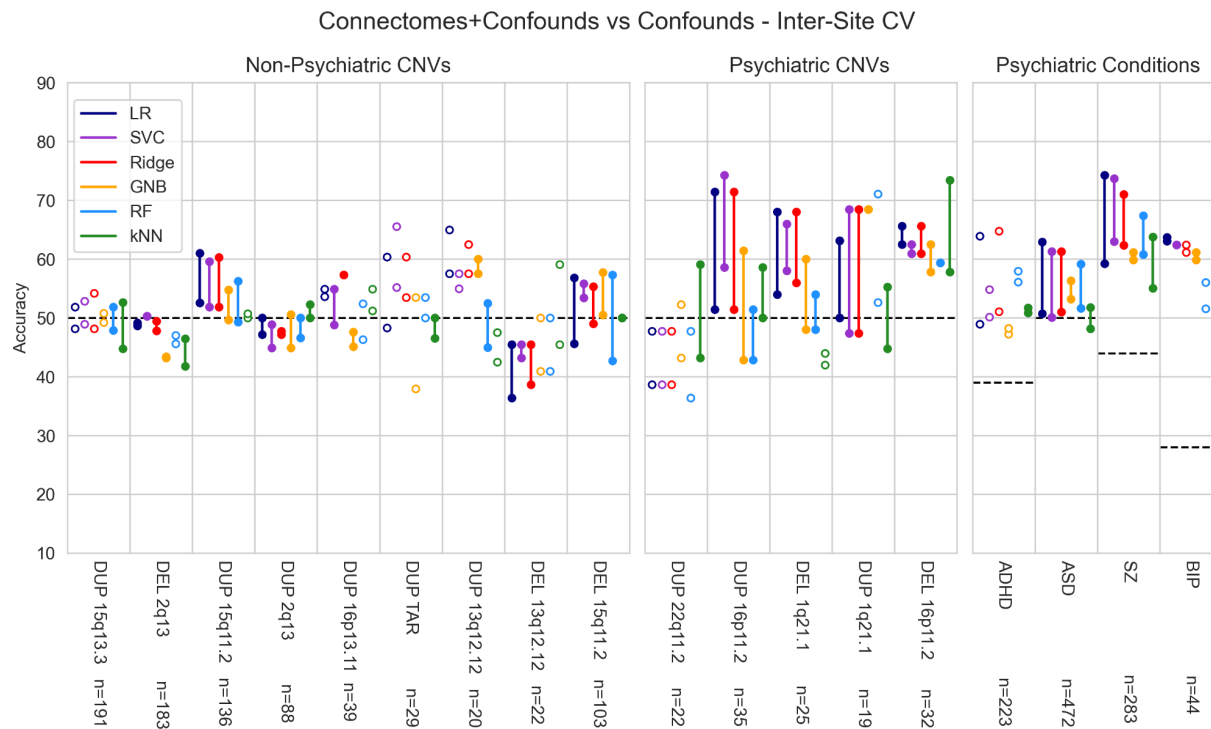


Figure 9 - Performance accuracy of automated diagnosis using inter-site cross-validation. The x axis represents the different datasets and diagnostic tasks included in this benchmark, along with total sample sizes (number of subjects). The y axis shows the accuracy of prediction using 6 different ML techniques. For each dataset and technique, the top point shows prediction using both connectomes and confounds, while the bottom point shows prediction for confounds only. If the connectome prediction outperformed the confounds, points are filled and connected by a line, and otherwise they are not. Chance level of prediction is indicated by a dashed line for each condition.

We next aimed to evaluate if the automatic diagnosis of conditions included in the benchmark could generalise to new sites of data collection. Using the same ML techniques on connectomes trained independently for each condition, we trained the classifiers on data from all but one site of collection and evaluated the classifier on the left out site (inter-site cross validation), again comparing the performance of each algorithm to a dummy model based solely on confounding effects. DEL22q11.2 was excluded from this analysis as the available data was collected entirely at one site. A number of conditions could not be diagnosed with substantially better accuracy than the dummy models using inter-site cross-validation. Some of those

conditions were already at chance level with intra-site cross-validation, however. For the other, above-chance level models, only BIP and DEL16p11.2 failed to generalise using inter-site, and those conditions featured a low number of sites: 2 and 5, respectively. The remaining above-chance conditions (SZ, ASD, DUP 16p11.2, DEL 1q21.1, DUP 1q21.1) featured between 6 and 28 sites of data collection, and performed as well or better than with intra-site cross-validation. Notably, DEL 1q21.1 (n=25) and DUP 1q21.1 (n=19) were able to generalise even while having a smaller sample size than DEL 16p11.2 (n=32) and lower prediction accuracy in the first setting. Overall, the heterogeneity of the data available for training in terms of the number of sites of data collection is crucial for successful automated diagnosis generalising to data collected at new sites.

4.4 - Discussion

We applied common ML models to automatically diagnose a wide range of CNVs and psychiatric conditions using intra- and inter-site cross-validation. Using intra-site cross-validation, we observed that most psychiatric conditions and CNVs associated with a psychiatric condition could be predicted above chance accuracy, but not CNVs without such an association. Inter-site cross-validation revealed that the heterogeneity of the data, not only the sample size, is crucial for generalising to data collected at new sites.

Regarding prediction using intra-site cross-validation, the accuracies obtained for the psychiatric conditions are on par with what is found in the literature. In their review of individual subject prediction of psychiatric conditions, Arbabshirani (Arbabshirani et al. 2017) reported accuracy values published in the literature using rs-fMRI data ranging from 70-91% for ASD, 62-100% for SZ and 54-90% for ADHD. However, the highest accuracies reported are likely misleading due to small sample sizes. Nielsen and colleagues (Nielsen et al. 2013) estimate that average reported prediction accuracy for ASD using single site rs-fMRI data is 80%, but when they attempted to replicate those results using the full multi-site ABIDE I dataset they obtained 60% - which is line with our accuracy using ABIDE I and II. Our accuracy for SZ fell within the wide range reported by Arbabshirani, and we highlight a similar multi-site study by Zeng and colleagues (Zeng et al. 2018) who reported obtaining 85% using data pooled across sites (n = 734, including 357 SZ from seven sites). Regarding BIP there are very few examples of classification studies using rs-fMRI, Wang and colleagues (H. Wang et al. 2022) achieved 83.7% accuracy predicting people with a diagnosis vs typical control, and Wang and colleagues (Y. Wang et al. 2020) achieved 80.5%, which are close to the accuracy we reported. While our prediction accuracy for ADHD fell within the reported range, the model including both connectomes and confounds failed to outperform the confounds only model. This result is supported by evidence from the ADHD-200 competition, in which the highest prediction score of clinical diagnosis was achieved using phenotypic data alone, including age, and in the absence of brain data (ADHD-200 Consortium 2012).

Obtaining standard prediction accuracy on psychiatric conditions validated our pipeline and lends credibility to our results in automatically diagnosing CNVs, for which this is the first study to date to our knowledge. The prediction accuracy for CNVs broadly follows the trend of clinical effect size seen in chapter 3. However, psychiatric conditions can be predicted better than effect size would suggest relative to CNVs, implying that prediction of CNVs could be improved with larger sample sizes. It was already noted that effect size in univariate regression models do not necessarily align well with performance of multivariate classifiers (Bzdok and Ioannidis 2019; Shmueli 2010; Lo et al. 2015). The failure to beat chance level when predicting CNVs with no association to a psychiatric condition serves as a negative control, further lending credibility to our results. Notably, DEL 22q11.2 reached the best prediction accuracy, which was significantly higher than the easiest to predict psychiatric conditions.

Regarding prediction using inter-site cross-validation, we found that heterogeneity, and not simply sample size, is an important factor in generalising to data from new sites. This finding agrees with the results by Orban and colleagues (Orban et al. 2018). Using data from 191 people diagnosed with SZ and matched controls collected from six scanning sites, Orban found that increasing the heterogeneity of the training set by including data from different sites improved accuracy of the model when using inter-site cross-validation and verified that the effect was not attributable to increased sample size. Orban also found that inter-site classification reached similar accuracy to intra-site classification when 5 out of the 6 sites were used for classifier training, which is in line with our findings that prediction accuracy for conditions with a large number of sites (6 to 28) was comparable using intra- and inter-site cross-validation. There are few examples of multi-site prediction studies evaluating inter-site cross-validation. Zeng and colleagues (Zeng et al. 2018) obtained 81% using inter-site cross-validation predicting SZ in their dataset (described above). Hensfield and colleagues (Hensfield et al. 2018) obtained 65% using inter-site cross-validation on ABIDE I ($n = 1045$, including 505 ASD from 17 sites), while Abraham and colleagues (Abraham et al. 2017) report 67% also using ABIDE with fewer subjects ($n = 871$).

Chapter 5 - Confound-Isolating Cross-Validation

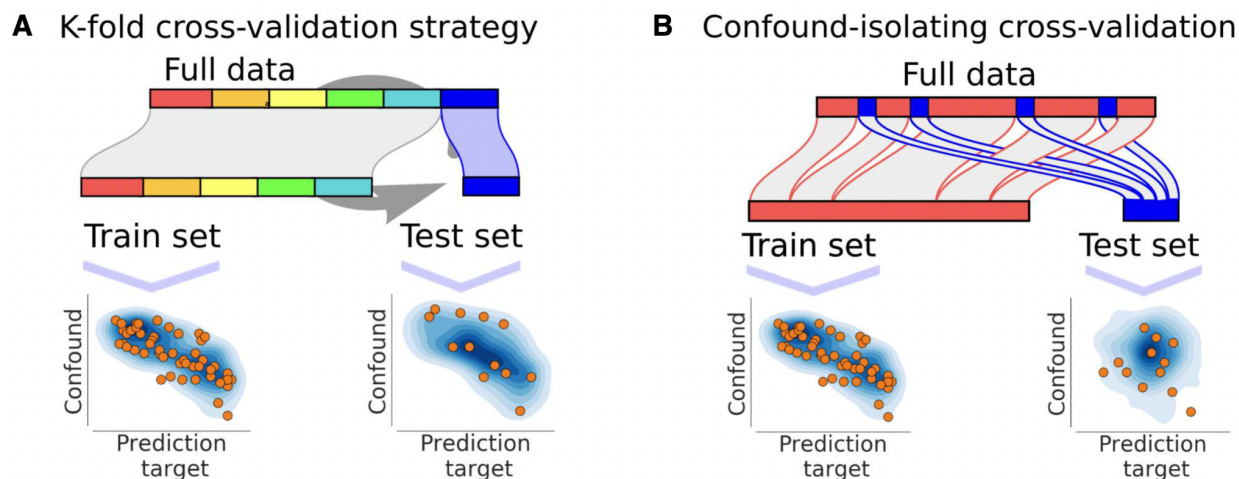


Figure 10 - Confound-isolating cross-validation. Image from (Chyzhyk et al. 2022).

5.1 - Intro

In this chapter, we introduce a strategy called confound-isolating cross-validation for managing confounding factors in the prediction of clinical conditions from connectomes. Previously, we used a simple intra-site cross-validation strategy. With such cross-validation, prediction accuracy may reflect biases in confound distribution across the condition groups. To address this issue, we compared the accuracy of two models: one trained on confounds and connectomes together, and a baseline model trained from confounds alone. This approach requires comparing two models with marked differences in the dimension of the input, which is much larger for the connectomes+confounds model. This difference in dimensionality is not an issue for the classifiers we used in chapter 4, thanks to their limited number of internal parameters and regularisation procedures. However, in the following chapters, we use MLPs, whose number of parameters quickly grows with the dimension of the input. For this reason, we chose to use models predicting conditions of interest from connectomes alone and not to include a baseline confounds model in these experiments. A connectome-only model may however leverage biases in confound distribution of a particular dataset in the prediction of clinical labels. In order to avoid this possibility, (Chyzhyk et al. 2022) proposed to use test sets that are balanced with respect to a single confounding factor, and termed that approach confound-isolating cross-validation. While we apply a different algorithm to adapt the method to the case with multiple confounds, we keep the terminology as we feel it succinctly communicates the aim. In the ideal case, in which a test set of subjects is perfectly balanced, the prediction accuracy of a confound model should be exactly 50%, and thus any prediction in excess of that chance level can be considered free from confounding (Chyzhyk et al. 2022) (see Figure 10).

While the ideal case of confound-isolating cross-validation is attractive, in reality it may not always be possible to create perfectly balanced test sets. In this chapter, we applied established approaches combined with brute force search in order to create approximately balanced test sets for each condition used in our benchmark. We demonstrate the effectiveness of confound-isolating cross-validation with these balanced test sets by showing that prediction accuracy of confounds-only models was near chance level. We next aimed to compare the performance of diagnosis based on connectomes only, using confound-isolating vs intra-site cross-validation. We hypothesised that the prediction accuracy would decrease with confound-isolating cross-validation, specifically in the conditions where the confound-only model performed well, as some of the prediction accuracy of the confounds+connectome models could be attributed to confounds biases.

In this and the following chapters, we limit the study to conditions that were predicted above chance in benchmark study: psychiatric conditions except ADHD and psychiatric CNVs except DEL15q11.2.

5.2 - Methods

5.2.1 - Automated balanced test set generation

In order to create test sets that were balanced with respect to confounds (age, head motion, global signal, scanning site and sex) we used a hybrid of two methods: General Class Balancer (described in section 2.2) and propensity score matching as implemented in `pymatch` which was written in support of (Miroglio et al. 2018). Propensity score matching first calculates a so-called propensity score as the output of a logistic regression model aiming to predict the class of interest from the confounds. Intuitively, subjects with the same propensity score have confound characteristics that are equally suggestive of the class of interest. Matching is then implemented iteratively by pairing subjects in the condition group with corresponding controls featuring the closest available propensity score, up to a threshold (Miroglio et al. 2018; Inacio et al. 2015).

5.2.2 - Iterative generation of test sets

Since propensity score matching doesn't exactly match categorical confounds, our preferred method to generate balanced test sets was General Class Balancer. This approach produced viable test sets in most conditions (ASD, BIP, SZ, DEL 22q11.2, DUP 22q11.2, DUP 16p11.2, DUP 1q21.1). For these conditions, we created 5 unique test sets for each dataset with the following procedure:

1. identify a balanced subset of the dataset (test set),
2. check if there was enough remaining data for a training set (50-80% of original data),

3. check if the subset was a duplicate of a previous selection.

Whenever the steps could not be completed successfully, a new candidate test set was generated, which was possible as General Class Balancer produces different results for different seeds of the random number generator. This whole procedure was repeated several times with different parameters of General Class Balancer (p-value cutoff for the mann-whitney U-test and minimum number of members in the test set) and thresholds for minimum amount of data left in as a training set, and we selected the final five folds as the ones where the accuracy of the confounds-only model was closest to chance in the benchmark.

DEL 1q21.1 and DEL16p11.2

General Class Balancer failed to find acceptable test sets in two conditions (DEL 1q21.1 and DEL16p11.2). In this case, we used propensity score matching and explicitly used ML prediction accuracy to evaluate and select test sets, by adding one step in our procedure. Specifically, we rejected test sets that reached an average accuracy outside 40-60% or maximum individual accuracy above 60%, as explained in the next paragraph.

5.2.3 - Evaluating test sets

A perfectly balanced test set would mean that a classifier trained to predict the condition based on the confounds would not outperform chance accuracy. Regardless of what it is possible to learn from confounds on the remaining data (training set), it should not generalise to the test set in which the classes are balanced with respect to those variables. In order to evaluate this expected behaviour, we repeated a variant of the benchmark study: for each condition we predicted from confounds using the same six ML classifiers and performed cross-validation using the five test sets.

5.2.4 - Connectomes alone benchmark

For each condition we predicted from connectomes alone using the same set of classifiers evaluated by confound-isolating cross-validation. We then compared these results to the previously established connectomes+confounds and confounds-alone baseline models, using intra-site cross-validation.

5.3 - Results

5.3.1 - Confounds models with balanced test datasets predict near chance level for most conditions

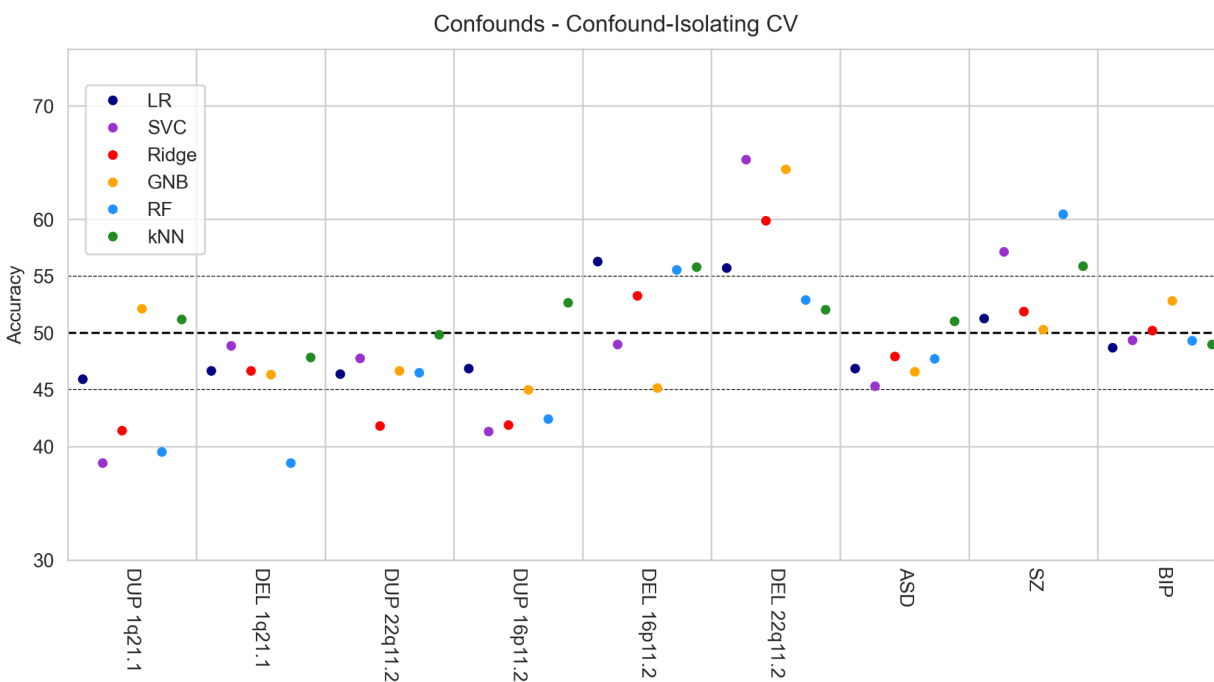


Figure 11 - Performance accuracy of automated diagnosis from confounds using confound-isolating cross-validation. The x axis represents the different datasets and diagnostic tasks included in this benchmark. The y axis shows the accuracy of prediction using 6 different ML techniques. Chance level of prediction (0.5) is indicated by a bold dashed line, the ideal range of 0.45-0.55 prediction is indicated by fine dashed lines.

We aimed to evaluate the impact of creating test sets for each condition that are balanced with respect to confounds, rather than mirroring the biases of confounds in the original sample. We first evaluated these balanced test sets in terms of the prediction accuracy of the confounds-only models, expecting them to perform at chance level. Five balanced test sets were generated for each condition using an intra-site cross-validation design, along with their complementary, unbalanced training sets. We then trained six ML classifiers on confounds alone and evaluated their performance on the balanced test sets, independently for the 9 conditions that could be diagnosed above chance in the intra-site benchmark. Prediction accuracy was below 60% for all conditions except SZ and DEL22q11.2. The most notable improvement was achieved on BIP, where prediction from confounds was indistinguishable from chance with balanced test sets, whereas in the previous intra-site benchmark, confounds alone reached nearly 80% accuracy. While imperfect (SZ, DEL22q11.2), balanced test sets can provide a means of evaluating automatic diagnosis reducing the bias from confounding factors.

5.3.2 - Diagnosis from connectomes alone reaches similar performance on balanced test sets than traditional intra-site cross-validation

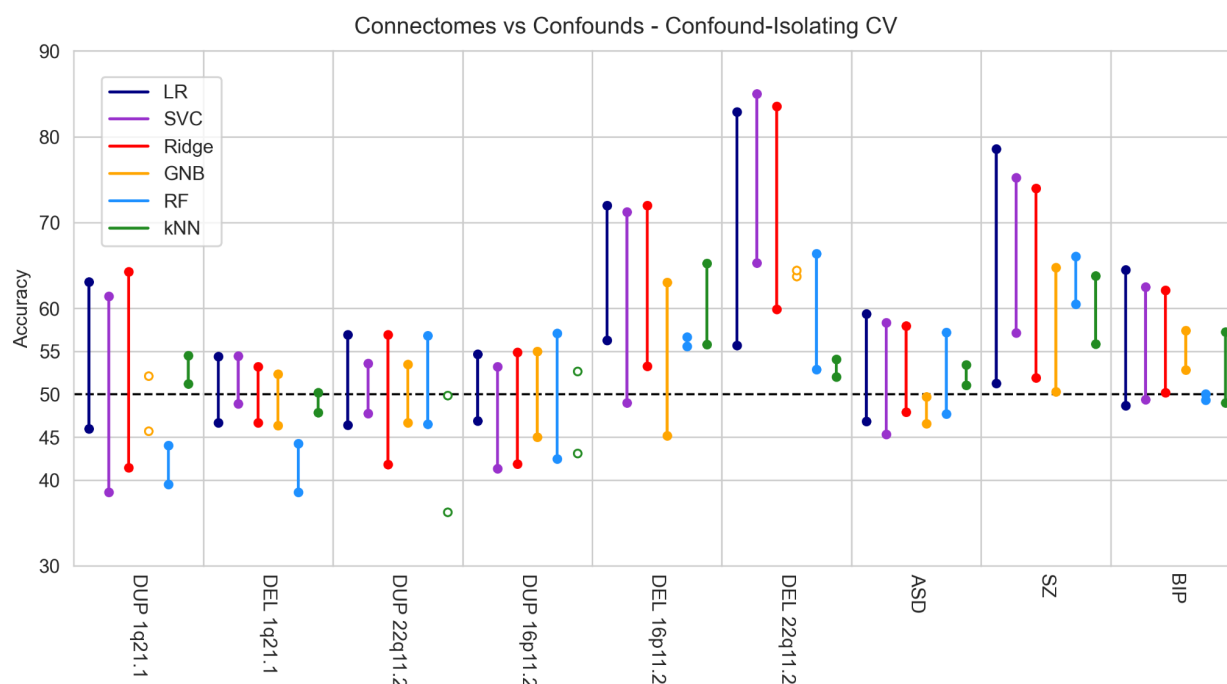


Figure 12 - Performance accuracy of automated diagnosis from connectomes using confound-isolating cross-validation. The x axis represents the different datasets and diagnostic tasks included in this benchmark. The y axis shows the accuracy of prediction using 6 different ML techniques. For each dataset and technique, the top point shows prediction using connectomes, while the bottom point shows prediction for confounds only. If the connectome prediction outperformed the confounds, points are filled and connected by a line, and otherwise they are not. Chance level of prediction (0.5) is indicated by a dashed line.

We next aimed to evaluate the performance of diagnosis based on connectomes only using confound-isolating cross-validation rather than intra-site cross-validation, as was done in our initial experiments. We trained six ML classifiers for this purpose, independently for each of the 9 conditions that could be diagnosed above chance in the intra-site benchmark. We observed that prediction with confound-isolating cross-validation reached very similar performance to what was achieved with intra-site cross-validation for most cases. In particular, the decrease in accuracy of automatic diagnosis for BIP matches the loss in prediction accuracy from confounds alone. There were two notable exceptions, DUP16p11.2 and DEL1q21.1, which decreased by roughly 10% accuracy without a major change in prediction accuracy from confounds. Overall, automatic diagnosis of conditions using connectomes alone on balanced test sets reached similar performance to those achieved with intra-site cross-validation.

5.4 - Discussion

We first demonstrated that it was possible to build an effective confound-isolating cross-validation for each condition, except for two cases, using a combination of established methods and exploration of parameters and random seeds to manually select good test sets. We next evaluated prediction from connectomes alone with confound-isolating cross-validation and found that performance was diminished for certain conditions but mostly led to similar results as the connectome+confounds model and intra-site cross-validation.

Regarding the creation of the test sets, failure to find a well balanced test set for DEL22q11.2 is surprising since the data was collected at a single site, however this highlights how difficult it is to recruit a sample of subjects balanced for age and sex for a rare condition (see Figure 19 in supplementary materials). Failure to find balanced test sets for SZ is surprising because of the large sample size, BIP in comparison with far fewer subjects had subsets that balanced almost perfectly. BIP however only has 2 sites to match across, while SZ has 12. ASD with many more sites than SZ (28) doesn't have the complication of matching for gender (all subjects are male). The gender ratio of each site in the SZ dataset is far from 50% (see Figure 20 in supplementary materials). This highlights the extent of heterogeneity in multi-site datasets and the challenge of isolating the effect of the condition in the presence of persistent confounding factors. It should be noted that the vast majority of ML works in rs-fMRI do not study the impact of confound biases, and rely on intra-site cross-validation on full, unselected samples. Leming and colleagues (Leming, Górriz, and Suckling 2020), who introduced the General Class Balancer we used, apply confound-isolating cross-validation in the prediction of ASD from rs-fMRI data, but do not report on the characteristics of the resulting test set. Chyzhyk and colleagues (Chyzhyk et al. 2022) show that their confound-isolating cross-validation algorithm successfully finds a test set in which the initial correlation between the confound and the target is removed in three settings: predicting age with motion as a confound in the CamCan and UKBB datasets, and predicting fluid intelligence with age as a confound in the CamCan dataset. We keep the heuristic evaluation of prediction accuracy from confounds rather than the correlation between the confounds and the target variable in the test sets, since balancing in the case of multiple confounds is more challenging than of a single confound (Chyzhyk et al. 2022).

Regarding the performance of automatic diagnosis using connectomes and confound-isolating cross-validation, we observed a loss in accuracy compared to a more standard intra-site cross validation scheme in a few cases, as expected. We also expected this loss to reflect the performance of the confounds-only model using intra-site cross-validation. The BIP condition did follow that predicted pattern, but this was not observed systematically. Chyzhyk and colleagues (Chyzhyk et al. 2022) found that, relative to results without controlling for confounds, using confound-isolating cross-validation reduced prediction of fluid intelligence on CamCan to chance level, significantly reduced the prediction accuracy of age on CamCan and only slightly that of age on UKBB (both predictions for age remained above chance level).

Leming and colleagues (Leming, Górriz, and Suckling 2020) do not compare the prediction accuracy they obtained using confound-isolating cross-validation with any other types of cross-validation.

Apart from losing prediction accuracy attributable to confounds, there are possibly more severe factors which stem from the requirements of finding balanced test sets for confound-isolating cross-validation. First, it takes more data to create a balanced sample than we would normally put in a test set (~40-50% of the dataset vs 25%), which leaves less available data for training. Next, by creating a balanced sample for our test set we are using excluded subjects as our training set. The training set is therefore necessarily unbalanced and different in distribution from the test set. When there is a large amount of well-sampled data available this problem is negligible, but in cases with small datasets it creates a significant hurdle. Snoek and colleagues (Snoek, Miletić, and Scholte 2019) examined a strategy similar to our implementation of confound-isolating cross-validation to predict brain size using gender as a confound. They first found a balanced sample of their dataset iteratively excluding outlying subjects until there was no correlation between brain size and gender, then split it into K-non-overlapping balanced subsets and applied K-fold cross-validation. While in this procedure there is no issue with unbalanced training data, they noted that excluding outliers amounts to removing examples on which the model would make poor predictions resulting in an inflated estimate of model performance. Although we created class-balanced datasets from the mega dataset for the CNVs, rather than excluding outlying subjects, our procedure aimed to find the closest matching control subject with respect to the confounds for each patient out of a very large pool of available controls (no patients were excluded). As a result, the CNV datasets were only roughly balanced with respect to confounds.

Chapter 6 - Multi-Task Prediction of Age and Sex

6.1 - Intro

In this chapter, we introduce our framework for multi-task learning and evaluate its performance in predicting simple targets (age and sex) across different sites of data collection. Multi-task learning is an ML framework in which multiple tasks are learned simultaneously by a shared model. When the tasks are well suited to each other, the model can leverage information shared across tasks which makes better use of data and the constraint of learning a pattern useful for multiple tasks improves generalizability. However, it can be difficult to decide which tasks to learn together (Standley et al. 2019), as incompatible tasks can compete and make multi-task learning deleterious for performance overall.

Therefore, before we delve into the complexity of multi-task learning across CNVs and psychiatric conditions, we evaluate the benefit of using multi-task learning where heterogeneity between tasks is only due to sites. We treat each site of data collection as a task and predict the same target (either sex or age) across them. We hypothesised that the model would be able to leverage data across tasks to learn a representation robust to inter-site noise that improves prediction overall. Each site of data collection consisted of subjects with markedly different age ranges (see Table 3 in section C.2), so we hypothesised this objective would be more difficult than sex prediction. For prediction of both sex and age we hypothesised that sites with small sample size would benefit from being trained with sites with large sample size. In contrast, for sites featuring a large number of subjects, we hypothesised the prediction performance would reach a plateau in accuracy (similar to the single task setting) and experience little benefit from multi-task learning.

6.2 - Methods

6.2.1 - Implementation

All models were implemented in Pytorch (Paszke et al. 2019). The code for multi-task learning was written using Snorkel (Ratner et al. 2017) as a reference.

6.2.2 - Architectures

The models used in this chapter are variants of a single multi-layer perceptron (MLP) architecture (see Figure 3), consisting of an MLP in which multi-task learning is implemented through hard parameter sharing. We elected to implement hard rather than soft parameter sharing, first because it is the most commonly used approach to multi-task learning and, second, because the reduction in parameters and hence capacity relative to single task learning

or soft parameter sharing is well suited to our high dimensional data. The first and second layers of the network (encoder) are shared and output an embedding that is common across tasks. The third layer is specific to each task and outputs either two values for binary classification (MLPconn) or a single value for regression (MLPconn_reg).

The MLPconn model is an MLP with the following configuration: 2080-256-64-2, as explained here. The input to the networks is a 1×2080 vector consisting of the upper triangular values of the symmetric connectome matrix, which is passed through two shared hidden layers with 256 and 64 units and finally to a task-specific output layer of 2 units for binary classification. Batch normalisation (Ioffe 2015) is applied after each hidden layer. In the single task setting all the layers of the network are specific to the given task.

The MLPconn_reg model is the same as MLPconn, but with the output layer modified for regression so that the configuration becomes: 2080-256-64-1.

6.2.3 - Training

We trained the multi-task learning model as follows for each epoch: first the batches of data are pooled across tasks and shuffled, next each batch is passed through the shared encoder and to the task-specific output layer it is associated with, the loss is calculated and back propagated through the task-specific and shared layers, finally the gradients are clipped to have a maximum magnitude of 1. In the single task setting, the training followed the same procedure except that the batches of data were not pooled across tasks and were fed through a fully task-specific network. We used small batch sizes (8) since we included small datasets, and models were trained for 100 epochs, roughly 50 epochs past plateau observing plateaus in the single task setting. We used the Adam optimizer (Kingma and Ba 2014)), Leaky ReLUs as an activation function, and dropout regularisation (Srivastava et al. 2014) with the default parameters (Paszke et al. 2019). The binary classification tasks were scored with the cross-entropy loss after applying the softmax function, and the regression tasks with the mean squared error (MSE).

6.2.4 - Predicting Sex & Age

We created the datasets by treating each site as a task with a common prediction target of either age or sex. We used only the control subjects from each site of data collection which had at least 30 participants. The datasets corresponding to the three sites of collection from the UKBB study are enormous (sample sizes 4569, 7943 and 17673), so we subsampled 1000 subjects from each, which dwarfs the next largest site (UCLA_DS1 with 94 subjects), for computational reasons. For the sex prediction task, we excluded sites that had insufficient female subjects (NYU, SZ1, SZ2, USM). The prediction was performed from the connectomes alone and evaluated using 5-fold cross validation with randomly split train and test sets. To predict sex, the MLPconn model was

used first in the single task setting to establish a baseline and then in the multi-task setting with all sites pooled together. For predicting age, the MLPconn_reg model was used again in single task and then multi-task across sites.

Site	N	Female	%
ADHD1	54	35	65
ADHD3	56	26	46
ADHD5	77	39	51
ADHD6	39	18	46
HSJ	39	25	64
NYU	66	0	0
SZ1	42	3	7
SZ2	41	2	45
SZ3	31	15	48
SZ6	35	12	34
Svip1	48	18	38
Svip2	36	17	47
UCLA_CB	43	22	51
UCLA_DS1	94	43	46
UKBB11025	17673	9342	53
UKBB11026	4569	2504	55
UKBB11027	7943	4414	56
USM	30	0	0

Table 2 - Female subjects by scanning site. Number of total and female subjects, and percentage female by site.

6.3 - Results

6.3.1 - Multi-task Learning improves prediction of sex for a majority of sites

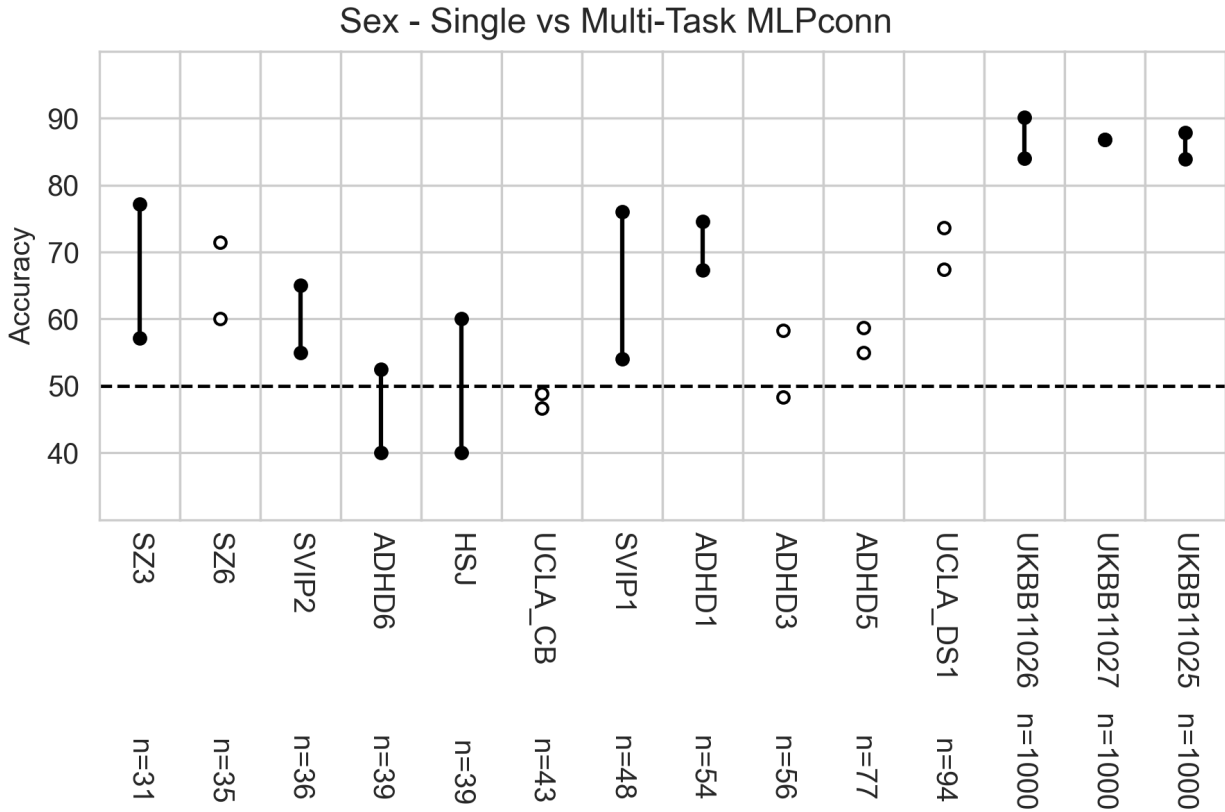


Figure 13 - Accuracy of sex prediction using single vs. multi-task learning. The x axis represents different sites of data collection included as prediction tasks, sites are ranked by sample size with the largest to the right. The y axis shows the accuracy of prediction using a neural network. For each task, the top point shows prediction using multi-task learning and the bottom point shows prediction on the task trained independently using the MLPconn architecture. If the multi-task prediction outperformed the single-task, points were filled and connected by a line, and otherwise they were not. Chance level of prediction (50) is indicated by a dashed line.

We aimed to evaluate the value of multi-task learning for heterogeneous data in a simple binary classification setting, where each site of data collection is treated as a task and the target of prediction, sex, is the same across tasks. Prediction accuracy improved for multi-task learning in a short majority of sites (9 out of 14). The mean accuracy in the multi-task setting (67.7) outperformed that of the single-task (62.8) substantially, but with much larger standard deviation (13.2 vs 11.7). Notably, prediction improved for the three sites with very large sample sizes from UKBB. Overall, multi-task learning across heterogeneous sites of data collection benefitted accuracy when the target of prediction was the same across tasks, but the effect was not systematic.

6.3.2 - Multi-task Learning improves prediction of age for a majority of sites

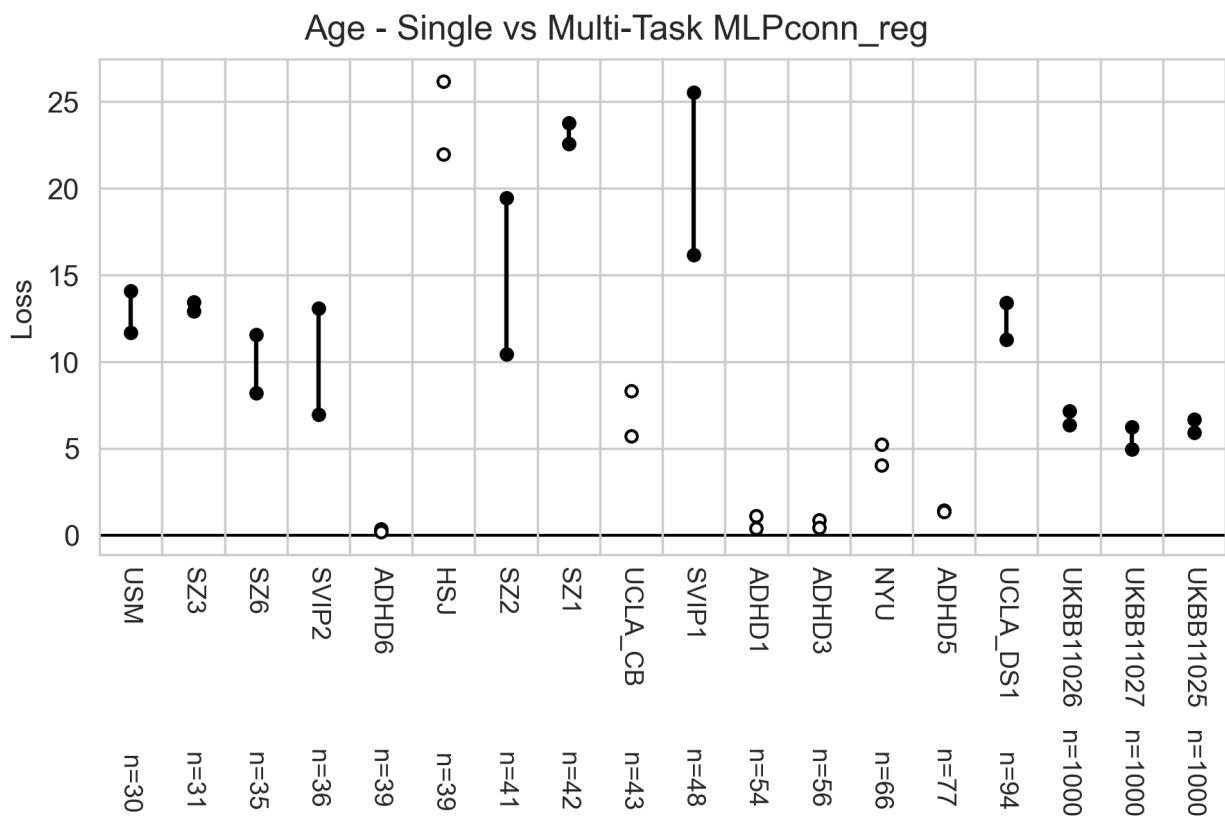


Figure 14 - Performance of age prediction using single vs. multi-task learning. The x axis represents different sites of data collection included as prediction tasks, sites are ranked by sample size with the largest to the right. The y axis shows the prediction loss using a neural network. For each task, the bottom point shows prediction using multi-task learning and the top point shows prediction on the task trained independently using the MLPconn_reg architecture. If the multi-task prediction achieved lower loss than the single-task, points were filled and connected by a line, and otherwise they were not.

We next aimed to evaluate if multi-task learning could improve prediction of age, once again using each site of data collection as a separate task. Each site of data collection consisted of subjects with markedly different age ranges (see Table 3 in section C.2), so we expected this objective to be more difficult than sex prediction. We used the same MLPconn architecture, modifying only the output for regression, and repeated the procedure, first training independent models on each site and then comparing them with a multi-task learning setting where all sites were trained concurrently. Prediction improved for a short majority of sites (11 out of 18). The mean loss in the multi-task setting (8.9 years²) outperformed that of the single-task (10.5 years²), but with larger standard deviation (10.47 vs 8.94). Again, prediction improved for the three sites with very large sample sizes from UKBB. Overall, multi-task learning benefitted prediction, even when the target of prediction was heterogeneously distributed across sites, but the effect was not systematic.

6.4 - Discussion

Applying multi-task learning across sites of data collection to predict age and sex, we observed that multi-task learning improved prediction relative to single task learning for a majority of sites. Contrary to our hypothesis, prediction always improved for sites with large sample size, whereas small sites showed varied performance that had no apparent relationship to sample size.

There are very few examples from the literature with which to compare our results. Ma and colleagues (Q. Ma et al. 2018) predict SZ across three sites using fMRI and show that prediction is improved relative to single task learning, however there is no impact of sample size since each site consists of 50 subjects diagnosed with SZ and 50 controls. Watanabe and colleagues (Watanabe et al. 2014) predict ADHD across seven sites using fMRI, but don't report on the sample size of the sites or the accuracy obtained per site. Hu and Zeng (Hu and Zeng 2019) predict SZ across three sites using structural MRI. They found that prediction improved for each site in multi-task vs single task learning, with the two smaller sites gaining more in accuracy than the largest (7% and 7% vs 3%). However, the differences in accuracy and sample size for each site were too close for any trend to be conclusive ($n = 269, 156, 325$, including $n = 137, 62, 144$ SZ for sites A, B and C respectively). Apart from prediction across sites, two other studies examined prediction of a common target across tasks (Xiao et al. 2020; Marquand et al. 2014), however both had the same amount of data for each task. Schulz and colleagues (Schulz et al. 2020) classified subjects into ten groups divided by sex and age using fMRI data in the UKBB with increasing sample sizes, and found that even with 8000 subjects and applying simple linear models the trend of prediction accuracy improving did not reach a plateau. In light of these results, the sites we consider to have large sample size ($n = 1000$) are still in the realm of small datasets for MLPs and rather than boost the tiny datasets ($n = 30-94$) they seem to simply dominate the training. Further studies are needed to assess the impact of sample size, as well as other factors such as heterogeneity of the samples, on multi-task learning prediction of a target across sites.

Chapter 7 - Multi-task Learning for Joint Diagnosis of CNVs and Psychiatric Conditions

7.1 - Intro

Having established a method for controlling for confounds (Chapter 5), and evaluated multi-task learning in a straight-forward setting (Chapter 6), we address our primary aim and apply multi-task learning to perform joint diagnosis across 9 CNVs and psychiatric conditions. As discussed previously (see sections 6.1 and 1.4), multi-task learning benefits from allowing similar tasks to share information during training.

The psychiatric conditions included in our dataset have high rates of comorbidity with each other as well as associations with the CNVs (see section 1.1). In addition to this, there is extensive overlap in genetic factors and symptomatology (Lichtenstein et al. 2009; Lee et al. 2013; Brainstorm Consortium et al. 2018; Taylor et al. 2021). Due to these relationships between conditions, we propose that the dataset we collected is a perfect candidate for multi-task learning. For example, the model might benefit from a feature that reflects social impairment learned from ASD that is also applicable to SZ, or a feature that reflects language impairment learned from DEL 16p11.2 that can help distinguish ASD. Huang and colleagues (Huang et al. 2022) reported promising results using multi-task learning across ASD, ADHD and SZ. Here we extend that work by examining a larger sample of conditions: including ASD, ADHD, and SZ, as well as BIP. Critically, we also incorporate rare CNVs with high effect size on both brain and behaviour in the multi-task learning study, that have never been previously studied in the ML context.

However, it can be difficult to determine a priori which tasks should be grouped together for multi-task learning, with similar tasks not necessarily learning better together (Standley et al. 2019). Here we test the hypothesis that the tasks included in our dataset overlap sufficiently to benefit from multi-task learning. We established a baseline by using an MLP for each task independently, and then train all the tasks concurrently in a model with a shared encoder (hard parameter sharing).

7.2 - Methods

The MLPconn model (architecture and training described in chapter 6) was used to predict each condition from connectomes alone first in the single task setting to establish a baseline, and then in the multi-task setting across conditions. The models were evaluated using confound-isolating cross-validation (described in chapter 5).

7.3 - Results

7.3.1 - Multi-task learning fails to improve automatic diagnosis across heterogeneous conditions

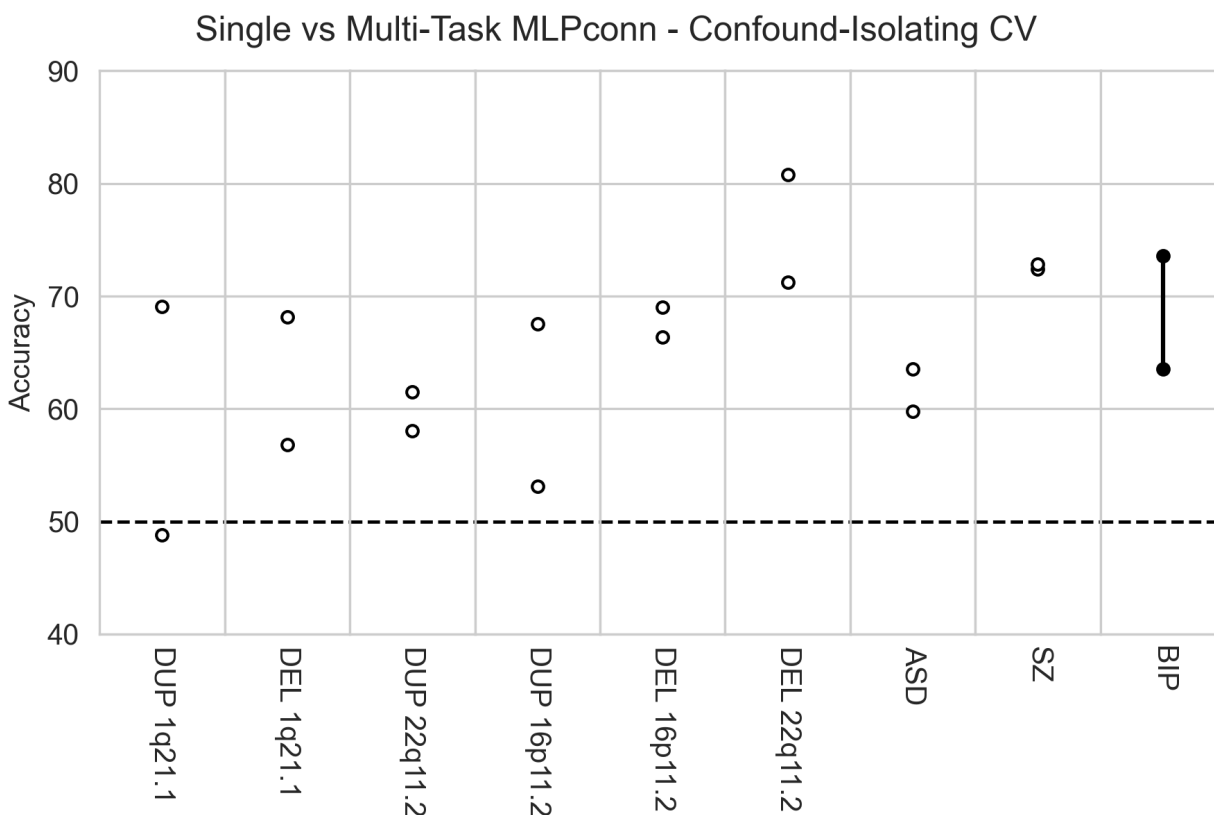


Figure 15 - Accuracy of automated diagnosis using single vs multi-task learning. The x axis represents different conditions included as prediction tasks. The y axis shows the accuracy of prediction using a neural network. For each task, the top point shows prediction using multi-task learning and the bottom point shows prediction on the task trained independently using the MLPconn architecture and confound-isolating cross-validation. If the multi-task prediction outperformed the single-task, points are filled and connected by a line, and otherwise they are not. Chance level of prediction (50) is indicated by a dashed line.

We aimed to improve automatic diagnosis of 9 CNVs and psychiatric conditions by leveraging shared information in datasets with limited sample size using a lightweight multi-task learning framework. First, we trained a simple feedforward neural network independently on each condition as a baseline and then trained a multi-task version of the model on all the conditions concurrently. Multi-task learning outperformed single-task learning for only BIP. For the remaining cases, multi-task learning suffered from negative transfer and performance accuracy actually decreased, contrary to our prediction. The performance of the neural network in the single-task setting outperformed the best of the simple ML models for ASD, DUP 16p11.2, DUP

22q11.2, DEL 1q21.1, and DUP 1q21.1. For the rest of the conditions, accuracy was in the range of that achieved previously.

7.4 - Discussion

Applying multi-task learning across CNVs and psychiatric conditions to perform automatic diagnosis, we observed that multi-task learning is deleterious for prediction accuracy overall. Contrary to our hypothesis, the tasks included in the concurrent training were too heterogeneous to benefit from multi-task learning implemented using hard parameter sharing.

The only existing studies in the literature to apply multi-task learning across conditions were conducted by Huang and colleagues (Huang, Liu, and Tan 2020; Huang et al. 2022). These studies proposed a soft parameter sharing model and compared it with MLPs in hard parameter sharing and single task learning. In the first study examining ASD and ADHD, they found that the hard parameter sharing model obtained accuracy that outperformed single task learning, but not their proposed model. However, in the follow-up study in which they added SZ they found that their proposed model still improved accuracy for each condition relative to single task learning, but that the hard parameter sharing model reduced prediction accuracy for ADHD and SZ. In light of these results, where negative transfer occurred with only three conditions, it is not surprising that multi-task learning failed to improve prediction accuracy for all but one of the 9 conditions included in our hard parameter sharing model. In the next chapter we thoroughly investigate the negative transfer we observed.

Chapter 8 - Negative Transfer Study

8.1 - Intro

In this chapter we investigate the negative transfer between tasks observed in the multi-task study on conditions. This is an exploratory analysis which was implemented as a post-mortem examination, following the failure of our original key hypothesis. Negative transfer is a phenomenon in multi-task learning in which instead of the desired outcome of similar tasks sharing a latent representation that benefits the learning of each, the tasks compete unproductively and interfere with each other's learning. Negative transfer can be mitigated through various soft parameter sharing schemes, in which parallel models for each task are linked and regularised together and allow for more flexibility between models at the cost of having a large number of parameters. However, we implemented hard parameter sharing - a stricter version of multi-task learning in which the models literally share a portion of their parameters - in this low sample size high dimensional data setting to reduce the capacity of the model and the potential of overfitting.

Since all the conditions were trained concurrently in the previous experiment, it is impossible to comment on the source of the negative transfer. We aimed to disentangle the negative transfer observed when predicting the 9 neurodevelopmental and psychiatric conditions concurrently in the multi-task setting by training the conditions together pairwise in different model and data settings. We varied the depth of the model (MLPconn_deeper) to increase its capacity, the input data to the model (MLPconcat) from connectomes alone to a concatenation of connectomes and confounding variables, and the type of layer (CNN) from fully connected to convolutional. This framework allowed us to see if certain conditions learned well or poorly together, as well as if the negative transfer behaved differently depending on the form of data input to the model or the model itself. We expected that some pairs of conditions learned consistently better or worse together regardless of data setting or model, and that conditions that have similarities in their effects on FC and symptoms would be learned better together.

8.2 - Methods

The implementation of the code and model training are as described in chapter 6.

8.2.1 - Architectures

MLPconcat

The MLPconcat model is exactly the same as the MLPconn model (described in chapter 6), with the input layer adapted to accept a concatenation of the upper triangular 1 x 2080 connectome

vector with the 1 x 58 confounds vector (age, head motion, global signal, scanning site and sex with categorical confounds one hot encoded). The result is an MLP model with configuration: 2183-256-64-2.

MLPconn_deeper

The MLPconn_deeper model is a version of the MLPconn model with two additional layers of width 64, one in the shared part of the model and another in the task specific part. The resulting configuration is 2080-256-64-64-64-2. The input to the model is the connectome vector.

CNN

The CNN model is adapted from (Leming and Suckling 2021). The input to the network is the upper triangle of the symmetric connectome matrix (2080 values) randomly permuted and formatted into a 40 x 52 matrix. The shared part of the model consists of a first convolution layer with 256 filters of shape 1 x 40 x 1, followed by two dense layers of 64 hidden units. The task-specific output layer has 2 units for binary classification. Batch normalisation (Ioffe and Szegedy 2015) is applied after each layer.

8.2.2 - Exploring Negative Transfer

In order to establish if certain tasks were learned better or worse together, we trained four different models to predict the conditions from the multi-task learning study first in the single task setting to establish a baseline, and then pairwise. The models evaluated were MLPconn which is the standard model, MLPconn_deeper which has two additional layers to vary model depth, MLPconcat modified for prediction from connectomes and confounds to vary the data context, and CNN a convolutional neural network with same depth as standard model to vary model type. The models were evaluated using confound-isolating cross-validation and Pearson's correlation was used to evaluate the similarity of the results across the four models.

8.3 - Results

8.3.1 - Negative transfer between conditions trained pairwise is stable across models and data settings

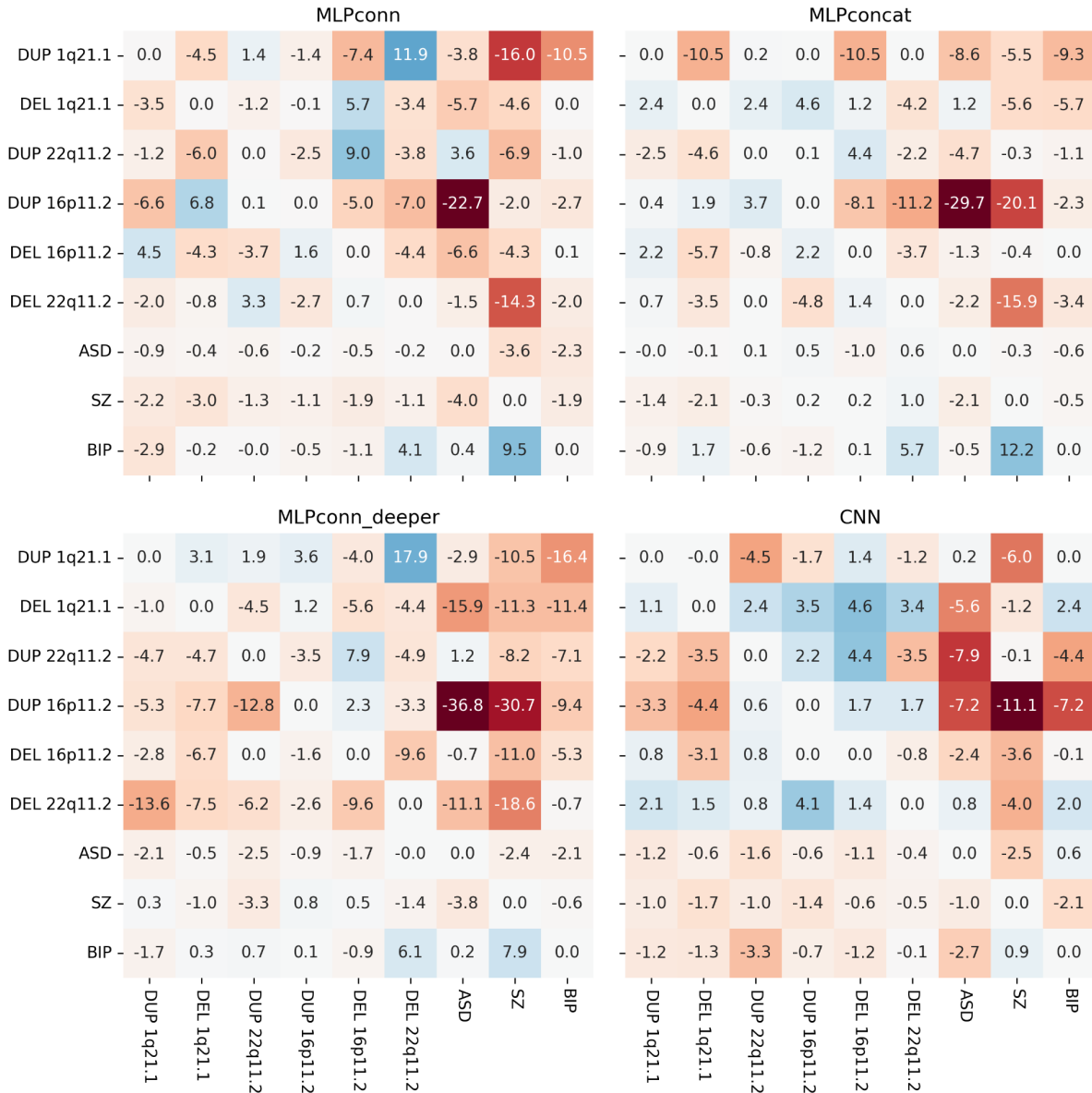


Figure 16 - Difference in accuracy from single-task baseline of conditions trained pairwise using multi-task learning in different data and model settings. The i,j th entry in the matrix is accuracy of condition in row i trained with condition in column j .

We aimed to disentangle the negative transfer observed when predicting the 9 neurodevelopmental and psychiatric conditions concurrently in the multi-task setting. We

trained a multi-task learning model on each pair of conditions in four different settings: simple feedforward network on connectomes alone (MLPconn), simple feedforward network on connectomes with confounds (MLPconcat), deeper feedforward network on connectomes alone (MLPconn_deeper), and a simple convolutional neural network (CNN) on connectomes alone. For each setting we also trained the models independently to establish a baseline. The matrix of pairwise accuracy for MLPconn showed a range of correlations with those for MLPconcat ($r = 0.71$), MLPconn_deeper ($r = 0.64$) and CNN ($r = 0.31$). We found that conditions with larger sample size (SZ, ASD) were not impacted by their partner, whereas for smaller sample size conditions the results were more variable. Apart from one case of consistent improvement (BIP being trained with SZ), the results were dependent on the model and data setting and did not reveal a clear pattern or point to a source for the observed negative transfer. Simple multi-task learning does not appear to be able to accommodate for the heterogeneity between the conditions, in spite of the shared latent task of decoding rs-fMRI data and overlapping effects of the conditions.

8.4 - Discussion

In the previous study, we trained a multi-task learning model to predict 9 CNVs and psychiatric conditions concurrently. We observed that the heterogeneity of the tasks was too large for hard parameter sharing multi-task learning framework to accommodate and the conditions interfered with rather than boosted each other's learning - a phenomenon called negative transfer.

Breaking down this complex objective, we implemented multi-task learning on the conditions pairwise using our primary model and three variations. This fine-grained exploration revealed a pattern of negative transfer between tasks that was stable across settings, ruling out any close changes to improve the model besides the multi-task learning framework itself. Specifically, in each pairing the prediction accuracy of one task was boosted at the expense of its partner - meaning that even with only two tasks the parameter sharing is too strict.

In a similar study of negative transfer, Standley and colleagues (Standley et al. 2019) used a standard encoder-decoder architecture and chose five tasks, that overlap enough to test if similar tasks train well together and that represented major task categories (one semantic, two 3D, and two 2D), from a large computer vision dataset (about 4 million examples). As in our study, the encoders were shared across tasks (hard parameter sharing), and they examined the matrix of prediction performance for tasks trained pairwise in different settings. The first setting explored a smaller capacity model, the second was a control condition, and the third a smaller dataset (about 200k examples).

When looking at a smaller dataset (with an amount of data more comparable to ours), Standley and colleagues (Standley et al. 2019) found most tasks suffer when trained with another task.

This is contrary to the notion that multi-task learning applies well to low-data scenarios, and could explain the poor performance observed in our study. They also found that tasks are more likely to benefit from being trained together when using a larger capacity network. The matrix of pairwise accuracy in the control setting was correlated with the low-data setting ($r = 0.558$), but unlike in our study it was not correlated with the modified capacity model ($r = 0.08$). As in our experiments, they found that certain tasks consistently improved the performance of their partner at their expense across. While they concluded that the relationships between tasks are dependent on the learning setup, this is difficult to confirm in our study where the correlations were high across settings. However, the change in dimension between networks in their study was much larger than ours and further experiments are needed. Finally, they found that similar tasks don't necessarily learn better together even with a much larger dataset, more complex model and better defined task relationships. This confirms what we observed for overlapping psychiatric conditions and is discouraging for future use of multi-task learning with hard parameter sharing for finding biologically relevant patterns across diagnostic categories.

In future studies the next step is to investigate soft parameter sharing schemes, in which information can be shared across models while leaving more flexibility for each model to learn its own representation. This approach proved successful in the studies by Huang and colleagues across psychiatric conditions (ASD, ADHD and SZ) (Huang, Liu, and Tan 2020; Huang et al. 2022), who report an increase in accuracy using their proposed method relative to both multi-task learning using hard parameter sharing and single task learning.

Chapter 9 - General Discussion

In this work we analysed a complex rs-fMRI dataset representing an unprecedented set of CNVs and psychiatric disorders through progressively more complex methods, first applying traditional statistics, next common machine learning algorithms, and finally multi-task learning using hard parameter sharing MLPs.

We first estimated the effect size of each condition on FC, CNVs with an association to a psychiatric condition, as well as BIP and SZ, were found to have large effect sizes. The other psychiatric conditions and CNVs had small-to-moderate effect sizes. Our results reflect previously reported effect sizes on FC and other imaging modalities for the same conditions (Assem et al. 2020; Moreau, Ching, et al. 2021; Van Rooij, Anagnostou, and Arango 2018; de Zwarte et al. 2019; van Erp et al. 2018; Hoogman et al. 2017, 2019). We then benchmarked the accuracy of common prediction algorithms on the same conditions, using intra- and inter-site cross-validation and evaluating accuracy relative to a confounds-only baseline model. Using intra-site cross-validation, we found that prediction accuracy for each condition broadly followed the trend of effect sizes, but that psychiatric conditions could be predicted better than effect size would indicate relative to the CNVs, implying that automatic diagnosis of CNVs could improve dramatically with sample size. Notably, we found that CNVs with no association to psychiatric conditions could not be predicted above chance level, which served as a validation of our approach - along with obtaining similar results to the ML literature for the psychiatric conditions (Arbabshirani et al. 2017; Nielsen et al. 2013; Zeng et al. 2018; H. Wang et al. 2022; Y. Wang et al. 2020). Inter-site cross-validation highlighted that increasing the number of sites of data collection, not only the sample size, is crucial for generalising to data collected at a new site, replicating the observation of Orban and colleagues (Orban et al. 2018).

Next, we described confound-isolating cross-validation, and compared the accuracy of prediction using our implementation of this strategy to the results of the benchmark and found that performance was diminished for certain conditions but mostly led to similar results. There is some debate about how to control for confounds in the predictive modelling context (Snoek, Miletić, and Scholte 2019; Chyzhyk et al. 2022; Dinga et al. 2020). While there are relatively well established procedures for the case of a single confound, the case of multiple confounds that we have attempted to address here is much more complex and the balanced test sets and corresponding training sets we found did not provide a perfectly unconfounded estimate of prediction accuracy. However, we demonstrated that they are acceptable in practice, and the accuracy of prediction from confounds in the intra- and inter-site benchmarks highlights the need to take confounds into consideration. Where it fits the aim of the study, comparing a model that includes confounds to a baseline model using only confounds is a simple, robust and easily interpretable approach.

We introduced our multi-task learning framework and evaluated its performance on a simple, well controlled benchmark predicting the same target (either age or sex) across sites and found that prediction was improved relative to single task learning for a majority of sites. Notably, prediction always improved for sites with large sample sizes, whereas small sites showed varied performance. This was contrary to our hypothesis that large sites would reach a plateau in accuracy in the single task setting and provide a stable representation that would improve the prediction for smaller sites. However, the sites we consider to have large sample size ($n = 1000$) are nowhere near the domain where models would begin to saturate in accuracy (Schulz et al. 2020) and rather than boost the tiny datasets ($n=30-94$) they appear to dominate the training. In the context of large imbalance in sample size between related tasks, transfer learning, in which a model is first trained on one task and then adapted to another, might offer a more promising framework. Transfer learning has been widely applied to MRI (Valverde et al. 2021), e.g. training a model on a very large dataset of natural images then tuning certain parameters for tumour segmentation in a small medical imaging dataset, and has shown promising results transferring from a broad range of fMRI data to brain decoding (Thomas, Ré, and Poldrack 2022).

Next, we applied multi-task learning across CNVs and psychiatric conditions to perform automatic diagnosis. Although we hypothesised that high rates of comorbidity (see section 1.1) and overlap in genetic factors and symptomatology (Lichtenstein et al. 2009; Lee et al. 2013; Brainstorm Consortium et al. 2018; Taylor et al. 2021) between the conditions in our dataset would allow them to benefit from multi-task learning, we observed these prediction tasks being learned concurrently was deleterious for accuracy overall. We then examined this phenomenon by implementing multi-task learning on the conditions pairwise using our primary model and three versions that varied the model capacity, input data, and type (from MLP to CNN). This fine-grained exploration did not reveal a clear pattern between conditions or source (model or data setting) to which we could attribute the negative transfer. Having ruled out immediate directions for modifying the existing framework, we turn to the issues of sample size and architecture choice for multi-task learning.

The sample sizes for the 9 CNVs and psychiatric conditions included in the multi-task learning study range from $n=19$ to 472 (cases only). These are much smaller than desired for either traditional statistical (Marek et al. 2022) or machine learning approaches, as seen in the limited prediction accuracy of CNVs in our benchmark study. Furthermore, while multi-task learning using hard parameter sharing was proposed to limit parameters and make better use of available data (Ruder 2017), evidence from our first multi-task study across sites with varying sample size as well as from Standley and colleagues (Standley et al. 2019) points to this framework being ineffective for small sample sizes.

Regarding the architecture, we implemented multi-task learning using hard parameter sharing and simple MLPs in order to first test a basic model with relatively limited parameters on our

dataset. However, our results show that this framework actually hurts prediction accuracy overall when applied across the conditions. Huang and colleagues (Huang et al. 2022) implemented a multicluster multigate mixture-of-experts (M-MMOE) model, which implements a form of soft parameter sharing, across ASD, ADHD and SZ to perform automatic diagnosis from connectomes. They found that their proposed approach improved accuracy for each condition relative to single task learning, whereas a hard parameter sharing MLP model did not. Ma and colleagues (J. Ma et al. 2018) found that the hard parameter sharing MLP model is sensitive to task relationships, and Standley and colleagues (Standley et al. 2019) showed that apparently similar tasks are not necessarily learned well together. These results clearly point to exploring different architectures in future studies, particularly those that explicitly learn relationships between tasks, Zhang and Yang (Y. Zhang and Yang 2017) provide a review of existing methods. In particular, a novel transformer-based multi-task learning architecture, which used a shared attention mechanism to model task relationships, outperformed state of the art multi-task CNN models on a range of computer vision tasks (Bhattacharjee et al. 2022).

Furthermore, there is debate about the use of connectomes for predictive modelling (Lurie et al. 2020) as information about brain dynamics is discarded when looking at correlation between time series of regions rather than the time series themselves, and various measures for computing connectomes can have a large impact on prediction accuracy (Abraham et al. 2017; Dadi et al. 2019). fMRI dynamics have been shown to provide features that outperform static connectomes on prediction of SZ and BIP (Rashid et al. 2016). Architectures that can directly model brain dynamics, such as transformers (Thomas, Ré, and Poldrack 2022) and graph convolutional networks (Y. Zhang, Farrugia, and Bellec 2022), provide another direction for future investigation.

In this study, we estimated effect sizes for 4 psychiatric conditions and 15 CNVs, 13 of which had never been previously investigated, then benchmarked prediction accuracy on these conditions using common machine learning methods, which is the first such study for all of the CNVs. We applied multi-task learning to predict a common target, across an unprecedented number of sites of data collection, and found that prediction was improved for a majority of sites. We applied multi-task learning across 9 psychiatric CNVs and psychiatric conditions to perform joint diagnosis, and found that prediction suffered overall and then explored the relationships between tasks. The hypothesis that high rates of comorbidity and genetic and behavioural overlap between CNVs and psychiatric conditions, and among psychiatric conditions, would translate into being closely related enough as prediction tasks to benefit from joint diagnosis was not supported. However, applying multi-task learning across sites showed that prediction can be improved when tasks are very tightly related, but the hypothesis that multi-task learning would benefit smaller sample sites by giving them access to the sample of larger sites by proxy was not supported. We implemented a lightweight hard parameter sharing model, but evidence from our results and the literature shows this framework is not well suited to heterogeneous tasks or, counterintuitively, to small sample sizes. While we believe there is potential to exploit

the similarities between CNVs and psychiatric conditions using methods that model relationships between tasks, small sample sizes for rare CNVs are a major limitation for the application of multi-task learning.

References

- Abraham, Alexandre, Michael P. Milham, Adriana Di Martino, R. Cameron Craddock, Dimitris Samaras, Bertrand Thirion, and Gael Varoquaux. 2017. "Deriving Reproducible Biomarkers from Multi-Site Resting-State Data: An Autism-Based Example." *NeuroImage* 147 (February): 736–45.
- ADHD-200 Consortium. 2012. "The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience." *Frontiers in Systems Neuroscience* 6 (September): 62.
- Arbabshirani, Mohammad R., Sergey Plis, Jing Sui, and Vince D. Calhoun. 2017. "Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls." *NeuroImage* 145 (Pt B): 137–65.
- Assem, Moataz, Matthew F. Glasser, David C. Van Essen, and John Duncan. 2020. "A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex." *Cerebral Cortex* 30 (8): 4361–80.
- Badhwar, Amanpreet, Yannik Collin-Verreault, Pierre Orban, Sebastian Urchs, Isabelle Chouinard, Jacob Vogel, Olivier Potvin, Simon Duchesne, and Pierre Bellec. 2020. "Multivariate Consistency of Resting-State fMRI Connectivity Maps Acquired on a Single Individual over 2.5 Years, 13 Sites and 3 Vendors." *NeuroImage* 205 (January): 116210.
- Barkley, Russell A. 2002. "Major Life Activity and Health Outcomes Associated with Attention-Deficit/hyperactivity Disorder." *The Journal of Clinical Psychiatry* 63 Suppl 12: 10–15.
- Bassett, Danielle S., Brent G. Nelson, Bryon A. Mueller, Jazmin Camchong, and Kelvin O. Lim. 2012. "Altered Resting State Complexity in Schizophrenia." *NeuroImage* 59 (3): 2196–2207.
- Bayer, Johanna M. M., Richard Dinga, Seyed Mostafa Kia, Akhil R. Kottaram, Thomas Wolfers, Jinglei Lv, Andrew Zalesky, Lianne Schmaal, and Andre Marquand. 2022. "Accommodating Site Variation in Neuroimaging Data Using Normative and Hierarchical Bayesian Models." *NeuroImage* 264 (December): 119699.
- Bellec, Pierre, Yassine Benhajali, Felix Carbonell, Christian Dansereau, Geneviève Albouy, Maxime Pelland, Cameron Craddock, et al. 2015. "Impact of the Resolution of Brain Parcels on Connectome-Wide Association Studies in fMRI." *NeuroImage* 123 (December): 212–28.
- Bellec, Pierre, Sébastien Lavoie-Courchesne, Phil Dickinson, Jason P. Lerch, Alex P. Zijdenbos, and Alan C. Evans. 2012. "The Pipeline System for Octave and Matlab (PSOM): A Lightweight Scripting Framework and Execution Engine for Scientific Workflows." *Frontiers in Neuroinformatics* 6 (April): 7.
- Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57 (1): 289–300.
- Bernanke, Joel, Alex Luna, Le Chang, Elizabeth Bruno, Jordan Dworkin, and Jonathan Posner. 2022. "Structural Brain Measures among Children with and without ADHD in the

- Adolescent Brain and Cognitive Development Study Cohort: A Cross-Sectional US Population-Based Study.” *The Lancet. Psychiatry* 9 (3): 222–31.
- Bhattacharjee, Deblina, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. 2022. “MuT: An End-to-End Multitask Learning Transformer.” *arXiv [cs.CV]*. arXiv. <http://arxiv.org/abs/2205.08303>.
- Biederman, Joseph, Carter R. Petty, Ronna Fried, Roselinde Kaiser, Chrystina R. Dolan, Steven Schoenfeld, Alysa E. Doyle, Larry J. Seidman, and Stephen V. Faraone. 2008. “Educational and Occupational Underattainment in Adults with Attention-Deficit/hyperactivity Disorder: A Controlled Study.” *The Journal of Clinical Psychiatry* 69 (8): 1217–22.
- Brainstorm Consortium, Verner Anttila, Brendan Bulik-Sullivan, Hilary K. Finucane, Raymond K. Walters, Jose Bras, Laramie Duncan, et al. 2018. “Analysis of Shared Heritability in Common Disorders of the Brain.” *Science* 360 (6395). <https://doi.org/10.1126/science.aap8757>.
- Bzdok, Danilo, and John P. A. Ioannidis. 2019. “Exploration, Inference, and Prediction in Neuroscience and Biomedicine.” *Trends in Neurosciences* 42 (4): 251–62.
- Chong, Huey Yi, Siew Li Teoh, David Bin-Chia Wu, Surachai Kotirum, Chiun-Fang Chiou, and Nathorn Chaiyakunapruk. 2016. “Global Economic Burden of Schizophrenia: A Systematic Review.” *Neuropsychiatric Disease and Treatment* 12 (February): 357–73.
- Chyzyk, Darya, Gaël Varoquaux, Michael Milham, and Bertrand Thirion. 2022. “How to Remove or Control Confounds in Predictive Models, with Applications to Brain Biomarkers.” *GigaScience* 11 (March). <https://doi.org/10.1093/gigascience/giac014>.
- Colella, Stefano, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes, and Jiannis Ragoussis. 2007. “QuantiSNP: An Objective Bayes Hidden-Markov Model to Detect and Accurately Map Copy Number Variation Using SNP Genotyping Data.” *Nucleic Acids Research* 35 (6): 2013–25.
- Crawford, Karen, Matthew Bracher-Smith, David Owen, Kimberley M. Kendall, Elliott Rees, Antonio F. Pardiñas, Mark Eion, et al. 2019. “Medical Consequences of Pathogenic CNVs in Adults: Analysis of the UK Biobank.” *Journal of Medical Genetics* 56 (3): 131–38.
- Dadi, Kamalaker, Mehdi Rahim, Alexandre Abraham, Darya Chyzyk, Michael Milham, Bertrand Thirion, Gaël Varoquaux, and Alzheimer’s Disease Neuroimaging Initiative. 2019. “Benchmarking Functional Connectome-Based Predictive Models for Resting-State fMRI.” *NeuroImage* 192 (May): 115–34.
- Dadi, Kamalaker, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J. Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. 2020. “Fine-Grain Atlases of Functional Modes for fMRI Analysis.” *NeuroImage* 221 (November): 117126.
- Di Martino, Adriana, David O’Connor, Bosi Chen, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Joshua H. Balsters, et al. 2017. “Enhancing Studies of the Connectome in Autism Using the Autism Brain Imaging Data Exchange II.” *Scientific Data* 4 (March): 170010.
- Di Martino, A., C-G Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, et al. 2014. “The Autism Brain Imaging Data Exchange: Towards a Large-Scale Evaluation of the Intrinsic Brain Architecture in Autism.” *Molecular Psychiatry* 19 (6): 659–67.

- Dinga, Richard, Lianne Schmaal, Brenda W. J. Penninx, Dick J. Veltman, and Andre F. Marquand. 2020. "Controlling for Effects of Confounding Variables on Machine Learning Predictions." *bioRxiv*. <https://doi.org/10.1101/2020.08.17.255034>.
- Dong, Qunxi, Jie Zhang, Qingyang Li, Junwen Wang, Natasha Laporé, Paul M. Thompson, Richard J. Caselli, Jieping Ye, Yalin Wang, and Alzheimer's Disease Neuroimaging Initiative. 2020. "Integrating Convolutional Neural Networks and Multi-Task Dictionary Learning for Cognitive Decline Prediction with Longitudinal Images." *Journal of Alzheimer's Disease: JAD* 75 (3): 971–92.
- Eloyan, Ani, John Muschelli, Mary Beth Nebel, Han Liu, Fang Han, Tuo Zhao, Anita D. Barber, et al. 2012. "Automated Diagnoses of Attention Deficit Hyperactive Disorder Using Magnetic Resonance Imaging." *Frontiers in Systems Neuroscience* 6 (August): 61.
- Erp, Theo G. M. van, Esther Walton, Derrek P. Hibar, Lianne Schmaal, Wenhao Jiang, David C. Glahn, Godfrey D. Pearlson, et al. 2018. "Cortical Brain Abnormalities in 4474 Individuals With Schizophrenia and 5098 Control Subjects via the Enhancing Neuro Imaging Genetics Through Meta Analysis (ENIGMA) Consortium." *Biological Psychiatry* 84 (9): 644–54.
- Esteban, Oscar, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, et al. 2019. "fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI." *Nature Methods* 16 (1): 111–16.
- Finn, Emily S., Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. 2015. "Functional Connectome Fingerprinting: Identifying Individuals Using Patterns of Brain Connectivity." *Nature Neuroscience* 18 (11): 1664–71.
- Glover, Gary H. 2011. "Overview of Functional Magnetic Resonance Imaging." *Neurosurgery Clinics of North America* 22 (2): 133–39, vii.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.
- Gordon, Evan M., Timothy O. Laumann, Adrian W. Gilmore, Dillan J. Newbold, Deanna J. Greene, Jeffrey J. Berg, Mario Ortega, et al. 2017. "Precision Functional Mapping of Individual Human Brains." *Neuron* 95 (4): 791–807.e7.
- Grattarola, Daniele. 2017. "Deep Feature Extraction for Sample-Efficient Reinforcement Learning." Unpublished. <https://doi.org/10.13140/RG.2.2.30267.31527>.
- Hahn, Sage, Max M. Owens, Dekang Yuan, Anthony C. Juliano, Alexandra Potter, Hugh Garavan, and Nicholas Allgaier. 2022. "Performance Scaling for Structural MRI Surface Parcellations." Performance Scaling for Structural MRI Surface Parcellations. April 3, 2022. https://sahahn.github.io/parc_scaling/.
- Harvey, Annabelle, and Guillaume Dumas. 2022. "PyNM: A Lightweight Python Implementation of Normative Modeling." *Journal of Open Source Software* 7 (80): 4321.
- Hay, Simon I., Amanuel Alemu Abajobir, Kalkidan Hassen Abate, Cristiana Abbafati, Kaja M. Abbas, Foad Abd-Allah, Rizwan Suliankatchi Abdulkader, et al. 2017. "Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 333 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016." *The Lancet* 390 (10100): 1260–1344.

- Heinsfeld, Anibal Sólón, Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. 2018. "Identification of Autism Spectrum Disorder Using Deep Learning and the ABIDE Dataset." *NeuroImage. Clinical* 17: 16–23.
- He, Lili, Hailong Li, Jinghua Wang, Ming Chen, Elveda Gozdas, Jonathan R. Dillman, and Nehal A. Parikh. 2020. "A Multi-Task, Multi-Stage Deep Transfer Learning Model for Early Prediction of Neurodevelopment in Very Preterm Infants." *Scientific Reports* 10 (1): 15072.
- Hoogman, Martine, Janita Bralten, Derrek P. Hibar, Maarten Mennes, Marcel P. Zwiers, Lianne S. J. Schweren, Kimm J. E. van Hulzen, et al. 2017. "Subcortical Brain Volume Differences in Participants with Attention Deficit Hyperactivity Disorder in Children and Adults: A Cross-Sectional Mega-Analysis." *The Lancet. Psychiatry* 4 (4): 310–19.
- Hoogman, Martine, Ryan Muetzel, Joao P. Guimaraes, Elena Shumskaya, Maarten Mennes, Marcel P. Zwiers, Neda Jahanshad, et al. 2019. "Brain Imaging of the Cortex in ADHD: A Coordinated Analysis of Large-Scale Clinical and Population-Based Samples." *The American Journal of Psychiatry* 176 (7): 531–42.
- Huang, Zhi-An, Rui Liu, and Kay Chen Tan. 2020. "Multi-Task Learning for Efficient Diagnosis of ASD and ADHD Using Resting-State fMRI Data." In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7.
- Huang, Zhi-An, Rui Liu, Zexuan Zhu, and Kay Chen Tan. 2022. "Multitask Learning for Joint Diagnosis of Multiple Mental Disorders in Resting-State fMRI." *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.
- Hu, Dewen, and Ling-Li Zeng. 2019. "Multi-Task Learning of Structural MRI for Multi-Site Classification." In *Pattern Analysis of the Human Connectome*, edited by Dewen Hu and Ling-Li Zeng, 205–26. Singapore: Springer Singapore.
- Huguet, Guillaume, Catherine Schramm, Elise Douard, Lai Jiang, Aurélie Labbe, Frédérique Tihy, Géraldine Mathonnet, et al. 2018. "Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples." *JAMA Psychiatry* 75 (5): 447–57.
- Inacio, Maria C. S., Yuexin Chen, Elizabeth W. Paxton, Robert S. Namba, Steven M. Kurtz, and Guy Cafri. 2015. "Statistics in Brief: An Introduction to the Use of Propensity Scores." *Clinical Orthopaedics and Related Research* 473 (8): 2722–26.
- Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/1502.03167>.
- Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods." *Biostatistics* 8 (1): 118–27.
- Jonas, Rachel K., Caroline A. Montojo, and Carrie E. Bearden. 2014. "The 22q11.2 Deletion Syndrome as a Window into Complex Neuropsychiatric Disorders over the Lifespan." *Biological Psychiatry* 75 (5): 351–60.
- Jønch, Aia Elise, Elise Douard, Clara Moreau, Anke Van Dijck, Marzia Passeggeri, Frank Kooy, Jacques Puechberty, et al. 2019. "Estimating the Effect Size of the 15Q11.2 BP1-BP2 Deletion and Its Contribution to Neurodevelopmental Symptoms: Recommendations for Practice."

- Journal of Medical Genetics* 56 (10): 701–10.
- Katzman, Martin A., Timothy S. Bilkey, Pratap R. Chokka, Angelo Fallu, and Larry J. Klassen. 2017. “Adult ADHD and Comorbid Disorders: Clinical Implications of a Dimensional Approach.” *BMC Psychiatry* 17 (1): 302.
- Kawahara, Jeremy, Colin J. Brown, Steven P. Miller, Brian G. Booth, Vann Chau, Ruth E. Grunau, Jill G. Zwicker, and Ghassan Hamarneh. 2017. “BrainNetCNN: Convolutional Neural Networks for Brain Networks; towards Predicting Neurodevelopment.” *NeuroImage* 146 (February): 1038–49.
- Kearney, Hutton M., Erik C. Thorland, Kerry K. Brown, Fabiola Quintero-Rivera, Sarah T. South, and Others. 2011. “Working Group of the American College of Medical Genetics Laboratory Quality Assurance C. American College of Medical Genetics Standards and Guidelines for Interpretation and Reporting of Postnatal Constitutional Copy Number Variants.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 13 (7): 680–85.
- Khosla, Meenakshi, Keith Jamison, Gia H. Ngo, Amy Kuceyeski, and Mert R. Sabuncu. 2019. “Machine Learning in Resting-State fMRI Analysis.” *Magnetic Resonance Imaging* 64 (December): 101–21.
- Khosla, Meenakshi, K. Jamison, Amy Kuceyeski, and M. Sabuncu. 2018. “Ensemble Learning with 3D Convolutional Neural Networks for Connectome-Based Prediction.” *arXiv.org*. <https://www.semanticscholar.org/paper/df2a223be0a4ef887ca5874e9dd0c2fb501238a5>.
- Kim, Junghoe, Vince D. Calhoun, Eunsoo Shim, and Jong-Hwan Lee. 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 (Pt A): 127–46.
- Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/1412.6980>.
- Kong, Ru, Qing Yang, Evan Gordon, Aihuiping Xue, Xiaoxuan Yan, Csaba Orban, Xi-Nian Zuo, et al. 2021. “Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior.” *Cerebral Cortex* 31 (10): 4477–4500.
- Lee, S. Hong, Stephan Ripke, Benjamin M. Neale, Stephen V. Faraone, Shaun M. Purcell, Roy H. Perlis, Bryan J. Mowry, et al. 2013. “Genetic Relationship between Five Psychiatric Disorders Estimated from Genome-Wide SNPs.” *Nature Genetics* 45 (9): 984–94.
- Leming, Matthew, Juan Manuel Górriz, and John Suckling. 2020. “Ensemble Deep Learning on Large, Mixed-Site fMRI Datasets in Autism and Other Tasks.” *International Journal of Neural Systems* 30 (7): 2050012.
- Leming, Matthew, and John Suckling. 2021. “Deep Learning for Sex Classification in Resting-State and Task Functional Brain Networks from the UK Biobank.” *NeuroImage* 241 (November): 118409.
- Liang, Wei, Kai Zhang, Peng Cao, Xiaoli Liu, Jinzhu Yang, and Osmar Zaiane. 2021. “Rethinking Modeling Alzheimer’s Disease Progression from a Multi-Task Learning Perspective with Deep Recurrent Neural Network.” *Computers in Biology and Medicine* 138 (November): 104935.

- Lichtenstein, Paul, Benjamin H. Yip, Camilla Björk, Yudi Pawitan, Tyrone D. Cannon, Patrick F. Sullivan, and Christina M. Hultman. 2009. “Common Genetic Determinants of Schizophrenia and Bipolar Disorder in Swedish Families: A Population-Based Study.” *The Lancet* 373 (9659): 234–39.
- Li, Jian, Anand A. Joshi, and Richard M. Leahy. 2020. “A NETWORK-BASED APPROACH TO STUDY OF ADHD USING TENSOR DECOMPOSITION OF RESTING STATE FMRI DATA.” *Proceedings / IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE International Symposium on Biomedical Imaging 2020* (April): 544–48.
- Li, Jingwei, Ru Kong, Raphaël Liégeois, Csaba Orban, Yanrui Tan, Nanbo Sun, Avram J. Holmes, Mert R. Sabuncu, Tian Ge, and B. T. Thomas Yeo. 2019. “Global Signal Regression Strengthens Association between Resting-State Functional Connectivity and Behavior.” *NeuroImage* 196 (August): 126–41.
- Linn, Kristin A., Bilwaj Gaonkar, Jimit Doshi, Christos Davatzikos, and Russell T. Shinohara. 2016. “Addressing Confounding in Predictive Models with an Application to Neuroimaging.” *The International Journal of Biostatistics* 12 (1): 31–44.
- Lo, Adeline, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. 2015. “Why Significant Variables Aren’t Automatically Good Predictors.” *Proceedings of the National Academy of Sciences* 112 (45): 13892–97.
- Lurie, Daniel J., Daniel Kessler, Danielle S. Bassett, Richard F. Betzel, Michael Breakspear, Shella Kheilholz, Aaron Kucyi, et al. 2020. “Questions and Controversies in the Study of Time-Varying Functional Connectivity in Resting fMRI.” *Network Neuroscience* (Cambridge, Mass.) 4 (1): 30–69.
- Ma, Jiaqi, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. “Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts.” In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1930–39. KDD ’18. New York, NY, USA: Association for Computing Machinery.
- Ma, Qiongmin, Tianhao Zhang, Marcus V. Zanetti, Hui Shen, Theodore D. Satterthwaite, Daniel H. Wolf, Raquel E. Gur, et al. 2018. “Classification of Multi-Site MR Images in the Presence of Heterogeneity Using Multi-Task Learning.” *NeuroImage. Clinical* 19 (May): 476–86.
- Marek, Scott, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, et al. 2022. “Reproducible Brain-Wide Association Studies Require Thousands of Individuals.” *Nature* 603 (7902): 654–60.
- Marquand, Andre F., Michael Brammer, Steven C. R. Williams, and Orla M. Doyle. 2014. “Bayesian Multi-Task Learning for Decoding Multi-Subject Neuroimaging Data.” *NeuroImage* 92 (100): 298–311.
- Marshall, Christian R., Daniel P. Howrigan, Daniele Merico, Bhooma Thiruvahindrapuram, Wenting Wu, Douglas S. Greer, Danny Antaki, et al. 2017. “Contribution of Copy Number Variants to Schizophrenia from a Genome-Wide Study of 41,321 Subjects.” *Nature Genetics* 49 (1): 27–35.
- McElroy, Susan L. 2004. “Diagnosing and Treating Comorbid (complicated) Bipolar Disorder.” *The Journal of Clinical Psychiatry* 65 Suppl 15: 35–44.

- Mellema, Cooper J., Kevin P. Nguyen, Alex Treacher, and Albert Montillo. 2022. "Reproducible Neuroimaging Features for Diagnosis of Autism Spectrum Disorder with Machine Learning." *Scientific Reports* 12 (1): 3057.
- Merikangas, Kathleen R., Robert Jin, Jian-Ping He, Ronald C. Kessler, Sing Lee, Nancy A. Sampson, Maria Carmen Viana, et al. 2011. "Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health Survey Initiative." *Archives of General Psychiatry* 68 (3): 241–51.
- Meszlényi, Regina J., Krisztian Buza, and Zoltán Vidnyánszky. 2017. "Resting State fMRI Functional Connectivity-Based Classification Using a Convolutional Neural Network Architecture." *Frontiers in Neuroinformatics* 11 (October): 61.
- Miroglio, Ben, David Zeber, Jofish Kaye, and Rebecca Weiss. 2018. "The Effect of Ad Blocking on User Engagement with the Web." In *Proceedings of the 2018 World Wide Web Conference*, 813–21. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.
- Modenato, Claudia, Kuldeep Kumar, Clara Moreau, Sandra Martin-Brevet, Guillaume Huguet, Catherine Schramm, Martineau Jean-Louis, et al. 2021. "Effects of Eight Neuropsychiatric Copy Number Variants on Human Brain Structure." *Translational Psychiatry* 11 (1): 399.
- Moreau, Clara A., Christopher Rk Ching, Kuldeep Kumar, Sebastien Jacquemont, and Carrie E. Bearden. 2021. "Structural and Functional Brain Alterations Revealed by Neuroimaging in CNV Carriers." *Current Opinion in Genetics & Development* 68 (June): 88–98.
- Moreau, Clara A., Annabelle Harvey, Kuldeep Kumar, Guillaume Huguet, Sebastian G. W. Urchs, Elise A. Douard, Laura M. Schultz, et al. 2023. "Genetic Heterogeneity Shapes Brain Connectivity in Psychiatry." *Biological Psychiatry* 93 (1): 45–58.
- Moreau, Clara A., Kuldeep Kumar, Annabelle Harvey, Guillaume Huguet, Sebastian Urchs, Laura M. Schultz, Hanad Sharmarke, et al. 2022. "Brain Functional Connectivity Mirrors Genetic Pleiotropy in Psychiatric Conditions." *Brain: A Journal of Neurology*, September. <https://doi.org/10.1093/brain/awac315>.
- Moreau, Clara A., Armin Raznahan, Pierre Bellec, Mallar Chakravarty, Paul M. Thompson, and Sebastien Jacquemont. 2021. "Dissecting Autism and Schizophrenia through Neuroimaging Genomics." *Brain: A Journal of Neurology* 144 (7): 1943–57.
- Moreau, Clara A., Sebastian G. W. Urchs, Kumar Kuldeep, Pierre Orban, Catherine Schramm, Guillaume Dumas, Aurélie Labbe, et al. 2020. "Mutations Associated with Neuropsychiatric Conditions Delineate Functional Brain Connectivity Dimensions Contributing to Autism and Schizophrenia." *Nature Communications* 11 (1): 5272.
- Moreno-De-Luca, D., S. J. Sanders, A. J. Willsey, J. G. Mulle, J. K. Lowe, D. H. Geschwind, M. W. State, C. L. Martin, and D. H. Ledbetter. 2013. "Using Large Clinical Data Sets to Infer Pathogenicity for Rare Copy Number Variants in Autism Cohorts." *Molecular Psychiatry* 18 (10): 1090–95.
- Moreno-Küstner, Berta, Carlos Martín, and Loly Pastor. 2018. "Prevalence of Psychotic Disorders and Its Association with Methodological Issues. A Systematic Review and Meta-Analyses." *PloS One* 13 (4): e0195687.

- Ngo, Duc-Ky, Minh-Trieu Tran, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. 2020. "Multi-Task Learning for Small Brain Tumor Segmentation from MRI." *NATO Advanced Science Institutes Series E: Applied Sciences* 10 (21): 7790.
- Niarchou, Maria, Samuel J. R. A. Chawner, Joanne L. Doherty, Anne M. Maillard, Sébastien Jacquemont, Wendy K. Chung, Leeanne Green-Snyder, et al. 2019. "Psychiatric Disorders in Children with 16p11.2 Deletion and Duplication." *Translational Psychiatry* 9 (1): 8.
- Nielsen, Jared A., Brandon A. Zielinski, P. Thomas Fletcher, Andrew L. Alexander, Nicholas Lange, Erin D. Bigler, Janet E. Lainhart, and Jeffrey S. Anderson. 2013. "Multisite Functional Connectivity MRI Classification of Autism: ABIDE Results." *Frontiers in Human Neuroscience* 7 (September): 599.
- Nilearn. n.d. "9.3. From Neuroimaging Volumes to Data Matrices: The Masker Objects." Nilearn Documentation. Accessed April 23, 2023. https://nilearn.github.io/manipulating_images/masker_objects.html.
- Noble, Stephanie, Dustin Scheinost, and R. Todd Constable. 2019. "A Decade of Test-Retest Reliability of Functional Connectivity: A Systematic Review and Meta-Analysis." *NeuroImage* 203 (December): 116157.
- Noble, Stephanie, Marisa N. Spann, Fuyuze Tokoglu, Xilin Shen, R. Todd Constable, and Dustin Scheinost. 2017. "Influences on the Test-Retest Reliability of Functional Connectivity MRI and Its Relationship with Behavioral Utility." *Cerebral Cortex* 27 (11): 5415–29.
- Orban, Pierre, Christian Dansereau, Laurence Desbois, Violaine Mongeau-Pérusse, Charles-Édouard Giguère, Hien Nguyen, Adrianna Mendrek, Emmanuel Stip, and Pierre Bellec. 2018. "Multisite Generalizability of Schizophrenia Diagnosis Classification Based on Functional Brain Connectivity." *Schizophrenia Research* 192 (February): 167–71.
- Orban, Pierre, Martin Desseilles, Adrianna Mendrek, Josiane Bourque, Pierre Bellec, and Emmanuel Stip. 2017. "Altered Brain Connectivity in Patients with Schizophrenia Is Consistent across Cognitive Contexts." *Journal of Psychiatry & Neuroscience: JPN* 42 (1): 17–26.
- Pan, Sinno Jialin, and Qiang Yang. 2010. "A Survey on Transfer Learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59.
- Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *arXiv [cs.LG]*. arXiv. <https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>.
- Patel, Krishna R., Jessica Cherian, Kunj Gohil, and Dylan Atkinson. 2014. "Schizophrenia: Overview and Treatment Options." *P & T: A Peer-Reviewed Journal for Formulary Management* 39 (9): 638–45.
- Phipson, Belinda, and Gordon K. Smyth. 2010. "Permutation P-Values Should Never Be Zero: Calculating Exact P-Values When Permutations Are Randomly Drawn." *Statistical Applications in Genetics and Molecular Biology* 9 (October): Article39.
- Poldrack, R. A., E. Congdon, W. Triplett, K. J. Gorgolewski, K. H. Karlsgodt, J. A. Mumford, F. W.

- Sabb, et al. 2016. "A Phenome-Wide Examination of Neural and Cognitive Function." *Scientific Data* 3 (December): 160110.
- Posner, Jonathan, Christine Park, and Zhishun Wang. 2014. "Connecting the Dots: A Review of Resting Connectivity MRI Studies in Attention-Deficit/hyperactivity Disorder." *Neuropsychology Review* 24 (1): 3–15.
- Rahim, Mehdi, Bertrand Thirion, Danilo Bzdok, Irène Buvat, and Gaël Varoquaux. 2017. "Joint Prediction of Multiple Scores Captures Better Individual Traits from Brain Images." *NeuroImage* 158 (September): 145–54.
- Rao, Anil, Joao M. Monteiro, Janaina Mourao-Miranda, and Alzheimer's Disease Initiative. 2017. "Predictive Modelling Using Neuroimaging Data in the Presence of Confounds." *NeuroImage* 150 (April): 23–49.
- Rao, Nikhil, Christopher Cox, Robert Nowak, and Timothy Rogers. 2013. "Sparse Overlapping Sets Lasso for Multitask Learning and Its Application to fMRI Analysis." *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/1311.5422>.
- Rashid, Barnaly, Mohammad R. Arbabshirani, Eswar Damaraju, Mustafa S. Cetin, Robyn Miller, Godfrey D. Pearlson, and Vince D. Calhoun. 2016. "Classification of Schizophrenia and Bipolar Patients Using Static and Dynamic Resting-State fMRI Brain Connectivity." *NeuroImage* 134 (July): 645–57.
- Ratner, Alexander, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. "Snorkel: Rapid Training Data Creation with Weak Supervision." *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* 11 (3): 269–82.
- Rees, Elliott, and George Kirov. 2021. "Copy Number Variation and Neuropsychiatric Illness." *Current Opinion in Genetics & Development* 68 (June): 57–63.
- Romero, Cato, Josefin Werme, Philip R. Jansen, Joel Gelernter, Murray B. Stein, Daniel Levey, Renato Polimanti, Christiaan de Leeuw, Mats Nagel, and Sophie van der Sluis. 2022. "Exploring the Genetic Overlap between Twelve Psychiatric Disorders." *Nature Genetics* 54 (12): 1795–1802.
- Rösler, Michael, Miguel Casas, Eric Konofal, and Jan Buitelaar. 2010. "Attention Deficit Hyperactivity Disorder in Adults." *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry* 11 (5): 684–98.
- Ruder, Sebastian. 2017. "An Overview of Multi-Task Learning in Deep Neural Networks." *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/1706.05098>.
- Saha, Sukanta, David Chant, Joy Welham, and John McGrath. 2005. "A Systematic Review of the Prevalence of Schizophrenia." *PLoS Medicine* 2 (5): e141.
- Sajatovic, Martha. 2005. "Bipolar Disorder: Disease Burden." *The American Journal of Managed Care* 11 (3 Suppl): S80–84.
- Sanders, Stephan J., Xin He, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Kaitlin E. Samocha, A. Ercument Cicek, Michael T. Murtha, et al. 2015. "Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci." *Neuron* 87 (6): 1215–33.
- Sanders, Stephan J., Mustafa Sahin, Joseph Hostyk, Audrey Thurm, Sebastien Jacquemont, Paul Avillach, Elise Douard, et al. 2019. "A Framework for the Investigation of Rare Genetic

- Disorders in Neuropsychiatry.” *Nature Medicine* 25 (10): 1477–87.
- Satterstrom, F. Kyle, Jack A. Kosmicki, Jiebiao Wang, Michael S. Breen, Silvia De Rubeis, Joon-Yong An, Minshi Peng, et al. 2020. “Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism.” *Cell* 180 (3): 568–84.e23.
- Schäfer, Thomas, and Marcus A. Schwarz. 2019. “The Meaningfulness of Effect Sizes in Psychological Research: Differences between Sub-Disciplines and the Impact of Potential Biases.” *Frontiers in Psychology* 10 (April): 813.
- Schulz, Marc-Andre, B. T. Thomas Yeo, Joshua T. Vogelstein, Janaina Mourao-Miranada, Jakob N. Kather, Konrad Kording, Blake Richards, and Danilo Bzdok. 2020. “Different Scaling of Linear Models and Deep Learning in UKBiobank Brain Images versus Machine-Learning Datasets.” *Nature Communications* 11 (1): 4238.
- Shmueli, Galit. 2010. “To Explain or to Predict?” *Schweizerische Monatsschrift Fur Zahnheilkunde = Revue Mensuelle Suisse D’odonto-Stomatologie / SSO* 25 (3): 289–310.
- Simonoff, Emily, Andrew Pickles, Tony Charman, Susie Chandler, Tom Loucas, and Gillian Baird. 2008. “Psychiatric Disorders in Children with Autism Spectrum Disorders: Prevalence, Comorbidity, and Associated Factors in a Population-Derived Sample.” *Journal of the American Academy of Child and Adolescent Psychiatry* 47 (8): 921–29.
- Simons Vip Consortium. 2012. “Simons Variation in Individuals Project (Simons VIP): A Genetics-First Approach to Studying Autism Spectrum and Related Neurodevelopmental Disorders.” *Neuron* 73 (6): 1063–67.
- Simon, Viktória, Pál Czobor, Sára Bálint, Agnes Mészáros, and István Bitter. 2009. “Prevalence and Correlates of Adult Attention-Deficit Hyperactivity Disorder: Meta-Analysis.” *The British Journal of Psychiatry: The Journal of Mental Science* 194 (3): 204–11.
- Snoek, Lukas, Steven Miletic, and H. Steven Scholte. 2019. “How to Control for Confounds in Decoding Analyses of Neuroimaging Data.” *NeuroImage* 184 (January): 741–60.
- Sønderby, Ida E., Christopher R. K. Ching, Sophia I. Thomopoulos, Dennis van der Meer, Daqiang Sun, Julio E. Villalon-Reina, Ingrid Agartz, et al. 2022. “Effects of Copy Number Variations on Brain Structure and Risk for Psychiatric Illness: Large-Scale Studies from the ENIGMA Working Groups on CNVs.” *Human Brain Mapping* 43 (1): 300–328.
- Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” 2014. https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer.
- Standley, Trevor, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2019. “Which Tasks Should Be Learned Together in Multi-Task Learning?” *arXiv [cs.CV]*. arXiv. <http://arxiv.org/abs/1905.07553>.
- Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.” *PLoS Medicine* 12 (3): e1001779.
- Syan, Sabrina K., Mara Smith, Benicio N. Frey, Raheem Remtulla, Flavio Kapczynski, Geoffrey B.

- C. Hall, and Luciano Minuzzi. 2018. "Resting-State Functional Connectivity in Individuals with Bipolar Disorder during Clinical Remission: A Systematic Review." *Journal of Psychiatry & Neuroscience: JPN* 43 (5): 298–316.
- Tabarestani, Solale, Mohammad Eslami, Mercedes Cabrerizo, Rosie E. Curiel, Armando Barreto, Naphtali Rishe, David Vaillancourt, et al. 2022. "A Tensorized Multitask Deep Learning Network for Progression Prediction of Alzheimer's Disease." *Frontiers in Aging Neuroscience* 14 (May): 810873.
- Taylor, Cora M., Rebecca Smith, Christopher Lehman, Marissa W. Mitchel, Kaitlyn Singer, W. Curtis Weaver, and Wendy Chung. 2021. *16p11.2 Recurrent Deletion*. University of Washington, Seattle.
- Thomas, Armin W., Christopher Ré, and Russell A. Poldrack. 2022. "Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data." *arXiv [q-bio.NC]*. arXiv. <http://arxiv.org/abs/2206.11417>.
- Traut, Nicolas, Katja Heuer, Guillaume Lemaître, Anita Beggiato, David Germanaud, Monique Elmaleh, Alban Bethegnies, et al. 2022. "Insights from an Autism Imaging Biomarker Challenge: Promises and Threats to Biomarker Discovery." *NeuroImage* 255 (July): 119171.
- Tsai, Jack, and Robert A. Rosenheck. 2013. "Psychiatric Comorbidity among Adults with Schizophrenia: A Latent Class Analysis." *Psychiatry Research* 210 (1): 16–20.
- Valverde, Juan Miguel, Vandad Imani, Ali Abdollahzadeh, Riccardo De Feo, Mithilesh Prakash, Robert Cizek, and Jussi Tohka. 2021. "Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review." *The Journal of Imaging Science and Technology / IS&T, the Society for Imaging Science and Technology* 7 (4). <https://doi.org/10.3390/jimaging7040066>.
- Van Rooij, D., E. Anagnostou, and C. Arango. 2018. "Cortical and Subcortical Brain Morphometry Differences between Patients with Autism Spectrum Disorder and Healthy Individuals across the Lifespan: Results from the" *American Journal of Psychiatry*. <https://ajp.psychiatryonline.org/doi/abs/10.1176/appi.ajp.2017.17010100>.
- Varoquaux, Gaël. 2018. "Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars." *NeuroImage* 180 (Pt A): 68–77.
- Venkataraman, Archana, Thomas J. Whitford, Carl-Fredrik Westin, Polina Golland, and Marek Kubicki. 2012. "Whole Brain Resting State Functional Connectivity Abnormalities in Schizophrenia." *Schizophrenia Research* 139 (1-3): 7–12.
- Wang, Huan, Rongxin Zhu, Shui Tian, Junneng Shao, Zhongpeng Dai, Li Xue, Yurong Sun, Zhilu Chen, Zhijian Yao, and Qing Lu. 2022. "Classification of Bipolar Disorders Using the Multilayer Modularity in Dynamic Minimum Spanning Tree from Resting State fMRI." *Cognitive Neurodynamics*, December. <https://doi.org/10.1007/s11571-022-09907-x>.
- Wang, Kai, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F. A. Grant, Hakon Hakonarson, and Maja Bucan. 2007. "PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data." *Genome Research* 17 (11): 1665–74.
- Wang, Xiangyang, Tianhao Zhang, Tiffany M. Chaim, Marcus V. Zanetti, and Christos Davatzikos. 2015. "Classification of MRI under the Presence of Disease Heterogeneity Using

- Multi-Task Learning: Application to Bipolar Disorder.” *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 9349 (October): 125–32.
- Wang, Ying, Kai Sun, Zhenyu Liu, Guanmao Chen, Yanbin Jia, Shuming Zhong, Jiyang Pan, Li Huang, and Jie Tian. 2020. “Classification of Unmedicated Bipolar Disorder Using Whole-Brain Functional Activity and Connectivity: A Radiomics Analysis.” *Cerebral Cortex* 30 (3): 1117–28.
- Wang, Zhaobin, Xiaocheng Zhou, Yuanyuan Gui, Manhua Liu, and Hui Lu. 2023. “Multiple Measurement Analysis of Resting-State fMRI for ADHD Classification in Adolescent Brain from the ABCD Study.” *Translational Psychiatry* 13 (1): 45.
- Watanabe, Takanori, Daniel Kessler, Clayton Scott, and Chandra Sripada. 2014. “Multisite Disease Classification with Functional Connectomes via Multitask Structured Sparse SVM.” 2014.
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=da3752254a91621ecbfa8bfe06d109bdffd45549>.
- Willsey, Helen Rankin, A. Jeremy Willsey, Belinda Wang, and Matthew W. State. 2022. “Genomics, Convergent Neuroscience and Progress in Understanding Autism Spectrum Disorder.” *Nature Reviews. Neuroscience* 23 (6): 323–41.
- Xiao, Li, Julia M. Stephen, Tony W. Wilson, Vince D. Calhoun, and Yu-Ping Wang. 2020. “A Manifold Regularized Multi-Task Learning Model for IQ Prediction From Two fMRI Paradigms.” *IEEE Transactions on Biomedical Engineering* 67 (3): 796–806.
- Yan, Chao-Gan, R. Cameron Craddock, Xi-Nian Zuo, Yu-Feng Zang, and Michael P. Milham. 2013. “Standardizing the Intrinsic Brain: Towards Robust Measurement of Inter-Individual Variation in 1000 Functional Connectomes.” *NeuroImage* 80 (October): 246–62.
- Yu, Chenglin, Dingnan Cui, Muheng Shang, Shu Zhang, Lei Guo, Junwei Han, Lei Du, and Alzheimer’s Disease Neuroimaging Initiative. 2021. “A Multi-Task Deep Feature Selection Method for Brain Imaging Genetics.” *arXiv [q-bio.GN]*. arXiv. <http://arxiv.org/abs/2107.00388>.
- Yu, Meichen, Kristin A. Linn, Philip A. Cook, Mary L. Phillips, Melvin McInnis, Maurizio Fava, Madhukar H. Trivedi, Myrna M. Weissman, Russell T. Shinohara, and Yvette I. Sheline. 2018. “Statistical Harmonization Corrects Site Effects in Functional Connectivity Measurements from Multi-Site fMRI Data.” *Human Brain Mapping* 39 (11): 4213–27.
- Zeidan, Jinan, Eric Fombonne, Julie Scora, Alaa Ibrahim, Maureen S. Durkin, Shekhar Saxena, Afqah Yusuf, Andy Shih, and Mayada Elsabbagh. 2022. “Global Prevalence of Autism: A Systematic Review Update.” *Autism Research: Official Journal of the International Society for Autism Research* 15 (5): 778–90.
- Zeng, Ling-Li, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, et al. 2018. “Multi-Site Diagnostic Classification of Schizophrenia Using Discriminant Deep Learning with Functional Connectivity MRI.” *EBioMedicine* 30 (April): 74–85.
- Zhang, Daoqiang, Dinggang Shen, and Alzheimer’s Disease Neuroimaging Initiative. 2012. “Multi-Modal Multi-Task Learning for Joint Prediction of Multiple Regression and Classification Variables in Alzheimer’s Disease.” *NeuroImage* 59 (2): 895–907.

- Zhang, Yu, Nicolas Farrugia, and Pierre Bellec. 2022. "Deep Learning Models of Cognitive Processes Constrained by Human Brain Connectomes." *Medical Image Analysis* 80 (August): 102507.
- Zhang, Yu, and Qiang Yang. 2017. "A Survey on Multi-Task Learning." *arXiv [cs.LG]*. arXiv. <http://arxiv.org/abs/1707.08114>.
- Zhou, Jiayu, Jun Liu, Vaibhav A. Narayan, Jieping Ye, and Alzheimer's Disease Neuroimaging Initiative. 2013. "Modeling Disease Progression via Multi-Task Learning." *NeuroImage* 78 (September): 233–48.
- Zwarte, Sonja M. C. de, Rachel M. Brouwer, Ingrid Agartz, Martin Alda, André Aleman, Kathryn I. Alpert, Carrie E. Bearden, et al. 2019. "The Association Between Familial Risk and Brain Abnormalities Is Disease Specific: An ENIGMA-Relatives Study of Schizophrenia and Bipolar Disorder." *Biological Psychiatry* 86 (7): 545–56.

Appendix A - Evaluating Confound-Isolating Cross-Validation in MTL Setting

A.1 - Intro

In Chapter 5 we introduced confound-isolating cross-validation and demonstrated that the balanced test sets we found for each condition resulted in chance level prediction of most the 9 CNVs and psychiatric conditions from confounds, and then used these test sets to evaluate automatic diagnosis from connectomes alone in Chapter 7. In this appendix we additionally demonstrate that confound-isolating cross-validation using these test sets is valid as an unconfounded estimate of prediction accuracy using MLPs in the single and multi-task setting.

A.2 - Methods

The MLPconf model was used to predict each condition from confounds (age, head motion, global signal, scanning site and sex) alone first in the single task setting to establish a baseline, and then in the multi-task setting across conditions. The models were evaluated using confound-isolating cross-validation (described in chapter 5), and training was conducted as is described in chapter 6.

A.2.1 - Architectures

MLPconf

The MLPconf model is similar to MLPconn (described in chapter 6), but with the input layer modified to take a 1×58 vector consisting of the confounding variables and the dimensions of the following layers reduced accordingly. The model configuration is 58-32-8-2. The confounds vector consists of the numerical variables (age, head motion, and mean connectivity) alongside the one-hot encoded categorical variables (sex and site).

A.3 - Results

A.3.1 - Confounds models consistently predict at chance level for most conditions using a neural network on balanced test sets

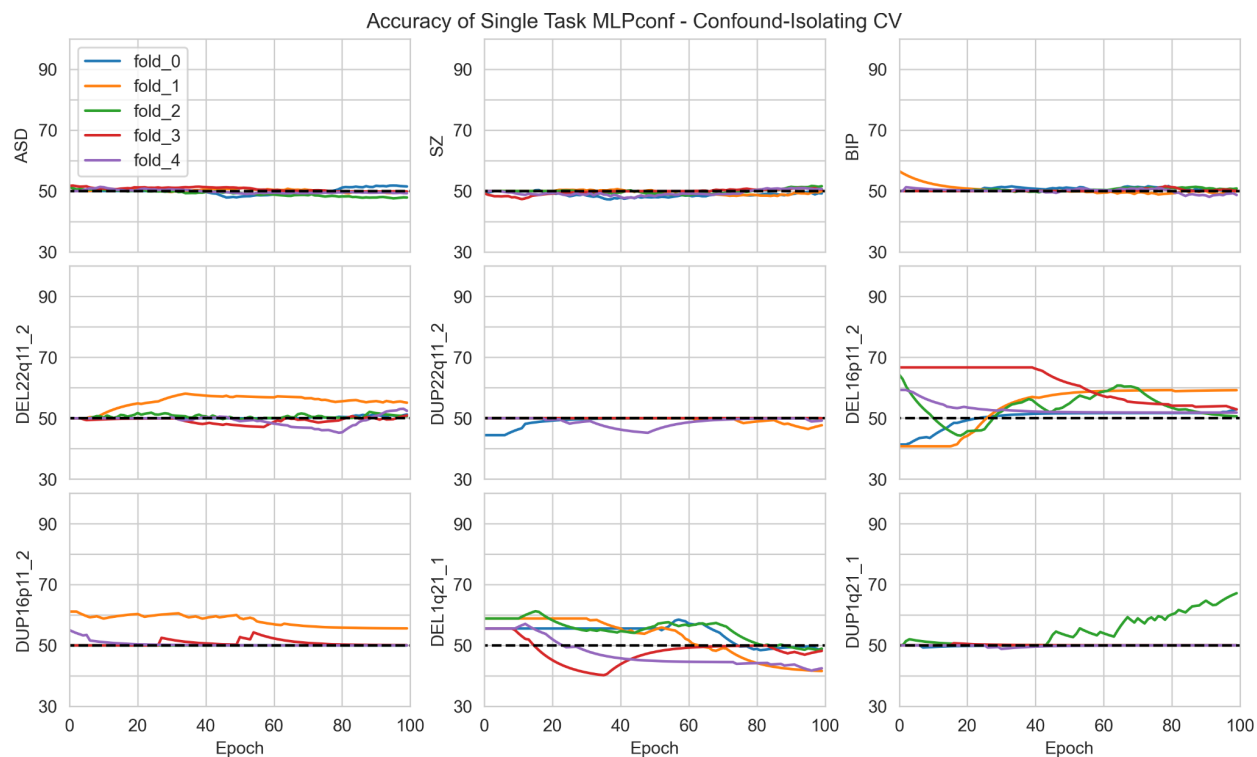


Figure 17 - Accuracy during training of MLPconf in the single-task setting. The x axis represents epochs of training and the y axis shows prediction accuracy. Each fold of cross-validation is indicated by a different coloured line for each condition. Plotted accuracy is smoothed using exponential moving averages for clarity.

We aimed to evaluate if the balanced test sets were valid in terms of preventing prediction from confounds using a neural network architecture, which is more flexible than the simple ML models. We trained a simple feedforward neural network on confounds to predict each condition independently (single-task setting). The mean final accuracy across folds and conditions was 51% with DEL16p11.2 reached the highest accuracy at 53%. For most cases, each fold of training results in chance level prediction. However, in one fold each DEL22q11.2, DEL16p11.2, DUP16p11.2, and DUP1q21.1 depart from chance with DUP1q21.1 reaching the highest accuracy 68% in fold_2. The mean variance in training accuracy across folds and conditions was 15%. The balanced test sets perform better as an unconfounded evaluation in the single-task neural network setting than in the intra-site benchmark.

A.3.2 - Confounds models predict at chance level using multi-task learning averaged across balanced test sets

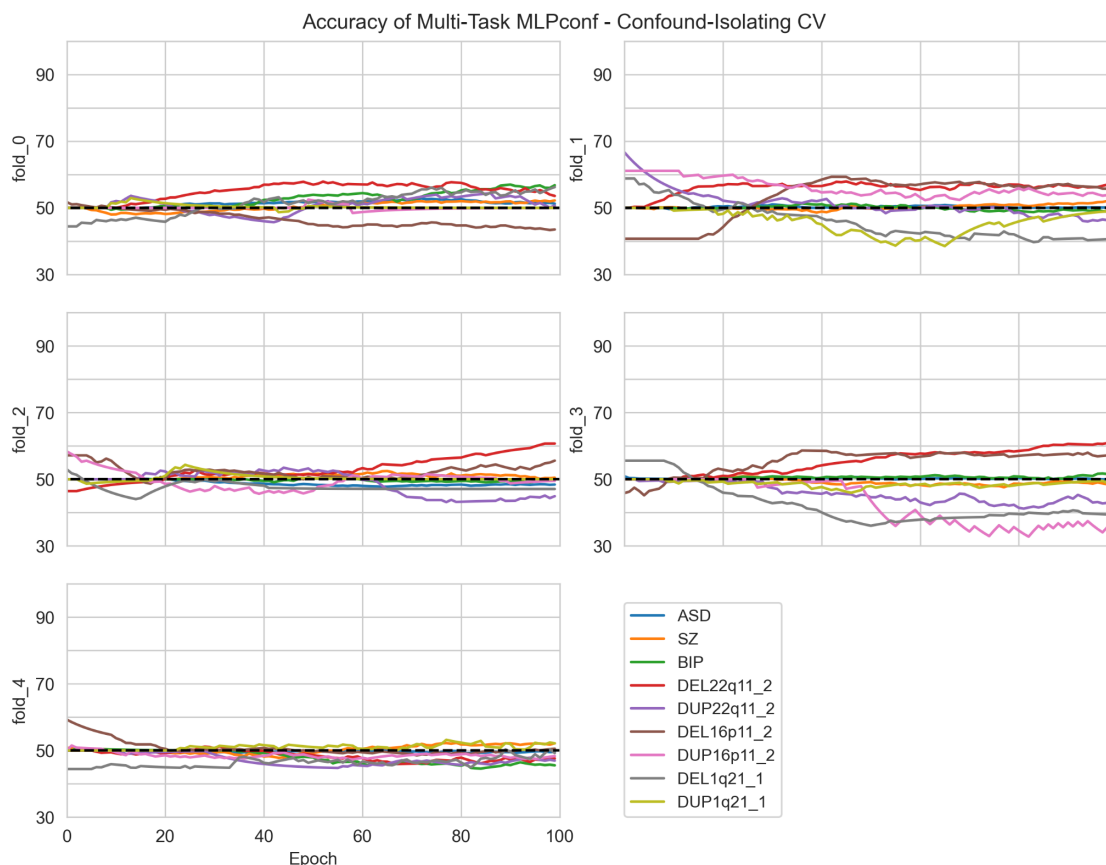


Figure 18 - Accuracy during training of MLPconf in the multi-task setting. The x axis represents epochs of training and the y axis shows prediction accuracy. Each condition is indicated by a different coloured line for each fold of cross-validation. Plotted accuracy is smoothed using exponential moving averages for clarity.

We next aimed to evaluate the balanced test sets using a neural network architecture in the multi-task setting. We trained a multi-task version of the neural network, where all the tasks are learned concurrently, on confounds alone to predict the conditions. The mean final accuracy across folds and conditions was 50%. DEL 22q11.2 reached the highest accuracy at 56%. The mean variance in prediction accuracy for each task across folds was 46%, much higher than in the single task setting. On average, the balanced test sets perform well as an unconfounded evaluation in the multi-task neural network setting.

A.4 - Discussion

Applying confound-isolating cross-validation to evaluate prediction of 9 CNVs and psychiatric conditions from confounds, we observed that the proposed approach successfully prevents

generalisation to the set set i.e. the balanced test sets are valid for unconfounded estimation in the single and multi-task setting using MLPs.

In the single task setting, the model makes above chance level prediction for some conditions in certain test sets (e.g. fold_2 of DUP1q21.1). However, the prediction doesn't remain above chance level when using multi-task learning for those conditions in the same test. There is a wide variance in prediction accuracy from confounds across cases in the multi-task learning setting, which results in close to chance level on average. Whether this effect is due to successful confound-isolating cross-validation or negative transfer between tasks is not clear. Either way, multi-task learning doesn't appear to present a deeper issue for confounding than single task learning.

Appendix B - PyNM

The following pages contain the publication accompanying the python package PyNM (Harvey and Dumas 2022) as published in the Journal of Open Source Software. See section 1.2.4 for a discussion.




PyNM: a Lightweight Python implementation of Normative Modeling

Annabelle Harvey ^{1,2} and Guillaume Dumas ^{2,3}

1 Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Université de Montréal, QC, Canada 2 Centre de Recherche du CHU Sainte-Justine, Université de Montréal, QC, Canada 3 Mila - Quebec AI Institute, Université de Montréal, QC, Canada

DOI: [10.21105/joss.04321](https://doi.org/10.21105/joss.04321)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Dan Foreman-Mackey](#) 

Reviewers:

- [@smkia](#)
- [@saigerutherford](#)

Submitted: 11 March 2022

Published: 08 December 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

The majority of studies in neuroimaging and psychiatry are focussed on case-control analysis ([Marquand et al., 2019](#)). However, case-control relies on well-defined groups which is more the exception than the rule in biology. Psychiatric conditions are diagnosed based on symptoms alone, which makes for heterogeneity at the biological level ([Marquand et al., 2016](#)). Relying on mean differences obscures this heterogeneity and the resulting loss of information can produce unreliable results or misleading conclusions ([Loth et al., 2021](#)).

Normative Modeling is an emerging alternative to case-control analyses that seeks to parse heterogeneity by looking at how individuals deviate from the normal trajectory. Analogous to normative growth charts, normative models map the mean and variance of a trait for a given population against a set of explanatory variables (usually including age). Statistical inferences at the level of the individual participant can then be obtained with respect to the normative range ([Marquand et al., 2019](#)). This framework can detect patterns of abnormality that might not be consistent across the population, and recasts disease as an extreme deviation from the normal range rather than a separate group.

PyNM is a lightweight python implementation of Normative Modeling making it approachable and easy to adopt. The package provides:

- Python API and a command-line interface for wide accessibility
- Automatic dataset splitting and cross-validation
- Five models from various back-ends in a unified interface that cover a broad range of common use cases
- Solutions for very large datasets and heteroskedastic data
- Integrated plotting and evaluation functions to quickly check the validity of the model fit and results
- Comprehensive and interactive tutorials

Statement of need

The basic idea underpinning Normative Modeling is to fit a model on the controls (or a subset of them) of a dataset, and then apply it to the rest of the participants. The difference between the model's prediction and the ground truth for the unseen participants relative to the variance around the prediction quantifies their deviation from the normal. While simple in concept, implementing Normative Modeling requires some care in managing the dataset and choosing an appropriate model.

In principle, any model that estimates both the mean and variance of the predictive distribution could be used for Normative Modeling. However, in practice, we impose more constraints.

First and foremost, the assumptions of the model must be met by the data. Second, it is important to distinguish between epistemic and aleatoric uncertainty. Epistemic or systematic uncertainty stems from how information about the distribution is collected, whereas aleatoric uncertainty is intrinsic to the distribution and represents the true variation of the population (Xu et al., 2021).

To the author's knowledge, PCNtoolkit (Marquand et al., 2021) is the only other available package for Normative Modeling. It implements methods that have been applied in a range of psychiatry and neuroimaging studies (Fraza et al., 2021; Kia et al., 2020, 2021; Rutherford, Fraza, et al., 2022), and is accompanied by thorough tutorials, a forum, and a framework for Normative Modeling in computational psychiatry (Rutherford, Kia, et al., 2022). While PCNtoolkit offers more advanced functionality, PyNM emphasizes being lightweight and easy to use, and implements different models than PCNtoolkit including a wrapper for the GAMLSS package from R, which is a powerful option for Normative Modeling (Dinga et al., 2021).

PyNM is intended to take users from their first steps in Normative Modeling to using advanced models on complex datasets. Crucially, it manages the dataset and has interactive tutorials – making it quick for new users to try the method either on their own data or on provided simulated data. The tutorials motivate the use of each model and highlight their limitations to help clarify which model is appropriate for what data, and built-in plotting and evaluation functions (Figure 1) make it simple to check the validity of the model output. The package includes five models from various backends in a unified interface, including a wrapper for GAMLSS (Rigby & Stasinopoulos, 2005) from R that is otherwise not yet available in python, and the selected models cover many settings including big data and heteroskedasticity.

Earlier versions of PyNM code were used in the following publications:

- Lefebvre et al. (2018)
- Maruani et al. (2019)
- Bethlehem et al. (2020)

Usage Example

```
from pynm.pynm import PyNM

# Load data
# df contains columns 'score', 'group', 'age', 'sex', 'site'
df = pd.read_csv('data.csv')

# Initialize pynm w/ data and confounds
m = PyNM(df, 'score', 'group', confounds = ['age', 'c(sex)', 'c(site)'])

# Run models
m.loess_normative_model()
m.centiles_normative_model()
m.gp_normative_model()
m.gamlss_normative_model()

# Collect output
data = m.data
```

Figures

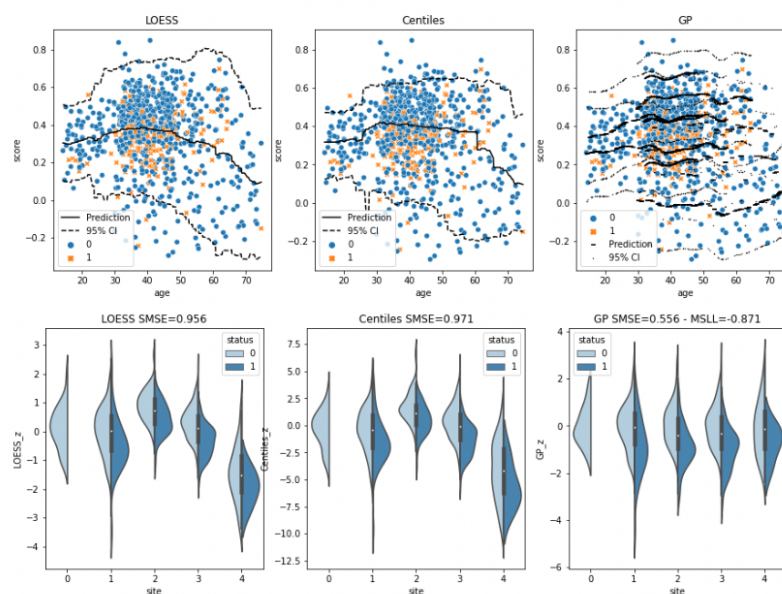


Figure 1: Output of built-in plotting function for model fit and residuals.

Acknowledgements

The development of this code has benefited from useful discussions with Andre Marquand, Thomas Wolfers, Eva Loth, Jumana Amad, Richard Bethlehem, and Michael Lombardo. The authors also want to thank the two reviewers Saige Rutherford (@saigerutherford) and Seyed Mostafa Kia (@smkia) for their insightful feedback.

Funding: This work is supported by IVADO, FRQS, CFI, MITACS, and Compute Canada.

References

- Bethlehem, R. A. I., Seidlitz, J., Romero-Garcia, R., Trakoshis, S., Dumas, G., & Lombardo, M. V. (2020). A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Communications Biology*, 3(1), 486. <https://doi.org/10.1038/s42003-020-01212-9>
- Dinga, R., Fraza, C. J., Bayer, J. M. M., Kia, S. M., Beckmann, C. F., & Marquand, A. F. (2021). Normative modeling of neuroimaging data using generalized additive models of location scale and shape. *bioRxiv*. <https://doi.org/10.1101/2021.06.14.448106>
- Fraza, C. J., Dinga, R., Beckmann, C. F., & Marquand, A. F. (2021). Warped bayesian linear regression for normative modelling of big data. *NeuroImage*, 245, 118715. <https://doi.org/10.1016/j.neuroimage.2021.118715>
- Kia, S. M., Huijsdens, H., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., & Marquand, A. F. (2020). Hierarchical bayesian regression for

- multi-site normative modeling of neuroimaging data. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, & L. Joskowicz (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2020* (pp. 699–709). Springer International Publishing. ISBN: 978-3-030-59728-3
- Kia, S. M., Huijsdens, H., Rutherford, S., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., & Marquand, A. F. (2021). Federated multi-site normative modeling using hierarchical bayesian regression. *bioRxiv*. <https://doi.org/10.1101/2021.05.28.446120>
- Lefebvre, A., Delorme, R., Delanoë, C., Amsellem, F., Beggiano, A., Germanaud, D., Bourgeron, T., Toro, R., & Dumas, G. (2018). Alpha waves as a neuromarker of autism spectrum disorder: The challenge of reproducibility and heterogeneity. *Frontiers in Neuroscience*, *12*. <https://doi.org/10.3389/fnins.2018.00662>
- Loth, E., Ahmad, J., Chatham, C., López, B., Carter, B., Crawley, D., Oakley, B., Hayward, H., Cooke, J., San José Cáceres, A., Bzdok, D., Jones, E., Charman, T., Beckmann, C., Bourgeron, T., Toro, R., Buitelaar, J., Murphy, D., & Dumas, G. (2021). The meaning of significant mean group differences for biomarker discovery. *PLOS Computational Biology*, *17*(11), 1–16. <https://doi.org/10.1371/journal.pcbi.1009477>
- 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>
- Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. *Biological Psychiatry*, *80*(7), 552–561. <https://doi.org/10.1016/j.biopsych.2015.12.023>
- Marquand, A. F., Rutherford, S., Kia, S. M., Wolfers, T., Frazza, C., Dinga, R., & Zabihi, M. (2021). *PCNToolkit (0.20)*. Zenodo. <https://doi.org/10.5281/zenodo.5207839>
- Maruani, A., Dumas, G., Beggiano, A., Traut, N., Peyre, H., Cohen-Freoua, A., Amsellem, F., Elmaleh, M., Germanaud, D., Launay, J.-M., Bourgeron, T., Toro, R., & Delorme, R. (2019). Morning plasma melatonin differences in autism: Beyond the impact of pineal gland volume. *Frontiers in Psychiatry*, *10*. <https://doi.org/10.3389/fpsy.2019.00011>
- Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554. <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
- Rutherford, S., Frazza, C., Dinga, R., Kia, S. M., Wolfers, T., Zabihi, M., Berthet, P., Worker, A., Verdi, S., Andrews, D., Han, L. K., Bayer, J. M., Dazzan, P., McGuire, P., Mocking, R. T., Schene, A., Sripada, C., Tso, I. F., Duval, E. R., ... Marquand, A. F. (2022). Charting brain growth and aging at high spatial precision. *eLife*, *11*, e72904. <https://doi.org/10.7554/eLife.72904>
- Rutherford, S., Kia, S. M., Wolfers, T., Frazza, C., Zabihi, M., Dinga, R., Berthet, P., Worker, A., Verdi, S., Ruhe, H. G., Beckmann, C. F., & Marquand, A. F. (2022). The normative modeling framework for computational psychiatry. *Nature Protocols*, *17*(7), 1711–1734. <https://doi.org/10.1038/s41596-022-00696-5>
- Xu, B., Kuplicki, R., Sen, S., & Paulus, M. P. (2021). The pitfalls of using gaussian process regression for normative modeling. *PLOS ONE*, *16*(9), 1–14. <https://doi.org/10.1371/journal.pone.0252108>

Appendix C - Supplementary Materials

C.1 - Demographics by Site

	N		Age		Global Signal		Motion	
	Total	(F)	Mean	(SD)	Mean	(SD)	Mean	(SD)
ABIDEII-BNI_1	16	(0)	21.44	(2.45)	0.37	(0.17)	0.23	(0.04)
ABIDEII-ETH_1	32	(0)	22.98	(4.65)	0.43	(0.14)	0.17	(0.04)
ABIDEII-GU_1	57	(0)	11.03	(1.55)	0.35	(0.11)	0.16	(0.04)
ABIDEII-IP_1	14	(0)	22.71	(6.75)	0.37	(0.11)	0.12	(0.04)
ABIDEII-IU_1	8	(0)	23.75	(6.69)	0.55	(0.24)	0.12	(0.03)
ABIDEII-KKI_1	27	(0)	10.31	(1.48)	0.36	(0.11)	0.16	(0.05)
ABIDEII-NYU_1	68	(0)	9.50	(3.85)	0.38	(0.14)	0.19	(0.03)
ABIDEII-OHSU_1	57	(0)	11.25	(2.12)	0.41	(0.11)	0.13	(0.04)
ABIDEII-OILH_2	23	(0)	22.52	(3.86)	0.46	(0.15)	0.12	(0.03)
ABIDEII-SDSU_1	43	(0)	12.98	(3.19)	0.33	(0.10)	0.11	(0.04)
ABIDEII-SU_2	16	(0)	11.03	(1.19)	0.38	(0.14)	0.17	(0.07)
ABIDEII-TCD_1	36	(0)	15.76	(3.04)	0.36	(0.13)	0.18	(0.03)
ABIDEII-UCD_1	21	(0)	14.75	(1.89)	0.40	(0.15)	0.14	(0.05)
ABIDEII-UCLA_1	20	(0)	11.41	(2.34)	0.40	(0.18)	0.14	(0.04)
ABIDEII-USM_1	27	(0)	20.90	(8.11)	0.36	(0.14)	0.17	(0.05)
ABIDEII-U_MIA_1	19	(0)	9.99	(2.06)	0.34	(0.13)	0.13	(0.04)
ADHD1	85	(42)	10.93	(1.72)	0.41	(0.10)	0.15	(0.04)
ADHD3	74	(34)	10.27	(1.35)	0.36	(0.13)	0.16	(0.05)
ADHD4	26	(13)	18.76	(3.21)	0.36	(0.11)	0.14	(0.05)
ADHD5	180	(66)	11.47	(2.94)	0.41	(0.12)	0.13	(0.04)
ADHD6	59	(23)	9.16	(1.22)	0.39	(0.14)	0.14	(0.04)
Cardiff	14	(7)	40.01	(8.03)	0.38	(0.15)	0.12	(0.05)
HSJ	53	(31)	31.80	(16.54)	0.38	(0.11)	0.21	(0.07)
KKI	36	(0)	10.27	(1.34)	0.30	(0.15)	0.19	(0.05)

LEUVEN_1	11	(0)	22.36 (4.50)	0.31 (0.09)	0.17 (0.04)
MAX_MUN	20	(0)	20.60 (11.99)	0.38 (0.14)	0.19 (0.06)
NYU	125	(0)	14.98 (6.53)	0.35 (0.12)	0.17 (0.04)
OLIN	17	(0)	16.82 (3.09)	0.40 (0.13)	0.24 (0.04)
SDSU	21	(0)	14.67 (1.59)	0.34 (0.12)	0.13 (0.06)
SZ1	84	(8)	33.57 (11.89)	0.38 (0.14)	0.22 (0.05)
SZ10	18	(3)	34.06 (9.89)	0.32 (0.09)	0.11 (0.04)
SZ2	82	(11)	33.35 (8.89)	0.41 (0.13)	0.14 (0.05)
SZ3	62	(24)	32.82 (7.74)	0.40 (0.16)	0.15 (0.05)
SZ4	50	(17)	31.70 (10.23)	0.43 (0.12)	0.15 (0.04)
SZ5	34	(8)	36.74 (9.98)	0.36 (0.10)	0.14 (0.03)
SZ6	70	(31)	30.37 (7.81)	0.37 (0.11)	0.15 (0.04)
SZ7	28	(8)	37.00 (8.33)	0.40 (0.12)	0.13 (0.04)
SZ8	28	(10)	31.07 (7.04)	0.36 (0.10)	0.12 (0.04)
SZ9	28	(4)	31.61 (9.04)	0.34 (0.13)	0.09 (0.04)
Svip1	86	(31)	26.03 (16.22)	0.32 (0.11)	0.21 (0.09)
Svip2	60	(30)	24.29 (14.02)	0.32 (0.11)	0.18 (0.06)
TRINITY	49	(0)	17.18 (3.64)	0.34 (0.10)	0.19 (0.04)
UCLA_1	52	(0)	13.58 (2.32)	0.37 (0.12)	0.16 (0.06)
UCLA_2	11	(0)	12.46 (1.93)	0.33 (0.13)	0.16 (0.04)
UCLA_CB	96	(47)	15.40 (6.98)	0.38 (0.13)	0.16 (0.07)
UCLA_DS1	154	(74)	32.26 (9.20)	0.40 (0.12)	0.15 (0.05)
UCLA_DS2	83	(27)	33.36 (9.35)	0.41 (0.15)	0.16 (0.06)
UKBB11025	18195	(9613)	63.42 (7.50)	0.42 (0.13)	0.19 (0.05)
UKBB11026	4685	(2582)	65.61 (7.53)	0.43 (0.13)	0.18 (0.05)
UKBB11027	8175	(4537)	64.77 (7.45)	0.42 (0.13)	0.19 (0.05)
UM_1	20	(0)	13.24 (2.92)	0.30 (0.09)	0.18 (0.04)
USM	64	(0)	21.86 (6.94)	0.39 (0.13)	0.18 (0.05)
YALE	33	(0)	12.76 (2.86)	0.31 (0.10)	0.17 (0.04)

Table 3 - Demographics by scanning site. The first two columns are the number of total subjects, and of female subjects (in parentheses) for each scanning site. The remaining columns show the mean age, global signal, and head motion, with standard deviation (in parentheses).

C.2 - Effect Size Table

Condition	Effect Size	p	FDR
DUP 15q13.3	0.11	0.86	0
DEL 2q13	0.11	0.87	0
DUP 15q11.2	0.16	0.036	0
DUP 2q13	0.18	3.1e-1	0
DUP 16p13.11	0.26	3.7e-1	0
DUP TAR	0.28	8.0e-1	0
DUP 13q12.12	0.31	9.7e-1	0
DEL 13q12.12	0.34	5.0e-1	0
DEL 15q11.2	0.20	1.6e-2	1
DUP 16p11.2	0.38	4.8e-3	7
DUP 22q11.2	0.43	4.2e-2	2
DEL 1q21.1	0.44	5.8e-3	12
DUP 1q21.1	0.49	1.2e-2	4
DEL 16p11.2	0.57	2e-4	273
DEL 22q11.2	0.66	2e-4	17
ADHD	0.15	2e-4	0
ASD	0.16	2e-4	106
SZ	0.31	2e-4	479
BIP	0.43	2e-4	57

Table 4 - Effect size of conditions on FC. Effect size is defined as the mean of the top decile connection in the FC profile. The second column shows the empirical p-value and the third how many individual connections in the FC profile were found to be significantly altered with respect to the control population after FDR correction ($q < 0.05$) (see Chapter 3.3.1).

C.3 - Confound-Isolating Cross-Validation

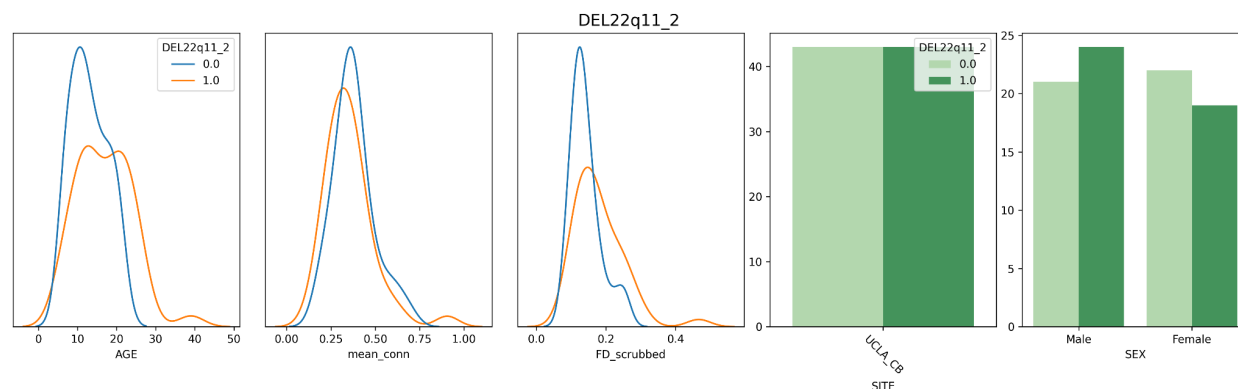


Figure 19 - DEL 22q11.2 - Distribution of confounds by class. Distribution of age, head motion, global signal, scanning site and sex plotted for DEL 22q11.2 carriers and controls (see Chapter 5.4).

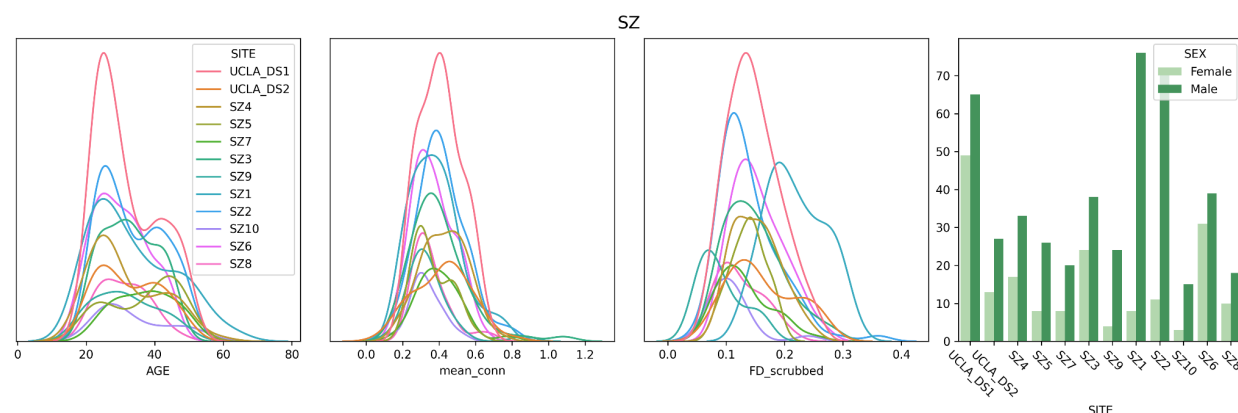


Figure 20 - SZ - Distribution of confounds by site. Distribution of age, head motion, global signal, scanning site and sex plotted for the SZ dataset by scanning site (see Chapter 5.4).

C.4 - Single vs Multi-task - Sex

Site	Single Task	Multi- Task
SZ3	57.14 (15.65)	77.14 (16.290)
SZ6	71.43 (20.20)	60.00 (18.63)
Svip2	55.00 (16.96)	65.00 (18.54)
ADHD6	40.00 (22.91)	52.50 (10.46)
HSJ	40.00 (20.00)	60.00 (24.04)
UCLA_CB	48.89 (11.33)	46.67 (24.09)
Svip1	54.00 (13.56)	76.00 (20.74)
ADHD1	67.27 (10.91)	74.55 (20.73)

ADHD3	58.33	(14.91)	48.33	(6.97)
ADHD5	58.75	(6.37)	55.00	(8.15)
UCLA_DS1	73.68	(4.71)	67.37	(8.65)
UKBB11026	84.00	(2.65)	90.10	(2.58)
UKBB11027	86.80	(1.69)	86.80	(1.82)
UKBB11025	83.90	(1.66)	87.80	(2.51)
Mean	62.80	(11.68)	67.66	(13.16)

Table 5 - Accuracy of sex prediction for each site using MLPs in single and multi-task learning. Standard deviation is reported in parentheses. See Chapter 6.3.1.

C.5 - Single vs Multi-Task Learning - Age

Site	Single Task		Multi- Task	
USM	14.07	(2.36)	11.67	(12.90)
SZ3	13.45	(5.320)	12.91	(5.51)
SZ6	11.57	(5.71)	8.21	(4.69)
Svip2	13.10	(5.00)	6.95	(2.88)
ADHD6	0.19	(0.09)	0.37	(0.21)
HSJ	21.97	(7.69)	26.17	(17.95)
SZ2	19.45	(9.52)	10.45	(7.75)
SZ1	23.76	(13.13)	22.56	(24.87)
UCLA_CB	5.71	(2.01)	8.33	(4.06)
Svip1	25.54	(12.55)	16.16	(6.69)
ADHD1	0.41	(0.09)	1.13	(0.40)
ADHD3	0.44	(0.23)	0.87	(0.84)
NYU	4.02	(1.03)	5.25	(3.44)
ADHD5	1.36	(0.57)	1.42	(0.76)
UCLA_DS1	13.41	(4.37)	11.27	(4.31)
UKBB11026	7.15	(0.95)	6.36	(0.64)
UKBB11027	6.25	(0.46)	4.94	(0.12)
UKBB11025	6.67	(0.57)	5.92	(0.55)
Mean	10.47	(3.98)	8.94	(5.48)

Table 6 - Performance of age prediction for each site using MLPs in single and multi-task learning. Loss is measured in MSE, standard deviation is reported in parentheses. See Chapter 6.3.2.

C.6 - Single vs Multi-Task Learning - Conditions

Condition	Single Task		Multi- Task	
DUP 1q21.1	69.05	(5.43)	48.81	(8.16)
DEL 1q21.1	68.17	(9.05)	56.80	(8.48)
DUP 22q11.2	61.53	(6.51)	58.06	(10.23)
DUP 16p11.2	67.56	(17.09)	53.11	(18.25)
DEL 16p11.2	69.00	(4.09)	66.38	(9.55)
DEL 22q11.2	80.81	(6.16)	71.24	(7.58)
ASD	63.51	(1.38)	59.75	(1.77)
SZ	72.86	(3.70)	72.38	(2.77)
BIP	63.53	(2.75)	73.58	(5.87)

Table 7 - Accuracy of automatic diagnosis for each condition using MLPs in single and multi-task learning. Standard deviation is reported in parentheses. See Chapter 7.3.1.