Challenges in Multi-Task Learning for fMRI-Based Diagnosis: Benefits for Psychiatric Conditions and CNVs Would Likely Require Thousands of Patients

Annabelle Harvey[1,2,3], Clara A. Moreau[4,5], Kuldeep Kumar[3], Guillaume Huguet[3], Sebastian G.W. Urchs[6], Hanad Sharmarke[2], Khadije Jizi[3], Charles-Olivier Martin[3], Nadine Younis[3], Petra Tamer[3], Jean-Louis Martineau[3], Pierre Orban[7,8], Ana Isabel Silva[9,10], Jeremy Hall[11], Marianne B.M. van den Bree[11,12,13], Michael J. Owen[11,12,13], David E.J. Linden[11,12,13,14], Sarah Lippé[15,3], Carrie E. Bearden[16,17], Guillaume Dumas[18,19], Sébastien Jacquemont[3,20]* and Pierre Bellec[2,1,15,19]*

* shared senior authorship

[1]Department of computer science and operational research, University of Montréal, Montréal, QC, Canada
[2]Centre de recherche de l'institut universitaire de gériatrie de Montréal, Montréal, QC, Canada
[3]Centre de recherche du CHU Sainte-Justine, Montréal, QC, Canada
[4]Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, USA
[5]Keck School of Medicine, University of Southern California, Marina del Rey, CA, USA
[6]Laboratory for Brain Simulation and Exploration, Université de Montréal, Montréal, QC, Canada
[7]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Montréal, QC, Canada
[8]Department of Psychiatry and Addictology, Université de Montréal, Montréal, QC, Canada
[9]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN, USA
[10]Neuroscience and Mental Health Innovation Institute, Cardiff University, Cardiff, UK
[11]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK
[12]Neuroscience and Mental Health Innovation Institute, Cardiff University, Cardiff, UK
[13]Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK
[14]Institute for Mental Health and Neuroscience, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, Netherlands
[15]Department of Psychology, Université de Montréal, Montréal, QC, Canada
[16]Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA
[17]Department of Psychology, University of California, Los Angeles, CA, USA
[18]Department of Psychiatry, Université de Montréal, Montréal, QC, Canada
[19]Mila – Québec AI Institute, Université de Montréal, Montréal, QC, Canada
[20]Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montréal, QC, Canada
Address correspondence to Pierre Bellec, PhD, at pierre.bellec@mila.quebec, or Annabelle Harvey, MSc, at annabelle.harvey@umontreal.ca.

Abstract

There is a growing interest in using machine learning (ML) models to perform automatic diagnosis of psychiatric conditions; however, generalising the prediction of ML models to completely independent data can lead to sharp decrease in performance. Patients with different psychiatric diagnoses have traditionally been studied independently, yet there is a growing recognition of neuroimaging signatures shared across them as well as rare genetic copy number variants (CNVs). In this work, we assess the potential of multi-task learning (MTL) to improve accuracy by characterising multiple related conditions with a single model, making use of information shared across diagnostic categories and exposing the model to a larger and more diverse dataset. As a proof of concept, we first established the efficacy of MTL in a context where there is clearly information shared across tasks: the same target (age or sex) is predicted at different sites of data collection in a large fMRI dataset compiled from multiple studies. MTL generally led to substantial gains relative to independent prediction at each site. Performing scaling experiments on the UK Biobank, we observed that performance was highly dependent on sample size: for large sample sizes (N>6000) sex prediction was better using MTL across three sites (N=K per site) than prediction at a single site (N=3K), but for small samples (N<500) MTL was actually detrimental for age prediction. We then used established machine learning methods to benchmark the diagnostic accuracy of each of the 7 CNVs (N=19-103) and 4 psychiatric conditions (N=44-472) independently, replicating the accuracy previously reported in the literature on psychiatric conditions. We observed that MTL hurt performance when applied across the full set of diagnoses, and complementary analyses failed to identify pairs of conditions which would benefit from MTL. Taken together, our results show that if a successful multi-task diagnostic model of psychiatric conditions were to be developed with resting-state fMRI, it would likely require datasets with thousands of patients across different diagnoses.

Keywords: Machine learning, multi-task learning, multi-site data, fMRI, CNVs, psychiatric conditions

1 - Introduction

There is a growing interest in using machine learning (ML) models to perform automatic diagnosis of psychiatric conditions. Unlike group-level mass-univariate analyses, ML models identify multivariate patterns that characterise a condition by learning to distinguish patients from control subjects at the individual level. While many studies have reported promising results (Iyortsuun et al., 2023), generalising the prediction of ML models to completely independent data can lead to sharp decrease in performance, as was recently evidenced for clinical trial stratification (Chekroud et al., 2024). This is due in large part to the massive biological heterogeneity that exists within the current diagnostic categories, which are based on behavioural symptoms alone (Pacheco et al., 2022). High rates of comorbidities (Katzman et al., 2017; McElroy, 2004; Simonoff et al., 2008; Tsai & Rosenheck, 2013) among psychiatric disorders, as well as genetic and symptom overlap (Romero et al., 2022), and shared neuroimaging signatures (Vanes & Dolan, 2021; Xie et al., 2023) supports the existence of latent factors that ignore diagnostic boundaries. Multi-task learning (MTL) is an ML framework that has the potential to improve prediction by characterising multiple related conditions with a single model, making use of information shared across diagnostic categories and exposing the model to a larger and more diverse dataset. Useful transdiagnostic information for MTL may also be found in rare genetic mutations, called copy number variants (CNVs), some of which confer a high risk for a range of psychiatric conditions (Marshall et al., 2017; Rees & Kirov, 2021; Sanders et al., 2019; Satterstrom et al., 2020). CNVs have a large impact on brain structure and function (Modenato et al., 2021; Moreau, Ching, et al., 2021; Moreau et al., 2020; Moreau, Raznahan, et al., 2021; Sønderby et al., 2022), which converges with neuroimaging brain signatures associated with psychiatric disorders (Moreau, Raznahan, et al., 2021). In this work, we aimed to assess the potential of MTL to exploit the relationships between a range of psychiatric and CNV conditions to improve the performance of automatic diagnosis. To this end, we compiled a resting state functional magnetic resonance imaging (rs-fMRI) dataset of 7 CNVs, which have never previously been studied in the ML context, and 4 psychiatric disorders.

ML models aim to identify patterns and mechanisms across regions in the brain that characterise a condition and provide scores for diagnosis at the individual level (Linn et al., 2016). While genetic variants conferring risk for psychiatric conditions have yet to be studied in the ML context, many studies have applied ML methods to automatically diagnose psychiatric conditions using rs-fMRI biomarkers, including schizophrenia (SZ) (Bassett et al., 2012; Kim et al., 2016; Venkataraman et al., 2012), autism spectrum disorder (ASD) (Abraham et al., 2017; Heinsfeld et al., 2018; Khosla et al., 2018; Nielsen et al., 2013; Traut et al., 2022), attention-deficit/hyperactivity disorder (ADHD) (Eloyan et al., 2012; J. Li et al., 2020; Z. Wang et al., 2023), and bipolar disorder (BIP) (Rashid et al., 2016; H. Wang et al., 2022). However, accuracy of automatic diagnosis reported in the literature should be interpreted with caution, as the majority of studies to date have been performed on data collected from a single site with

a small sample size (Orban et al., 2018). Sample size is a crucial factor in training ML models. Accuracy of prediction generally increases with the number of subjects in studies that examine the impact of sample size (Schulz et al., 2020; Traut et al., 2022). While more subjects allow the model to learn a better signature for a given condition, using small samples can yield deceptively high prediction accuracy due to overfitting (where a model memorises aspects of the dataset used to train it, but fails to generalise to new data). In a meta-analysis (including studies on ASD and SZ among other psychiatric diagnoses), Varoquaux found that studies with fewer subjects tended to report higher prediction accuracies (Varoquaux, 2018). Larger and more diverse samples are crucial to train properly evaluated models that can detect subtle patterns and generalise to the heterogeneity encountered in the clinical setting (Q. Ma et al., 2018; Traut et al., 2022). Given the evidence of latent factors shared across psychiatric conditions and CNVs, a logical next step is to develop ML models that can better exploit the available data by combining information across related categories.

MTL is an ML framework in which, rather than training a model on a single learning task (e.g. predicting ASD from rs-fMRI data), a model is trained on multiple related tasks concurrently. For example, predicting a diagnosis of ADHD as well as a diagnosis of ASD from resting-fMRI data (Huang et al., 2020). When the tasks are well grouped together, MTL can make 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. Although there are still few examples in neuroimaging, MTL has been applied across target clinical variables using rs-fMRI data (Rahim et al., 2017) and combined imaging modalities (Zhang et al., 2012), across timepoints to predict disease progression using cortical surface data (Zhou et al., 2013), across individuals to perform brain decoding using fMRI data (Marquand et al., 2014; Rao et al., 2013), across fMRI task conditions to predict intelligence quotient (IQ) (Xiao et al., 2020), and across sites (Hu & Zeng, 2019; Q. Ma et al., 2018) (Hu & Zeng, 2019; Q. Ma et al., 2018; Watanabe et al., 2014) and disease subtypes (X. Wang et al., 2015) to perform automatic diagnosis. Various deep learning architectures have also been applied to neuroimaging data using MTL (Dong et al., 2020; He et al., 2020; Liang et al., 2021; Ngo et al., 2020; Tabarestani et al., 2022; Yu et al., 2021). There are only two previous studies applying MTL across psychiatric conditions (Huang et al., 2022, 2020), the first examined ASD and ADHD and the second added SZ (prediction accuracy of 73.1, 72.7, and 84.9 respectively). In these studies, Huang and colleagues proposed the Multicluster Multigate Mixture of Experts (M-MMOE). The M-MMOE is a variant of the Multi-gate Mixture-of-Experts (MMOE) framework (J. Ma et al., 2018), in which MMOEs are combined across clusters of brain ROIs (see Supplementary Materials 10.3 for a more in depth description). While these results are in line with accuracies reported in the ML literature, they still fall short of being useful in clinical practice. Another limitation of these prior works is that they only combined a limited number of diagnostic categories in the MTL framework, in particular leaving out valuable but rare data on genetic risk for psychiatric conditions (CNVs) which might allow shared features that are subtle in psychiatric conditions to be learned more easily in highly impacted populations.

For this purpose, we used an rs-fMRI dataset consisting of subjects diagnosed with 7 CNVs and 4 psychiatric disorders, including a total of 2872 participants and 53 sites of data collection. We included CNVs that were previously found to be associated with psychiatric disorders: DEL 1q21.1, DUP 1q21.1, DEL 16p11.2, DUP16p11.2, DEL 22q11.2, DUP 22q11.2, and DEL 15q11.2 (Collins et al., 2022; Davies et al., 2020; Jønch et al., 2019; Marshall et al., 2017; Moreau et al., 2023; Sanders et al., 2015). In particular, DEL 22q11.2 and DEL 16p11.2 are rare examples of heritable (non de-novo) CNVs with severe clinical manifestations. Both variants have been found to have large clinical effect sizes (Crawford et al., 2019; Jonas et al., 2014; Moreau et al., 2023; Rees & Kirov, 2021; Willsey et al., 2022). 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). We included common psychiatric disorders, which have also been found to be associated with CNVs: autism spectrum disorder (ASD) associated with (16 different CNVs, schizophrenia (SZ) associated with 14 different CNVs (Satterstrom et al. 2020; Sanders et al. 2019; Marshall et al. 2017; Rees and Kirov 2021), and Bipolar (BIP) disorder and Attention-Deficit/Hyperactivity Disorder (ADHD) which are less frequently associated (Rees and Kirov 2021). The estimated prevalence of ASD is 1% of children worldwide (Zeidan et al., 2022)), of ADHD is 2.5% of the general population (Simon et al., 2009), of BIP is 1% worldwide (Merikangas et al., 2011), of SZ is 0.44% of the general population (Moreno-Küstner et al., 2018). Comorbidities between these conditions are extensive (Katzman et al., 2017; McElroy, 2004; Rösler et al., 2010; Sajatovic, 2005; Simonoff et al., 2008; Tsai & Rosenheck, 2013).

As a proof of concept, we first applied MTL in a context where there is clearly information shared across tasks: the same target (age or sex) is predicted at different sites of data collection where each site is treated as a task. Using the very large UK Biobank sample (30,185 subjects), we examined the impact of sample size on MTL accuracy. Next, we evaluated the potential benefits of MTL for prediction accuracy across the full set of psychiatric and genetic conditions in our dataset by treating each condition as a task. Finally, we explored the relationships between conditions by applying MTL to each pair using our standard model and several variant model architectures.

2 - Materials and Methods

2.1 - Ethics

The present secondary analysis project was approved by the research ethics review board at the Centre Hospitalier Universitaire Sainte-Justine.

		N		Age		FD			
	Condition	Total	(F)	Mean	(SD)	Mean	(SD)	**Sites**	**Dataset**
A	DEL 15q11.2	103	(55)	64.29	(7.44)	0.19	(0.06)	3	UKBB
	Controls	103	(55)	62.65	(7.51)	0.19	(0.05)	3	
	DUP 16p11.2	35	(14)	34.15	(19.53)	0.21	(0.09)	6	MRG, SVIP, UKBB
	Controls	35	(14)	32.04	(20.34)	0.18	(0.06)	6	
	DUP 22q11.2	22	(12)	39.43	(23.49)	0.19	(0.09)	5	DEFINE, MRG, UCLA,
	Controls	22	(12)	38.61	(25.81)	0.17	(0.06)	5	UKBB
	DEL 1q21.1	25	(12)	44.40	(18.87)	0.18	(0.07)	6	DEFINE, MRG, SVIP,
	Controls	25	(12)	50.89	(14.69)	0.21	(0.08)	6	UKBB
	DUP 1q21.1	19	(13)	50.86	(19.35)	0.21	(0.08)	7	DEFINE, MRG, SVIP,
	Controls	19	(13)	51.40	(22.31)	0.17	(0.04)	7	UKBB
	DEL 16p11.2	32	(13)	21.74	(20.14)	0.22	(0.09)	5	DEFINE, MRG, SVIP,
	Controls	32	(13)	31.64	(20.15)	0.19	(0.07)	5	UKBB
	DEL 22q11.2	43	(19)	16.86	(6.95)	0.18	(0.07)	1	UCLA
	Controls	43	(22)	13.00	(4.61)	0.14	(0.04)	1	
B	ADHD	223	(66)	14.71	(9.47)	0.15	(0.04)	7	ADHD-200, CNP
	Controls	353	183	17.68	(10.63)	0.14	(0.04)	7	
	ASD	472	(0)	14.71	(5.88)	0.17	(0.05)	28	ABIDE1, ABIDE2
	Controls	471	(0)	15.32	(6.58)	0.16	(0.05)	28	
	SZ	283	(73)	33.90	(9.22)	0.17	(0.06)	12	Orban, CNP
	Controls	355	(113)	31.85	(9.33)	0.14	(0.05)	12	
	BIP	44	(20)	35.02	(8.95)	0.17	(0.07)	2	CNP
	Controls	113	(52)	30.88	(8.59)	0.14	(0.04)	2	

Table 1 - Demographics by condition. A) Psychiatric CNVs, B) Psychiatric Conditions. The first two columns are the number of total subjects, and of female subjects (in parentheses). The intermediate columns show the mean age and framewise displacement (FD) (a measure of head motion, with standard deviation (in parentheses). The final column shows the number of scanning sites contributing to the dataset. See section 2.2 for dataset abbreviation definitions.

2.2 - Cohorts

The nine rs-fMRI datasets included four clinical CNV cohorts, five idiopathic neuropsychiatric datasets and one very large sample of unselected individuals. A majority of the datasets are compiled from different sites of data collection and studies. In total, rs-fMRI data from 2872 individuals were included, who were either neurotypical control subjects, individuals diagnosed with one of 7 CNVs associated with psychiatric disorders (Moreau et al., 2023), or one of 4 psychiatric disorders (ASD, SZ, BIP, ADHD) (see Table 1). The research ethics review boards of each relevant institution approved the study of the corresponding dataset.

2.2.1 - Clinical Genetic Datasets

Participants in the four clinical genetic rs-fMRI datasets were recruited for scanning based on the presence of a CNV regardless of the presentation of symptoms, along with matched control subjects. These four clinical CNV datasets included the Simons Variation in Individuals Project (SVIP) (Simons Vip Consortium, 2012), the DEFINE Neuropsychiatric-CNVs Project (DEFINE) (Cardiff, United Kingdom) (Drakesmith et al., 2019), the University of California, Los Angeles 22q11.2 CNV project (UCLA) (Jalbrzikowski et al., 2022; Lin et al., 2017; Schleifer et al., 2023), and the unpublished Montreal rare genomic disorder family project (MRG) (MRG, Canada).

2.2.2 - Psychiatric Conditions Cohorts

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

2.2.3 - Unselected Population

CNV carriers with available rs-fMRI data were identified in the UK Biobank (UKBB) (Sudlow et al., 2015) using PennCNV (K. Wang et al., 2007) and QuantiSNP (Colella et al., 2007) following previously published methods (Huguet et al., 2021; {Martineau et al., n.d.). The DNA was extracted from blood samples, the Affymetrix arrays were utilised, sharing common probes between them, with a scale of 50k on the UK BiLEVE Array and 450k on the UK Biobank Axiom Array (Wain et al., 2015).

2.3 - rs-fMRI Preprocessing

All datasets were preprocessed using the same parameters of NIAK (Bellec et al., 2012). The three first volumes of each run were suppressed to allow the magnetisation to reach equilibrium. Each fMRI dataset was corrected for inter-slice difference in acquisition time and the parameters of rigid-body motion were estimated for each time frame. Rigid-body motion

was estimated within as well as between runs. The median volume of one selected fMRI run for each subject was coregistered with a T1 individual scan, which was itself non-linearly transformed to the Montreal Neurological Institute (MNI) template (Fonov et al., 2011). The rigid-body transform, fMRI-to-T1 transform and T1-to-stereotaxic transform were all combined, and the functional volumes were resampled in the MNI space at a 3 mm isotropic resolution. The "scrubbing" method (Power et al., 2012) was used to remove the volumes with excessive motion (frame displacement greater than 0.5). The following nuisance parameters were regressed out from the time series at each voxel: slow time drifts (basis of discrete cosines with a 0.01 Hz high-pass cut-off), average signals in conservative masks of the white matter and the lateral ventricles as well as the first principal components (95% energy) of the six rigid-body motion parameters and their squares (Giove et al., 2009; Lund et al., 2006). The fMRI volumes were finally spatially smoothed with a 6 mm isotropic Gaussian blurring kernel. A more detailed description of the pipeline can be found on the NIAK website. Preprocessed data were visually controlled for the quality of the co-registration, head motion, and related artefacts.

2.4 - 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 (Dadi et al., 2020; Hahn et al., 2022; 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 functional connectivity (FC) in psychiatric disorders, both using ML (see previous references) as well as classical mass univariate regression (Bellec et al., 2015). 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.

2.5 - Multi-Task Learning

We explored using MTL to predict age and sex across sites, and to perform automatic diagnosis across conditions from connectomes using shared bottom neural network models. For age and sex prediction each site is treated as a task, and for automatic diagnosis each condition is treated as a task. In shared bottom models (often called hard parameter sharing), the first layers of the network are common to all tasks after which the model branches into a series of task-specific heads (see Figure 1B). We chose to implement this form of MTL rather than various soft-parameter sharing schemes, in which partially or entirely parallel networks have parameters regularised jointly, first because it is a very commonly used approach and, second, because the reduction in parameters and hence capacity is well suited to our high dimensional

data. We used a simple Multi-Layer Perceptron (MLP) architecture throughout our experiments, with either two outputs for binary classification tasks (MLPconn) or a single output for regression (MLPconn_reg), and added variants (described below) to explore the relationships between tasks. We also explored different parameters of the MLPconn and MLPconn_reg architectures as a sensitivity analysis, see Supplementary Materials 10.9. All models were implemented in Pytorch (Paszke et al., 2019) and the code for MTL was written using Snorkel (Ratner et al., 2017) as a reference.

The MLPconn model is an MLP with the following configuration: 2080-256-64-2. 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 & Szegedy, 2015) is applied after each 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.

The MLPconcat model is exactly the same as the MLPconn model, with the input layer adapted to take as input 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.

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.

The convolutional neural network (CNN) model is adapted from (Leming & 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 & Szegedy, 2015) is applied after each layer. This implementation of a CNN is not designed to take into account spatial or functional relationships between regions, see Supplementary Materials 10.10 for a more in depth discussion and comparison with other architectures.

The shared middle (SM) model is a variant of the shared bottom framework, in which each task has its own specific input layer, followed by layers shared across tasks, and finally the task-specific head (see Figure 1C). Specifically, it has the configuration 2080-256-256-64-2: the input to the model is the upper triangular 1 x 2080 connectome vector, the first layer with 256

units is task specific, followed by two shared layers with 256 and 64 units, then a task specific head with two output units. Batch normalisation (Ioffe & Szegedy, 2015) is applied after each layer.

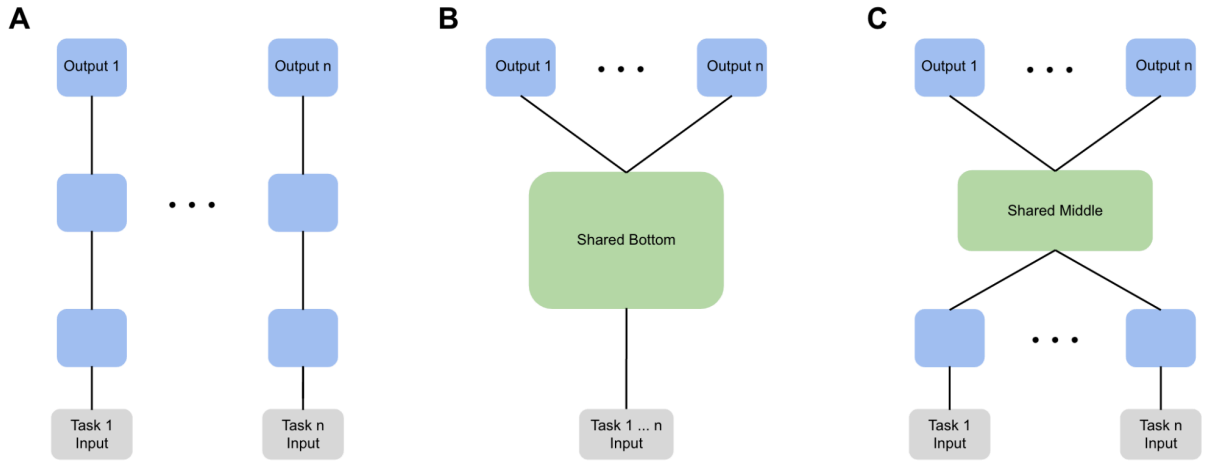

Figure 1 - A) Single task learning, B) Shared bottom model, C) Shared middle model.

2.6 - Training

We trained the MTL models as follows for each epoch: first, the batches of data are pooled across tasks and shuffled (see Figure 2); next, each batch is passed through the path it is associated with (through task-specific and shared layers), the loss is calculated and back propagated through the same 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 independent network. We used a small batch size (8) since we included small datasets, and models were trained for 100 epochs, roughly 50 epochs past observing plateaus in the single task setting. We used the Adam optimizer (Kingma & 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 trained using the cross-entropy loss after applying the softmax function, and the regression tasks with the mean squared error (MSE) (average of the squared differences between the predicted and actual scores). Classification tasks were additionally scored using prediction accuracy (number of correctly classified subjects divided by the total number of subjects), as well as the Area Under the Receiver Operating Characteristic (AUC) and F1-scores (see Supplementary Materials 10.5).

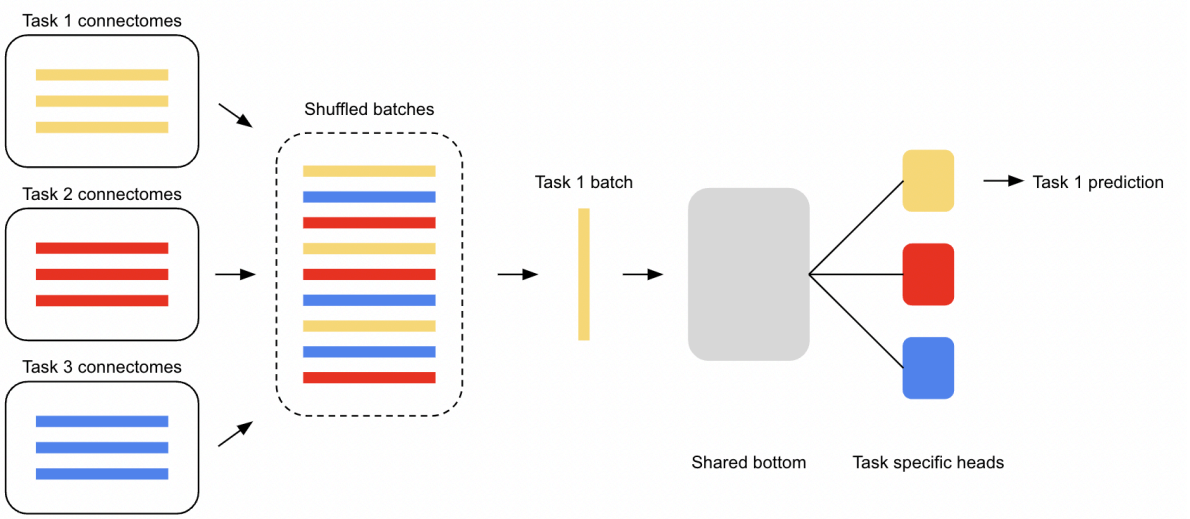

Figure 2 - The training process for MTL using the shared bottom model illustrated for three tasks. Connectomes from each task (conditions or sites of data collection) are shuffled and then fed to the model. Batches are fed to the model one at a time, and a batch from a given task is fed to its respective task specific head.

2.7 - Predicting Sex & Age

Before we delved into the complexity of MTL across CNVs and psychiatric conditions, we evaluated the benefit of using multi-task learning where heterogeneity between tasks is only due to sites. We treated each site of data collection as a task and predicted the same target (either sex or age) across them, the prediction was performed from the connectomes alone and evaluated using 5-fold cross validation. 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.

We first applied this approach using only the three sites from UKBB, since they have sample sizes (4569, 7943 and 17673) that are large enough that we could systematically quantify both the impact of sample size (similar to the experiments of (Schulz et al., 2020)) and the impact of MTL. Here we looked at two scenarios, in the first we sampled K subjects from each site and compared the prediction of single task models to a multi-task model across the three sites, effectively tripling the number of subjects available to the multi-task model. In the second, we compared a multi-task model across a sample of K subjects from each site to a single task model (on the largest site) with 3xK subjects, effectively keeping the sample size the same between the multi-task and single task model.

Next, we used the control subjects from each site of data collection in our sample that had at least 30 such participants. We subsampled 50 subjects from each of the very large UKBB datasets (sample sizes of 4569, 7943 and 17673) to place them within the range of the other sites.

For the sex prediction task, we excluded sites that had an insufficient number of female subjects (NYU, SZ1, SZ2, USM) (see Table 2 and figures in Supplementary Materials 10.7.1). For both the sex and age prediction, we performed ablation studies in which each site was dropped from the set of tasks as a sensitivity analysis, see Supplementary Materials 10.8.

Site	N	Female	%	Age	Dataset
ADHD1	54	35	65	10.93 (1.63)	ADHD-200
ADHD3	56	26	46	10.22 (1.27)	ADHD-200
ADHD5	77	39	51	12.25 (3.12)	ADHD-200
ADHD6	39	18	46	9.30 (1.25)	ADHD-200
HSJ	39	25	64	34.03 (16.10)	MRG
NYU	66	0	0	15.68 (6.22)	ABIDE1
SZ1	42	3	7	34.05 (10.90)	Orban
SZ2	41	2	5	31.54 (8.68)	Orban
SZ3	31	15	48	31.00 (8.19)	Orban
SZ6	35	12	34	29.03 (8.48)	Orban
Svip1	48	18	38	28.25 (16.56)	SVIP
Svip2	36	17	47	24.62 (12.44)	SVIP
UCLA_CB	43	22	51	13.00 (4.62)	UCLA
UCLA_DS1	94	43	46	31.10 (8.72)	CNP
UKBB11025	17673	9342	53	63.43 (7.50)	UKBB
UKBB11026	4569	2504	55	65.52 (7.54)	UKBB
UKBB11027	7943	4414	56	64.80 (7.45)	UKBB
USM	30	0	0	20.76 (7.21)	ABIDE1

Table 2 - Demographics of control subjects by scanning site. Number of total and female subjects, percentage female by site, and mean age in years with standard deviation in parentheses.

2.8 - Class Imbalances

The CNV datasets have major class imbalance, with far more controls than case subjects. Major class imbalances are problematic for predictive modelling; therefore for this context we created datasets using the General Class Balancer algorithm (Leming et al., 2020) which were balanced exactly with respect to diagnostic classes and approximately with respect to the distribution of confounding variables inside each diagnostic class (age, head motion, global signal, scanning site, and sex). The General Class Balancer algorithm exactly matches categorical variables, while continuous confounds are matched by recursively quantizing into smaller and smaller bins until subjects can be matched across bins while the distributions of the confound between classes are not found to be statistically different using a Mann Whitney U-test. For most of the

CNVs, we applied General Class Balancer with no modifications. The General Class Balancer algorithm repeatedly failed to find a match for a specific subject with DUP 16p11.2 when launched with different random seeds, we hand-selected 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; in this case we used all the subjects available without applying General Class Balancer. 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. The distribution of confounding variables for each of the balanced datasets are presented in Supplementary Materials 10.7.2

2.9 - Predicting CNVs & Psychiatric Conditions

In order to establish a baseline with which to compare our MTL results, we first performed automatic diagnosis in the single task setting (see Figure 1A), in which each task is learned by an independent model. In addition to the MLPconn model (described above), we evaluated three well-performing (Dadi et al., 2019) ML algorithms implemented in scikit-learn (Pedregosa et al., 2011): Support Vector Classifier (SVC) (linear kernel, C=100, and L_2 penalty), Logistic Regression (LR) (L_2 penalty), and Ridge Regression (Ridge). Next, we trained the MLPconn model in the MTL setting across all 7 CNVs and 4 psychiatric conditions. The models were evaluated using intra-site cross-validation (Orban 2018), in which the model is exposed to identical sites of data collection during training and testing to account for site effects. Specifically, five non-overlapping folds of training and test groups are built for each dataset such that they have roughly the same proportion of cases and controls from each site. Both the training and test groups feature every available site at each fold. The reported accuracy is the average of the model's performance across all folds.

2.10 - Study of Task Relationships

We added a fine grained analysis to explore the task relationships between the 7 CNVs and 4 psychiatric conditions by training the conditions together pairwise using our primary model (MLPconn) and four variations that explored different model capacity (MLPconn_deeper), input data (MLPconcat), encoder type (CNN), and parameter sharing scheme (SM). This framework allowed us to characterise whether relationships between tasks behaved differently depending on the context. Each model was first evaluated in the single task setting to establish a baseline.

3 - Results

3.1 - Multi-task learning of sex and age prediction across sites

3.1.1 - Multi-task learning across sites improves sex prediction in UK Biobank

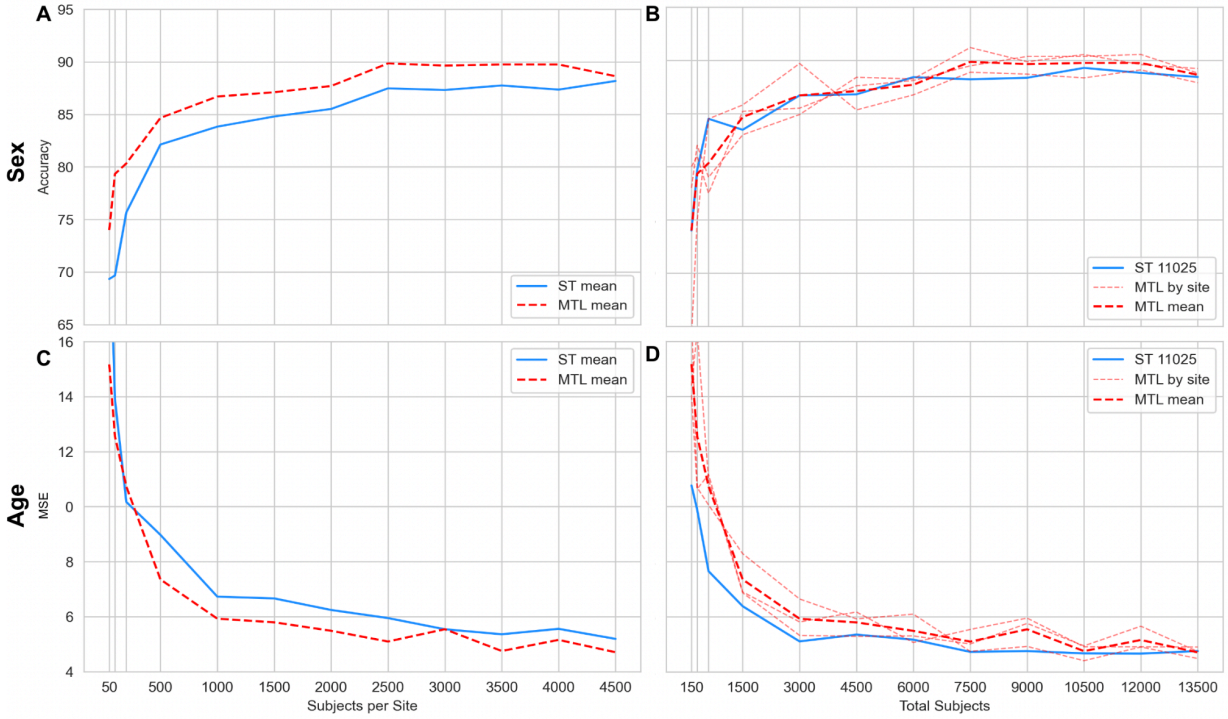

Figure 3. Prediction performance across sample sizes in the UKBB of single task (ST) vs. multi-task learning (MTL) applied to a common target (age or sex) across sites. In the first setting (A & C) we kept the same subjects across the ST & MTL model; i.e. for a sample of K subjects per site and ST model, the MTL model has access to 3 x K subjects in total. In the second setting (B & D) we aligned the sample size across models by comparing a MTL model with K subjects (across 3 sites) to a ST model with K subjects (using only the largest UKBB site where this was possible). A & C) Mean accuracy of sex prediction (A) or Mean Squared Error (MSE) of age prediction (C) across sites for ST and MTL models. B & D) Accuracy of sex prediction (B) or MSE of age prediction (D); mean performance across sites using MTL model (red dashed line), performance of MTL model at each individual site (pale red dashed line), and of ST model on UKBB site 11025 (blue solid line).

We first evaluated the benefit of using MTL in a simple binary classification setting where heterogeneity between tasks is only due to sites. We treated each site of data collection as a task and predicted sex across them. We applied this approach using only the three large sites (n = 4569, 7943 and 17673) from UKBB in order to compare single task learning and MTL at a range of sample sizes. First, we sampled K subjects from each site and compared the prediction of single task models to a multi-task model across the three sites (see Figure 3A). There was a clear

gain in performance for MTL in this setting, which could be observed for all sample sizes, although becoming small for N>4000. However, in this set up the MTL model effectively has access to triple the sample size which in and of itself improves prediction, therefore we also compared a MTL model across a sample of K subjects from each site to a single task model (only on the largest site which possessed sufficient subjects) with 3 x K subjects, effectively keeping the sample size the same across settings (see Figure 3B). For large sample sizes (N>7500) multi-task learning out-performed single task learning, meaning that combining heterogeneous data intelligently can actually improve the quality of prediction. Overall, multi-task learning across sites seems to be beneficial for sex prediction in the UK Biobank, although the largest gains are due to increased sample size.

3.1.2 - Multi-task learning across sites improves age prediction in UK Biobank for large sample sizes

We repeated the previous experiment using the three large UKBB sites, but this time using age as a prediction target, which is continuous and more challenging. In the first setting in which we sampled K subjects from each site for the single task models and used the same subjects (3 X K) for the MTL model (see Figure 3C), there was an advantage for MTL, but only for a relatively large sample size (N > 500 per site, 1500 total). For smaller sample sizes, single task and multi-task models had very similar performance. In the second setting in which we kept sample size fixed by comparing a MTL model across a sample of K subjects from each site to a single task model with 3 x K subjects (see Figure 3D), it became apparent that MTL models underperformed compared to the single task model, with the gap in performance decreasing with sample size, and becoming very small for N > 10k. Overall, MTL across sites seemed beneficial for age prediction across the UKBB sites, but only when MTL offered a boost in sample size relative to single task learning.

3.1.3 - Multi-task learning across sites improves sex prediction across cohorts

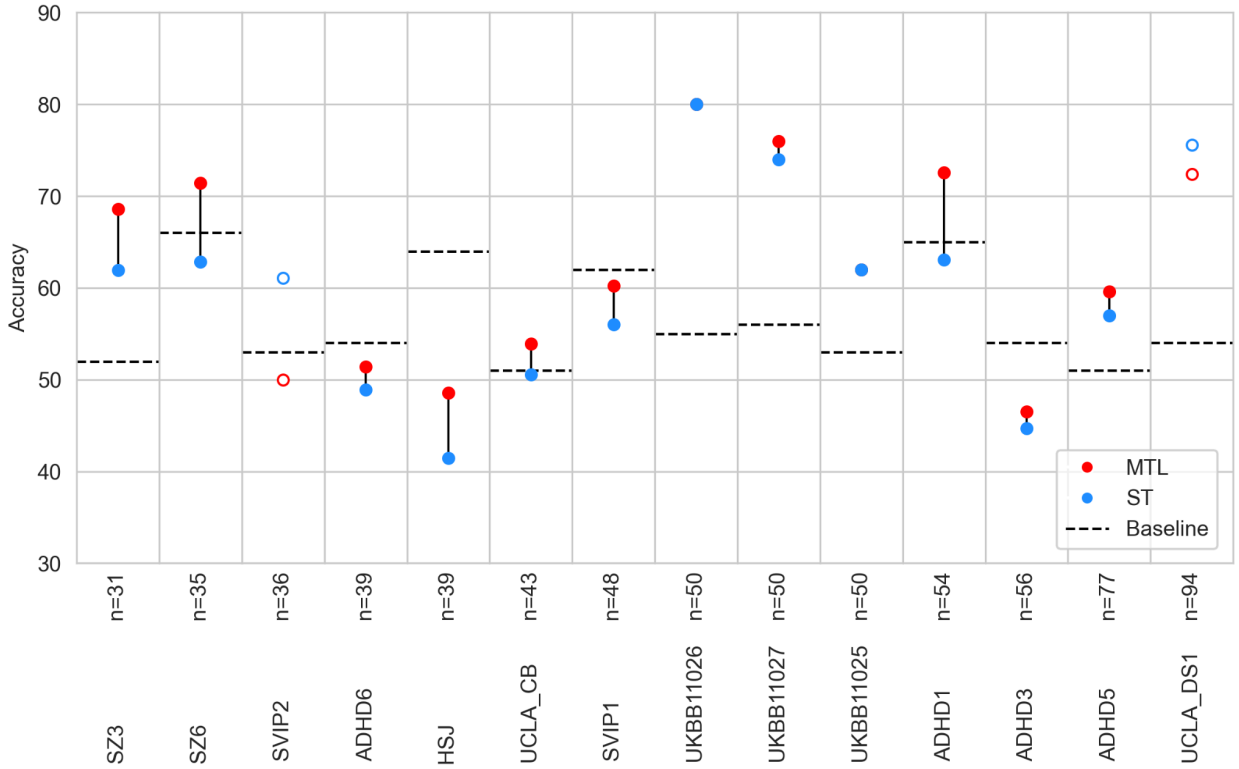

Figure 4 - Accuracy of sex prediction using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents different data collection sites included as prediction tasks. Sites are ranked by sample size, with the largest to the right. The y axis shows the accuracy of prediction, chance level of prediction is indicated by a black dashed line. For each task, the red point shows prediction using the MLPconn architecture in MTL, and the blue point shows prediction on the task trained independently using the MLPconn architecture. Where the red point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL prediction outperformed the ST, points were filled and connected by a line, and otherwise they were not.

Next, we used the control subjects from each site of data collection in our sample that had at least 30 participants and predicted sex across them using the MLPconn model, excluding sites that had an insufficient number of female subjects (NYU, SZ1, SZ2, USM). We subsampled 50 subjects from each of the very large UKBB datasets (n = 4569, 7943 and 17673) to place them within the range of the other sites. In this setting, MTL effectively pools subjects for a larger sample size at the price of greater heterogeneity. Prediction accuracy improved for MTL in a majority of sites (10 out of 14) (see Figure 4). The mean accuracy in the multi-task setting (62.4) outperformed that of the single-task (60.0) although not significantly (Wilcoxon's signed rank test) and with larger standard deviation (13.6 vs. 12.8) (see Supplementary Materials 10.4 - Figure 13 for the distribution across folds). Overall, MTL across heterogeneous sites of data collection benefitted accuracy for sex prediction.

3.1.4 - Multi-task learning across sites improves age prediction in a varied collection of samples

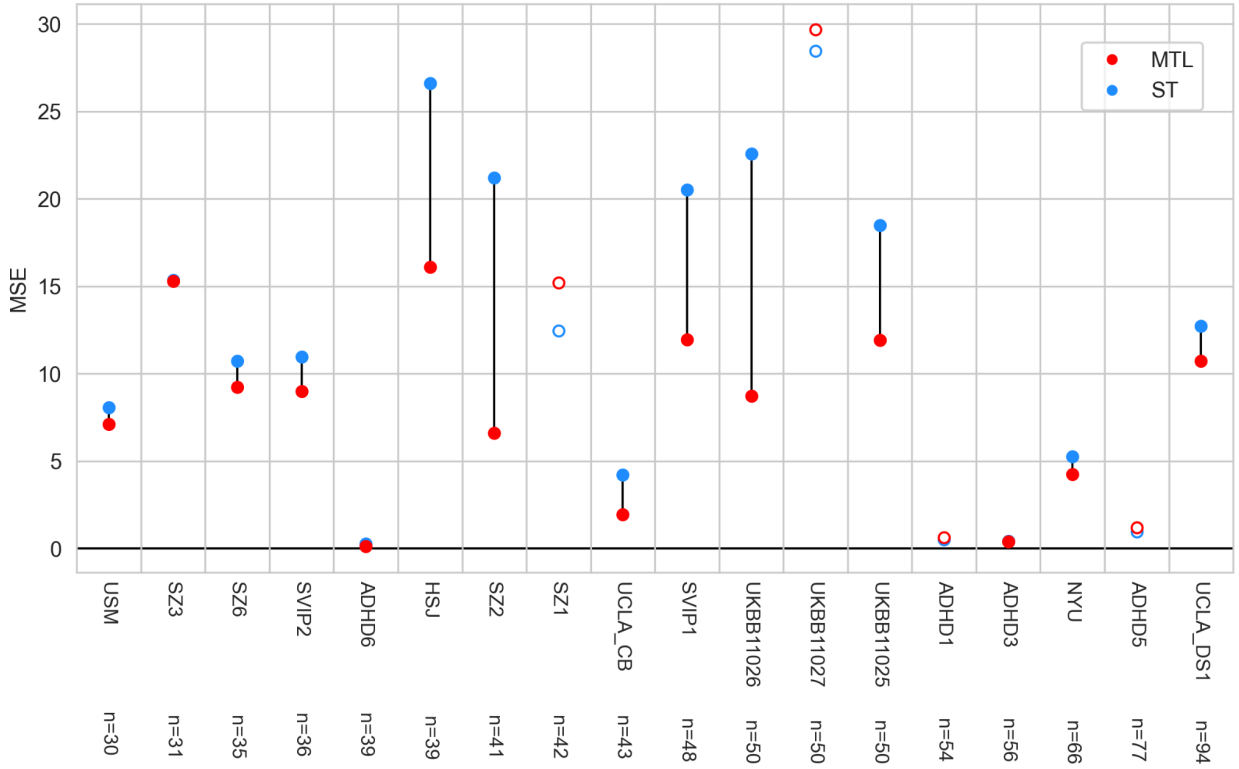

Figure 5 - Mean Squared Error (MSE) of age prediction using single (ST) vs. multi-task learning (MTL) in a varied collection of samples. The x axis represents different sites of data collection included as prediction tasks, sites are ranked by sample size, with the largest to the right. The y axis shows the prediction error. For each task, the red point shows prediction using the MLPconn_reg architecture in multi-task learning and the blue point shows prediction on the task trained independently using the MLPconn_reg architecture. Where the blue point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL prediction achieved lower loss than the ST, points were filled and connected by a line, and otherwise they were not.

Next, we predicted age across each site of data collection (control subjects only) in our sample that had at least 30 participants using the MLPconn_reg model, again subsampling 50 subjects from each of the very large UKBB datasets. Each site of data collection consisted of subjects with markedly different age ranges (see Table 2), so we expected this objective to be more difficult than sex prediction since MTL in this setting essentially works as a trade off between effectively increasing sample size at the cost of increased heterogeneity. Prediction improved for a large majority of sites (14 out of 18) (see Figure 5). The mean loss in the multi-task setting (8.9 $years_2$) significantly outperformed (Wilcoxon's signed rank test $p<2e-16$) that of the single-task (12.2 $years^2$), but with a larger standard deviation (6.0 vs 4.4) (see Supplementary section 10.4 - Figure 14 for the distribution across folds). Overall, MTL benefited prediction, even when the target of prediction was heterogeneously distributed across sites.

3.2 - Multi-task learning fails to improve automatic diagnosis across psychiatric conditions and genetic variants

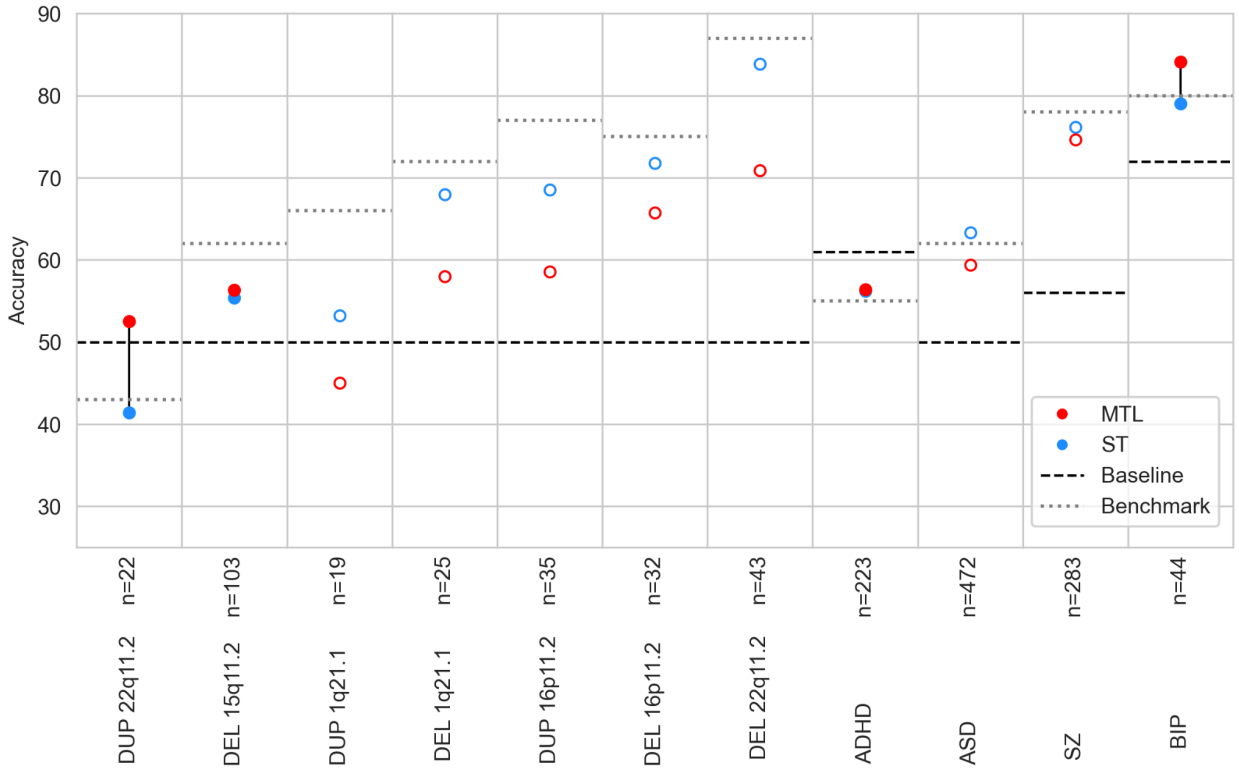

Figure 6 - Accuracy of automated diagnosis using single (ST) vs multi-task learning (MTL). For each task, the red point shows prediction using the MLPconn architecture in MTL and the blue point shows prediction on the task trained independently using the MLPconn architecture. Where either the blue point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL prediction outperformed the ST, points are filled and connected by a line, and otherwise they are not. The x axis represents different conditions included as prediction tasks. The y axis shows the accuracy of prediction, chance level of prediction is indicated by a black dashed line, and the best accuracy obtained in the single task learning benchmark (see Supplementary Materials 10.1) is indicated by a grey dotted line.

In order to establish a baseline with which to compare our MTL results, we first performed automatic diagnosis in the single task setting. In addition to the MLPconn model, we evaluated three ML algorithms: Support Vector Classifier (SVC), Logistic Regression (LR), and Ridge Regression (Ridge). 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,DUP 16p11.2, DEL 1q21.1) (see Supplementary Materials 10.1 - Figure 8). The other conditions were very challenging to predict, being near (or below) chance level. However, the prediction accuracy for CNVs broadly follows the trend of clinical effect size (Moreau et al., 2023) with CNVs with near chance level accuracy having small effect sizes. Overall, standard ML models seem capable of automatically diagnosing most of the CNVs and psychiatric conditions. Next, we aimed to improve the automatic diagnosis of the 7 CNVs and 4 psychiatric conditions by leveraging

shared information in datasets with limited sample size using a lightweight MTL framework to effectively increase the sample size available to the model. We trained the MLPconn model across all conditions and compared the performance to the same model trained on each condition independently. MTL outperformed single task learning in only 3 out of 11 conditions (see Figure 6), in the remaining cases performance accuracy actually decreased (see Supplementary section 10.4 Figure 15 for the distribution across folds).

3.3 - Task relationships are dominated by sample size and accuracy of diagnosis in single task setting

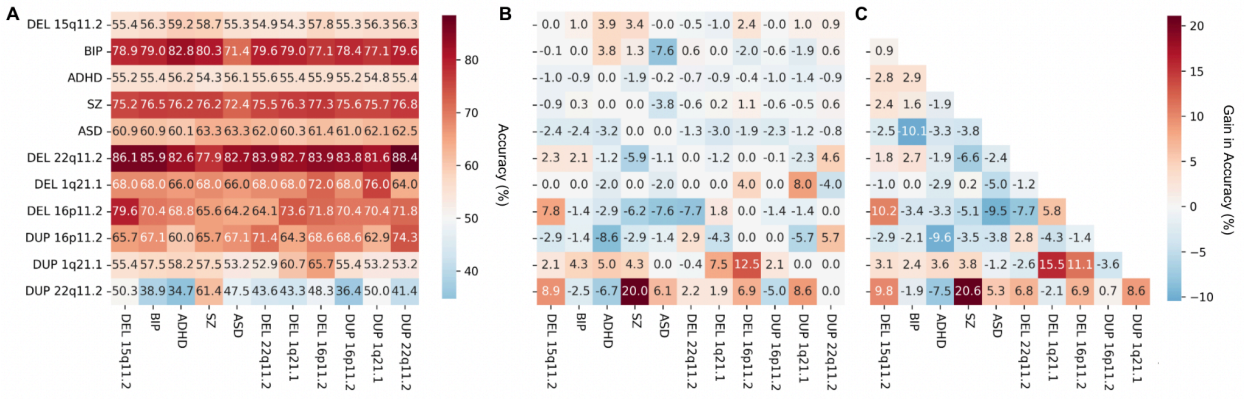

Figure 7 - In each matrix, the i,jth entry in the matrix is accuracy of condition in row i trained with condition in column j using the MLPconn model to perform automatic diagnosis. A: the matrix shows the raw accuracy achieved for each pair, the second matrix B represents the difference in accuracy from the single-task baseline, and the third C shows the overall gain relative to baseline for a pair ($B + B^T$).

We aimed to disentangle the complexity of performing automatic diagnosis on the 7 CNVs and 4 psychiatric conditions in the MTL setting by training the conditions together pairwise to gain insight on the relationships between tasks. Using the MLPconn model, we found that the conditions with high accuracy in the single task setting but small sample size (DEL 22q11.2 and DEL 16p11.2) suffered overall from being trained with a partner (see Figure 7B). In contrast, regardless of accuracy in the single task setting the conditions with larger sample size (SZ, ADHD, ASD) were not impacted by their partner (see Figure 7B). MTL appeared to benefit smaller sample size conditions with mid-range accuracy, but the results were not systematic. Certain pairs of conditions produced marked overall improvement in accuracy, notably DUP 22q11.2 + SZ and DUP 1q21.1 + DEL 1q21.1 (see Figure 7C). We repeated the study using four different models (MLPconn_deeper, MLPconcat, CNN, and SM) (see Supplementary Materials 10.2 - Figure 9), and correlated the matrix of change in accuracy from the single task baseline for each with that of the MLPconn model. We found a range of correlations (r = 0.52, 0.26, 0.18, 0.19 respectively), which showed that the relationships between tasks was not stable across contexts. Overall, relationships between tasks appear to be dominated by the available sample

size as well as the performance in the single task setting, rather than reflecting potentially meaningful biological relationships.

4 - Discussion

Using MTL to predict a common target (sex or age) across sites in the UK Biobank at a range of sample sizes, we found that for large sample sizes MTL can improve prediction even when compared to single task learning using the same number of subjects, but that MTL can be detrimental for small sample sizes (N<500). When we applied MTL to predict a common target (sex or age) across the full sample of sites in our dataset we found that it improved prediction. However, applying MTL across our diagnostic tasks (7 CNVs and 4 psychiatric conditions) was detrimental for performance overall. When we implemented MTL pairwise on the 7 CNVs and 4 psychiatric conditions in our dataset using our primary model and four variations, we found that the relationships between tasks were not stable across model architectures.

Repeating prediction across a range of sample sizes in the UK Biobank, we found that using MTL to predict a common target (sex or age) across data collection sites (3 x N subjects, N subjects per site from 3 sites) is not as strong as prediction at a single site with the same amount of data (3 x N subjects from 1 site), but is stronger in general than performing prediction at each site independently (N subjects from 1 site). These results are in line with findings by Schulz and colleagues (Schulz et al., 2020), who classified subjects into groups divided by sex and age using fMRI data and simple linear models in the UK Biobank. They found that prediction accuracy improved with increasing sample size, but did not investigate the effects of sites or MTL models. When predicting age, we saw that MTL did not necessarily improve prediction when sample sizes were small (N < 500). Standley and colleagues (Standley et al., 2019) also examined the impact of sample size on MTL, comparing a shared bottom model trained on a large dataset to the same model trained on a subsample (5% of the original data). They found that while MTL was overwhelmingly beneficial for prediction using the full sample, it hurt prediction overall when less data were used to train the model. Essentially, MTL allows a shared model to access more data than independent models could, but the combined data naturally introduces heterogeneity. While there is a possibility to benefit from this as regularisation across tasks, there exists a regime with modest amounts of data where MTL can be detrimental for performance.

When we applied MTL to predict a common target (sex or age) across the full sample of sites in our dataset we saw that performance was improved for a large majority of sites. While we expected that MTL would be beneficial for sex prediction, it was especially encouraging that prediction of age was improved since in light of the scaling experiment it was unclear if our sample sizes (N = 881 for MTL vs. N 30-94 for single task) could benefit from MTL. The most notable difference with the UK biobank experiment was that we had far more sites of data

collection (18 sites vs. 3 sites), and the sites were also more varied (retrospectively pooled across independent studies vs harmonised acquisitions from a single study). Our results suggest that there might be some benefit not only from increasing sample size but also from combining diverse information, in line with previous work on SZ diagnosis (Orban et al., 2018). Several studies have also reported improved prediction when using MTL across sites as applied to automatic diagnosis of SZ (Hu & Zeng, 2019; Q. Ma et al., 2018), and ADHD (Watanabe et al., 2014). While these studies did not include scaling experiments, in each case MTL was an improvement over pooling heterogeneous data. MTL learning thus appears as a viable alternative to data harmonisation across sites (El-Gazzar et al., 2023; Roffet et al., 2022; Y.-W. Wang et al., 2023), with the potential to improve prediction accuracy in the presence of moderate heterogeneity.

Applying MTL across our diagnostic tasks (7 CNVs and 4 psychiatric conditions), we saw that it was detrimental for performance overall. One possible conclusion is that the tasks are too heterogeneous to benefit from being learned together, and that rather than using a combined model across tasks we should pursue an approach that emphasises finding homogeneous subtypes within conditions, as has been explored in the context of high precision modelling (S. G. W. Urchs et al., 2020). It is however illuminating to look at our UK biobank scaling experiment and note that the difference between the combined sample size using MTL (N = 2872) and the single task sample sizes (N = 44 - 943) is not far from the threshold for MTL to consistently improve prediction of age (N > 500 single task, > 1500 MTL). As joint diagnosis is a much more complex objective (see Supplementary Materials section 10.6 for a more detailed analysis), it seems natural that it would require much more data to see a benefit from MTL. Although a speculation, our results thus suggest that the application of MTL to automated diagnosis across psychiatric conditions may be successful only if applied with over 500 patients per condition i.e. several thousand subjects in total. This parallels the conclusions of Marek and colleagues (Marek et al., 2022) who concluded that thousands of subjects are needed for reliable mass univariate brain/behaviour associations.

In the only comparable studies in the literature, Huang and colleagues (Huang et al., 2022, 2020) proposed a variant of the MMOE model (J. Ma et al., 2018) to perform joint diagnosis across psychiatric conditions. In their second study, their method showed marginal gains for each condition (N = 72 SZ, 358 ADHD, 505 ASD) relative to single task learning, while a shared bottom model reduced prediction accuracy for ADHD and SZ. We implemented additional experiments (see Supplementary Materials 10.3) to test if our results could be attributed to limitations of the shared bottom model, but found that it was not consistently outperformed by the MMOE. While there are surely improvements that can be made to the MTL framework, we emphasise that increasing the amount of data is a crucial aspect of future ML research in neuroimaging.

When we implemented MTL pairwise on the 7 CNVs and 4 psychiatric conditions in our dataset using our primary model and four variations, we found that the relationships between tasks were not stable across contexts. This is in line with findings by Standley and colleagues (Standley et al., 2019), who examined the matrix of prediction performance for tasks trained pairwise using a shared bottom model in different contexts (varying sample size and model capacity) and found that the relationships between tasks were dependent on the setting. This is contrary to our hypothesis that our MTL framework could uncover potentially meaningful biological relationships across the 7 CNVs and 4 psychiatric conditions in our dataset, which we would expect to be stable across different model architectures.

An important limitation of this study is that the datasets we had access to carry current biases in psychiatric research. It is well known that certain conditions (particularly ASD, ADHD, and SZ) are underrepresented among females, which reflects differences in understanding and diagnosing as well as prevalence (Attoe & Climie, 2023; Bierer et al., 2022; X. Li et al., 2022; Loomes et al., 2017; Pedersen et al., 2022).

We examined the important concept of using MTL to take advantage of information shared across biologically related conditions, possibly allowing automated diagnosis of conditions for which there is small amounts of available data and using traits that are easily learned in highly impacted CNV populations that could also apply to related psychiatric conditions. Although small sample size was a limitation of this study that was clear from the outset, the idea that the ability of MTL to make more efficient use of data would make it applicable to small datasets was not supported in our results. We found that applying MTL across conditions has potential, and although clever approaches to modelling, such as self-supervised (Caro et al., 2023) or transfer (Mahamud et al., 2023; Raghav et al., 2023) learning could potentially overcome the limitation of sample size, our results clearly show that increasing the amount of data is a crucial factor for improving prediction performance.

5 - Conclusion

In this paper, we examined the potential of MTL to combine multiple automatic diagnosis tasks in a large fMRI dataset compiled from multiple studies. In an initial proof of concept, we predicted a common target (sex or age) across sites of data collection, and showed that MTL can be beneficial for prediction accuracy, with the important caveat that benefits for age prediction only became apparent for large sample sizes (N>500 per site). We then benchmarked diagnostic accuracy of 7 CNVs and 4 psychiatric conditions using common machine learning methods, for each condition independently. None of the CNVs had previously been studied using machine learning, and prediction accuracy aligned with results from the literature otherwise. We then applied MTL to test if learning conditions with shared latent biological factors jointly could benefit prediction. Contrary to our hypothesis, we observed that MTL harmed prediction

accuracy overall. We further explored the behaviour of MTL by applying the framework on each pair of tasks and found that the relationships between tasks were not stable across varied contexts, which was evidence against the possibility of these relationships reflecting latent factors with biological meaning. Our scaling experiment with UK biobank suggests that MTL may become beneficial for automated diagnosis across neurodevelopmental conditions, but this would likely require larger sample sizes than what could be assembled in this study.

6 - Data and Code Availability

We thank all of the families at the participating Simons Variation in Individuals Project (SVIP) sites, as well as the Simons VIP Consortium. We appreciate obtaining access to imaging and phenotypic data on the SFARI Base. Approved researchers can obtain the Simons VIP population dataset described in this study by applying at https://base.sfari.org. We are grateful to all families who participated in the 16p11.2 European Consortium.
Data from UK Biobank were downloaded under the application 40980 and can be accessed via their standard data access procedure (see http://www.ukbiobank.ac.uk/register-apply). UK Biobank CNVs were called using the pipeline developed in Jacquemont Lab and described in https://github. com/labjacquemont/MIND-GENESPARALLELCNV. The final CNV calls are available from UK Biobank returned datasets (Return ID: 3104, https://biobank.ndph.ox.ac.uk/ukb/dset.cgi?id=3104). ABIDE1, ABIDE2, COBRE, ADHD-200, CNP, and 16p11.2 SVIP data are publicly available at: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html, http://fcon_1000.projects.nitrc.org/indi/abide/abide_II.html, http://schizconnect.org/queries/new, http://fcon_1000.projects.nitrc.org/indi/adhd200/, https://www. openfmri.org/dataset/ds000030/, and https://www.sfari.org/funded-project/simons-variation-in-individuals-project-simons-vip/. The 22q11.2 UCLA raw data are currently available by request from the principal investigator (CEB). Raw imaging data for the Montreal rare genomic disorder family dataset are currently available by request from the principal investigator (SJ). The Cardiff raw data are not publicly available yet; contact the principal investigator for further information (DEJL). All processed connectomes are available through a request to the corresponding author (AH). Code for all analyses are available online through the GitHub platform: https://github.com/harveyaa/neuropsych_mtl.

7 - Author Contributions

AH, SJ and PB designed the overall study and drafted the manuscript. AH performed all analyses. CAM and SGWU pre-processed 90% of all the fMRI data and generated all individual functional connectomes. HS performed the UKBB fMRI pre-processing. KK, KJ, C-OM, NY, PT, and SL recruited/scanned patients for the Montreal rare genomic disorder family dataset. GH, and J-LM performed the CNV calling. PO preprocessed the schizophrenia data. AIS, JH, MBMB, MJO, and DEJL provided the Cardiff CNV fMRI data. CEB provided the UCLA 22q.11.2 fMRI data. GD contributed to the interpretation of the data. PB and SJ contributed equally to this work as joint senior authors. All authors provided feedback on the manuscript.

8 - Funding

This research was supported by a donation from the Courtois foundation (to PB), grants from the Canadian Institutes of Health Research (Canadian Consortium on Neurodegeneration in Aging, to PB, Grant No. CIHR 400528 to SJ)), the Institute of Data Valorization (IVADO PRF3, to PB and SJ), a grant from Compute Canada (scq-952 to SJ and gsf-624 to PB), the Brain Canada Multi-Investigator Research Initiative (MIRI, to SJ), Canada First Research Excellence Fund (to SJ), and Healthy Brain, Healthy Lives (to SJ). PB is a fellow (Chercheur boursier Junior 2) of the Fonds de recherche du Québec-Santé. SJ is a recipient of a Canada Research Chair in neurodevelopmental disorders, and a chair from the Jeanne et Jean Louis Levesque Foundation.
CAM is supported by AIMS-2-TRIALS, which received support from the Innovative Medicines Initiative 2 Joint undertaking under grant agreement (Grant No. 777394).
The Cardiff Copy Number Variant cohort was supported by the Wellcome Trust Strategic Award DEFINE and the National Centre for Mental Health with funds from Health and Care Research Wales (code 100202/Z/12/Z).
Data from the UCLA cohort provided by CEB (participants with 22q11.2 deletions or duplications and control subjects) was supported through grants from the National Institutes of Health (NIH) (Grant No. U54EB020403), the National Institute of Mental Health (Grant Nos. R01MH100900, R01MH085953, U01MH119736, and R21MH116473), and the Simons Foundation (SFARI Explorer Award).
Finally, data from another study was obtained through the OpenFMRI project (http://openfmri.org) from the Consortium for Neuropsychiatric Phenomics (CNP), which was supported by NIH Roadmap for Medical Research grants (Grant Nos. UL1-DE019580, RL1MH083268, RL1MH083269, RL1DA024853, RL1MH083270, RL1LM009833, PL1MH083271, and PL1NS062410). Finally, this work was supported by the Simons Foundation (Grant Nos. SFARI219193 and SFARI274424).

9 - Declaration of Competing Interests

MJO, JH, and MBMvdB have a research grant from Takeda Pharmaceuticals outside the scope of this work. JH is a founding director of the company Meomics (unrelated to this work). PB is a consultant in fMRI processing for NeuroRX Inc., outside of the scope of this work. All other authors report no biomedical financial interests or potential conflicts of interest.

10 - Supplementary Materials

10.1 - Single Task Learning Benchmark - Conditions

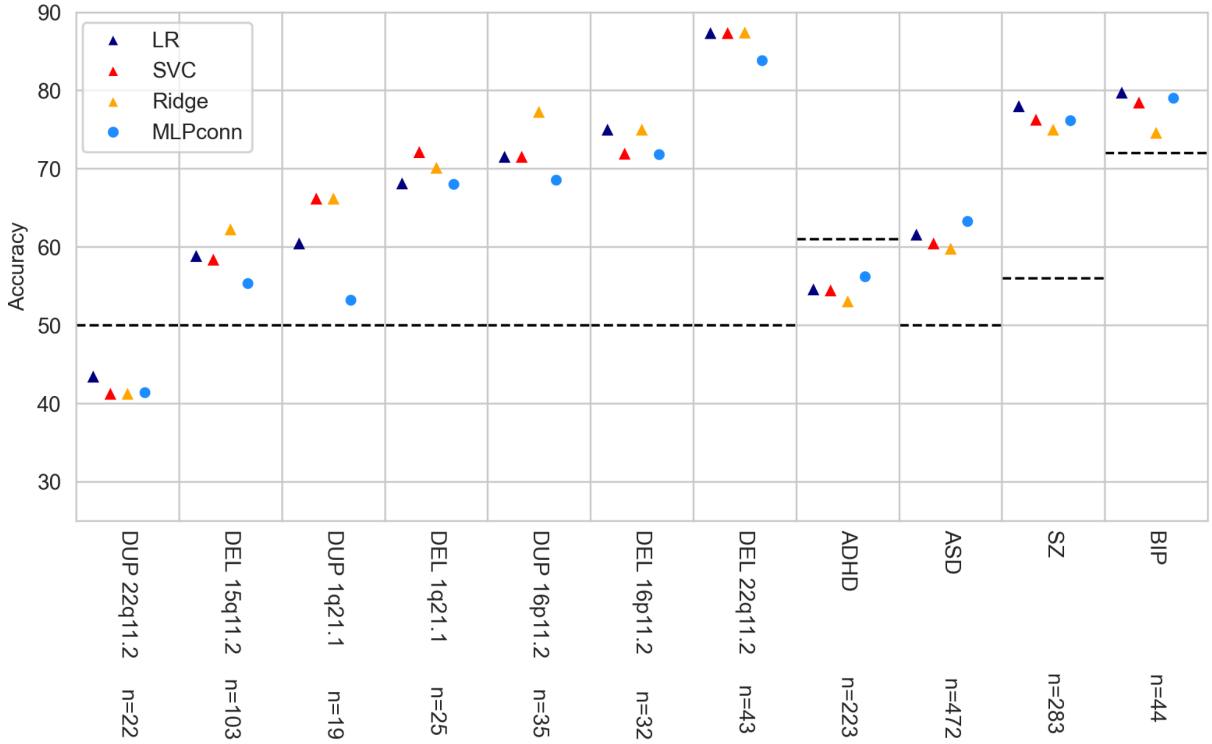

Figure 8 - Accuracy of automated diagnosis using single task learning. For each task, accuracy is shown for each of four models. LR: Logistic Regression, SVC: Support Vector Classifier, Ridge: Ridge Regression, and MLPconn. The x axis represents different conditions included as prediction tasks. The y axis shows the accuracy of prediction, chance level of prediction is indicated by a black dashed line.

10.2 - Study of Task Relationships using Variant Models

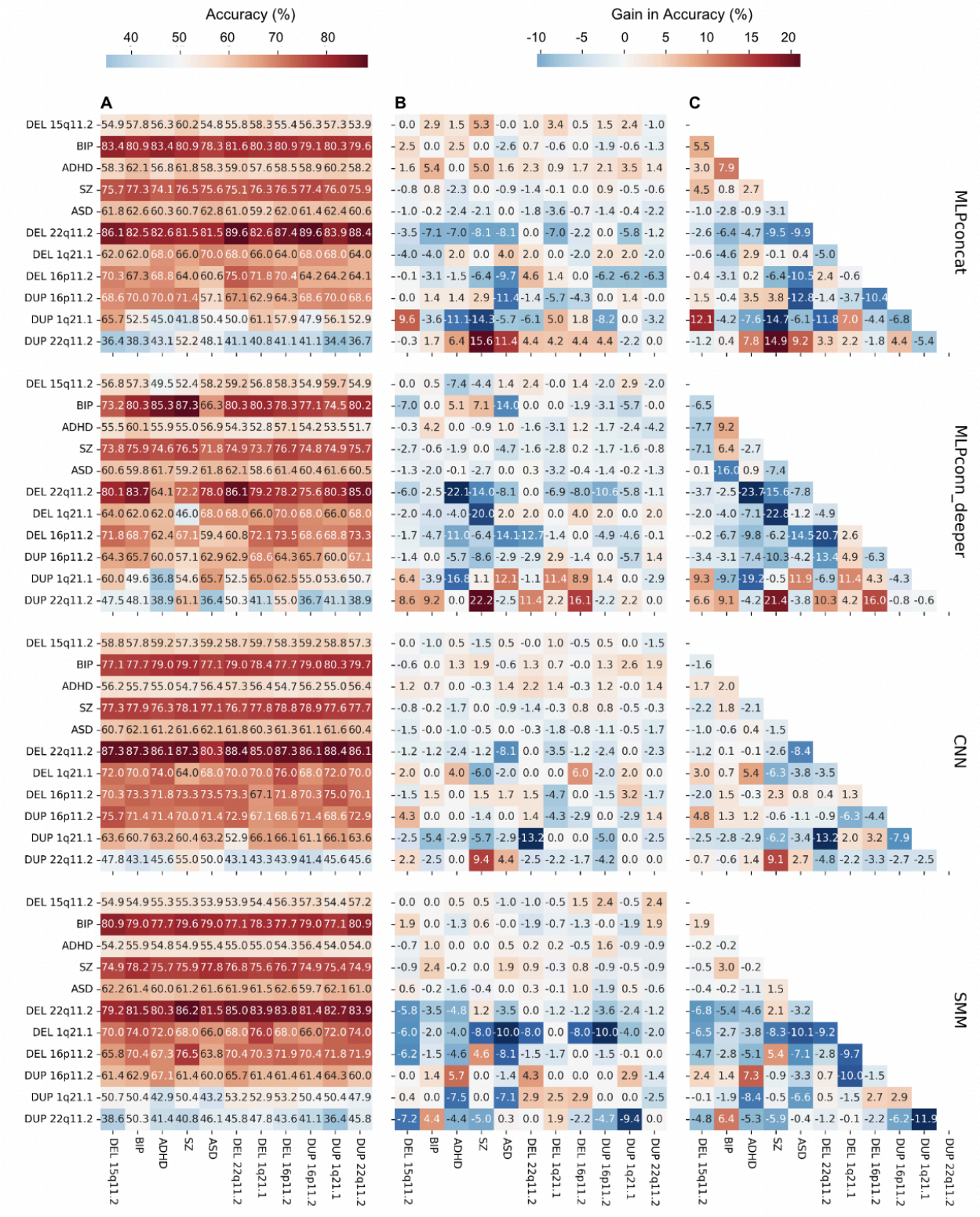

Figure 9 - In each matrix, the i,jth entry in the matrix is accuracy of condition in row i trained with condition in column j using the MLPconn model to perform automatic diagnosis. A: the matrix shows the raw accuracy achieved

for each pair, the second matrix B represents the difference in accuracy from the single-task baseline, and the third C shows the overall gain relative to baseline for a pair ($B + B^T$). Each row shows the results using the labelled model (MLPconcat, MLPconn_deeper, CNN, SMM), see the methods for details.

10.3 - Comparison with Huang and Colleagues

We implemented additional experiments to allow a closer comparison of our results to the only studies in the literature to apply MTL across conditions (Huang et al., 2022, 2020). In these studies, Huang and colleagues proposed the multicluster multigate mixture of experts model (M-MMOE). In the mixture of experts (MoE) model (Masoudnia & Ebrahimpour, 2014), expert submodels are shared across all tasks and combined by a single gate. In the multigate mixture of experts (MMOE) (J. Ma et al., 2018) rather than a single gate across experts, a gating network is added for each task (see Figure 10B). In the M-MMOE, brain ROIs are first clustered using a novel algorithm, and then each cluster receives an MMOE which are themselves combined as experts. We chose to use the MMOE model in our comparison since it is well established and allowed us to explore if the simple addition of multiple experts and gates could improve the results of MTL by allowing the model to learn task relationships, rather than the M-MMOE structure which is unique to the Huang and colleagues studies and introduces much more complexity.

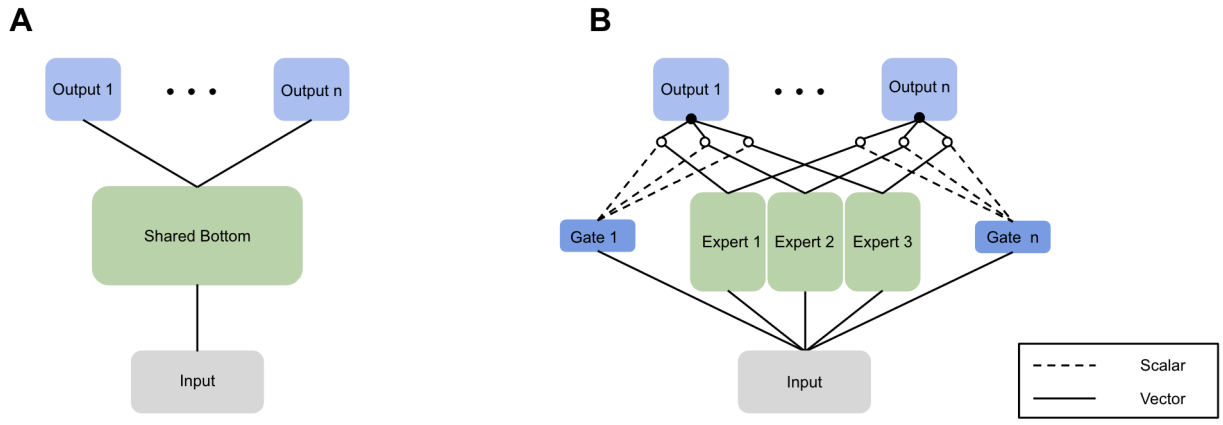

Figure 10 - A) Shared bottom model, B) MMOE.

We used the same encoder as in our MLPconn model as an expert, the same decoder as an output stack for each task, and followed the ratio of experts to tasks (2:1) used by Huang and colleagues. In detail, 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 each expert (two hidden layers with 256 and 64 units) as well as to the gating network (with N experts units), then the output from each expert is reweighted by the gate and summed, and passes finally to a task-specific output layer of 2 units for binary classification. Batch normalisation (Ioffe & Szegedy, 2015) is applied after each layer. Training was implemented as described in the methods. First, we applied the MMOE across the full sample of conditions (22 experts, 11 tasks),

next to eliminate the complication of the small CNV datasets we applied the MLPconn and MMOE (8 experts, 4 tasks) models across only the psychiatric conditions in the dataset. Finally, we implemented the task relationship experiment with the MMOE model.

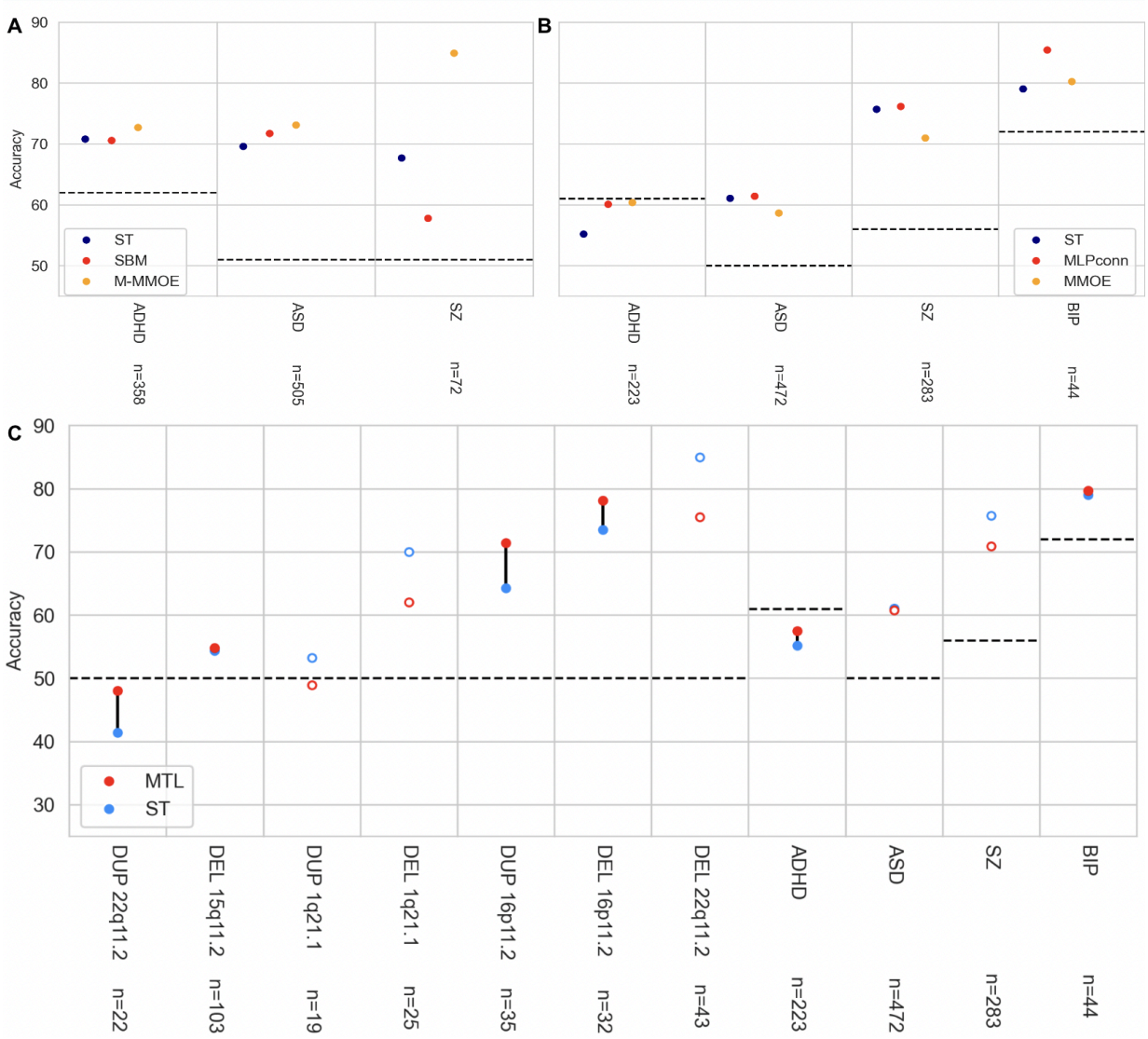

Figure 11 - A) Results reported in Huang et al. 2022 ST: Single Task, SBM: Shared Bottom Model, M-MMOE: variant of Multigate Mixture of Experts implemented by Huang and colleagues. Chance level indicated by a dashed line. B) Results of MMOE (8 experts) vs MLPconn across the psychiatric conditions C) Results of MMOE (22 experts) vs single task over the full set of conditions.

Using only the psychiatric conditions in the dataset, the MLPconn model improved prediction for all 4 conditions, whereas the MMOE helped accuracy for only 2 out of the 4 (see Figure 11B). This is contrary to the findings of Huang and colleagues, who reported marginal gains using their M-MMOE while the shared bottom model decreased accuracy (see Figure 11A). When we applied the MMOE across the full sample of conditions, we found that it performed slightly better than the MLPconn model on the full sample (accuracy improved for 6 out of 11 tasks vs. 4 out of 11) (see Figure 6). However, when examining task relationships using the MMOE (see

Figure 12) we saw a similar overall behaviour to the MLPconn model. Many pairs suffered a decrease in prediction accuracy when trained together, and the correlation with the results of the MLPconn model was high (r = 0.52).

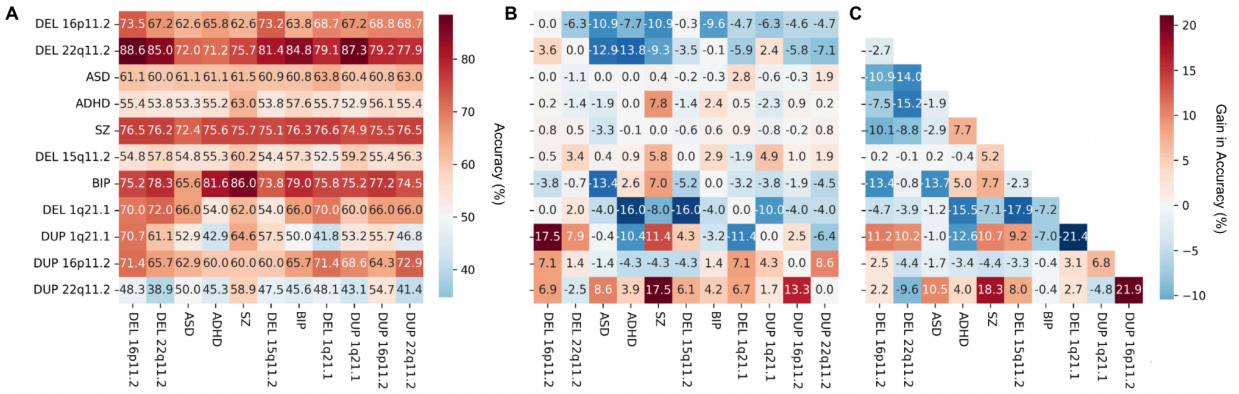

Figure 12 - In each matrix, the i,jth entry in the matrix is accuracy of condition in row i trained with condition in column j using the MLPconn model to perform automatic diagnosis. A: the matrix shows the raw accuracy achieved for each pair, the second matrix B represents the difference in accuracy from the single-task baseline, and the third C shows the overall gain relative to baseline for a pair ($B + B^T$).

10.4 - Distribution of Scores Across Folds of Cross-Validation

Here we present the distribution across the 5 folds of cross-validation of our results from sections 3.1.3, 3.1.4 and 3.2, rather than the average score, in order to give a better sense of the spread of MTL vs ST scores. We observed substantial variations in accuracy across folds, which was expected given the small sample size in each fold (less than 20 individuals, and as low as 6).

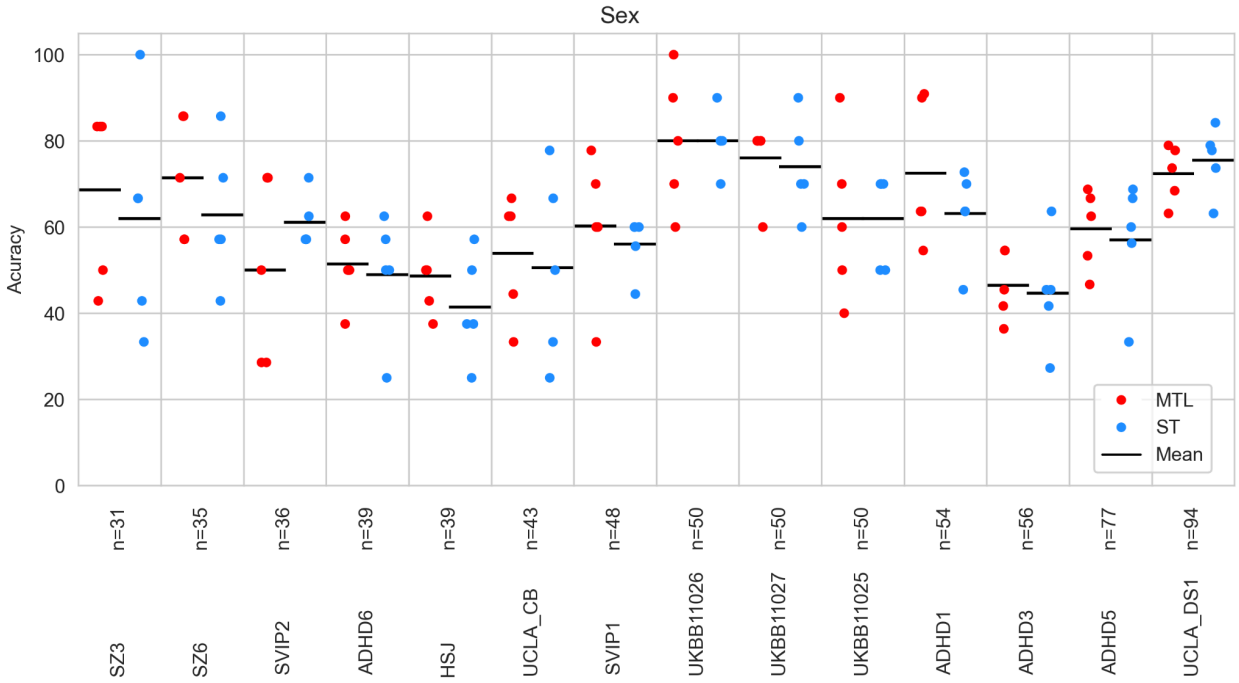

Figure 13 - Distribution of accuracy of sex prediction across 5-folds of k-fold cross-validation using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents different data collection sites included as prediction tasks. Sites are ranked by sample size, with the largest to the right. The y axis shows the accuracy of prediction. For each task, the red points show prediction using the MLPconn architecture in MTL, and the blue points show prediction using the MLPconn architecture in ST.

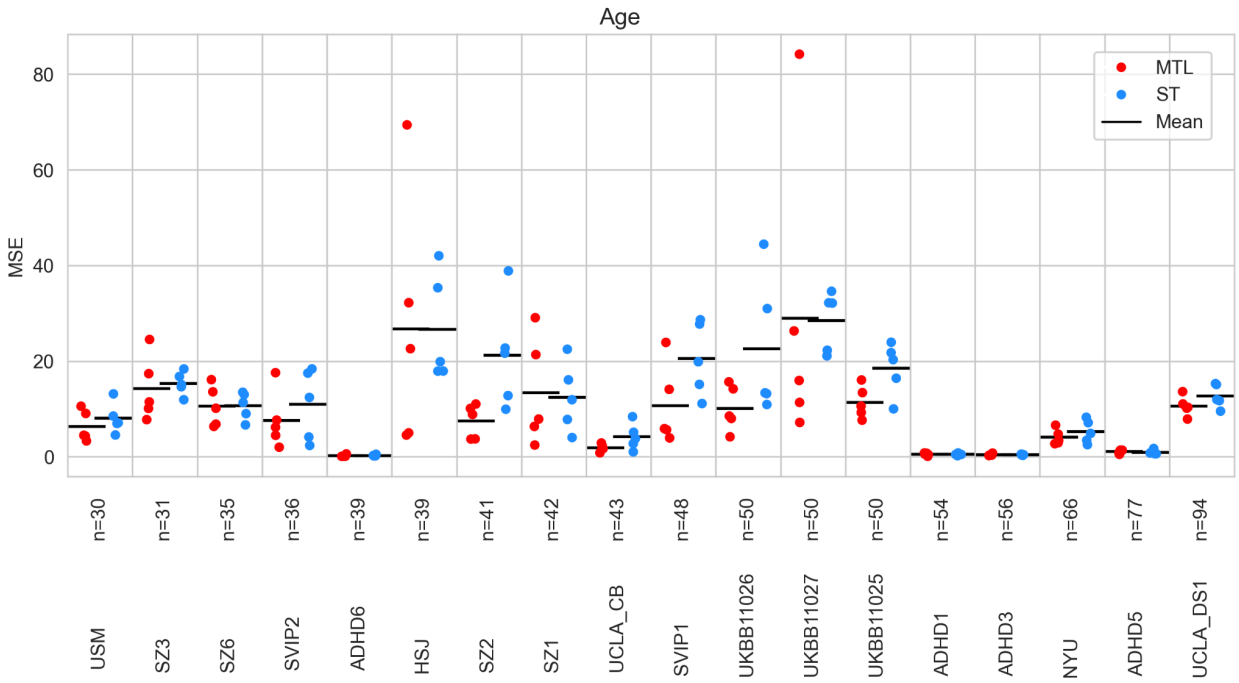

Figure 14 - Distribution of Mean Squared Error (MSE) of age prediction across 5-folds of k-fold cross-validation using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents different data collection sites included as prediction tasks. Sites are ranked by sample size, with the largest to the right. The y axis shows the prediction error. For each task, the red points show prediction using the MLPconn_reg architecture in MTL, and the blue points show prediction using the MLPconn_reg architecture in ST.

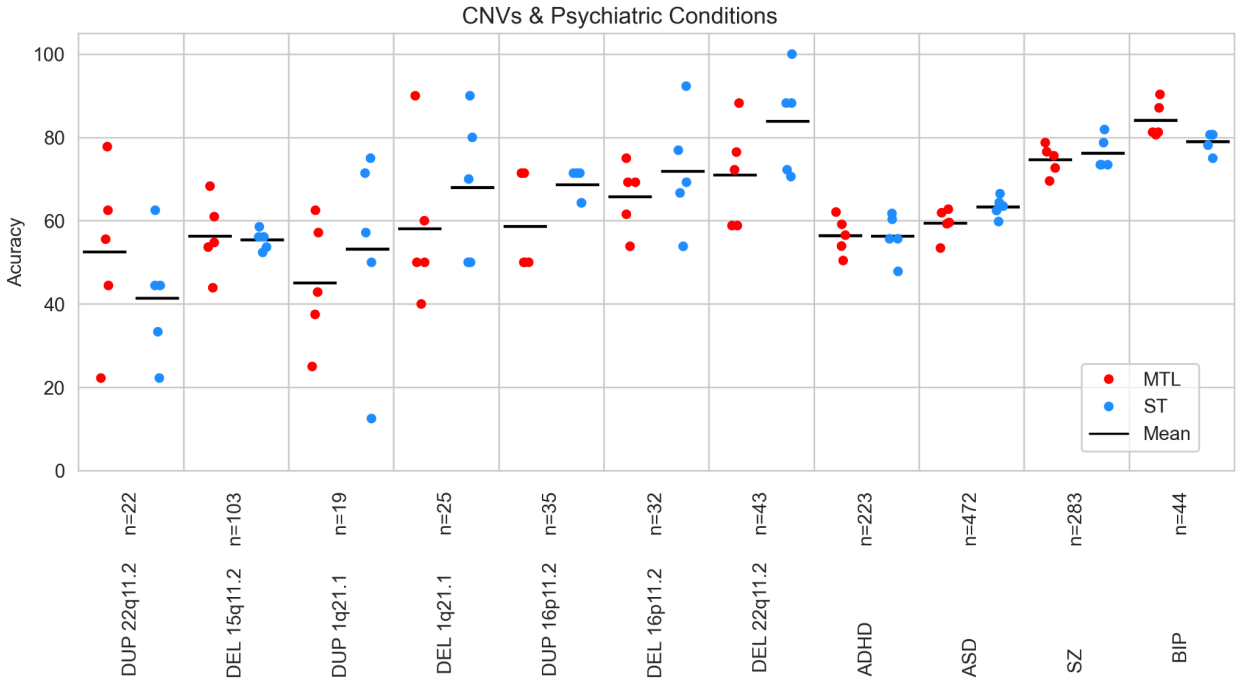

Figure 15 - Distribution of accuracy of automated diagnosis across 5-folds of k-fold cross-validation using single (ST) vs multi-task learning (MTL). The x axis represents different conditions included as prediction tasks. The y axis

shows the accuracy of prediction. For each task, the red points show prediction using the MLPconn architecture in MTL and the blue points show prediction using the MLPconn architecture in ST.

10.5 - AUC & F1 Score

Here we present the results of our classification prediction experiments, scored in sections 3.1.3 and 3.2 using prediction accuracy, using the Area Under the Receiver Operating Characteristic (AUC) (Nahm, 2022) and F1-scores (Taha & Hanbury, 2015) to provide a more comprehensive view of the model's performance. This is particularly relevant for the sex prediction study (section 3.1.3) in which the datasets have class imbalances (see Table 2).

These scores are derived from precision, recall (also called sensitivity), and specificity. Precision is defined as the number of true positives (subjects predicted as class 1 that are class 1) divided by the number of true positives plus false positives (subjects predicted as class 1 that are class 0). Recall is the number of true positives divided by the number of true positives plus false negatives (subjects predicted as class 0 that are class 1). Specificity is defined as the number of true negatives (subjects predicted as class 0 that are class 0) divided by the number of true negatives plus false positives. The F1-score is defined as the harmonic mean of precision and recall. In general classifiers output a continuous value which is turned into a binary prediction by comparing it to a threshold. The Receiver Operating Characteristic (ROC) curve plots recall vs. 1 - specificity at different thresholds. The AUC measures the overall performance of classification models using the area under the ROC curve. An AUC of 1 implies a perfect classifier and an AUC of 0.5 implies a random classifier. The qualitative conclusions of our experiments matched between AUC and accuracy scores, while F1 scores were more difficult to interpret.

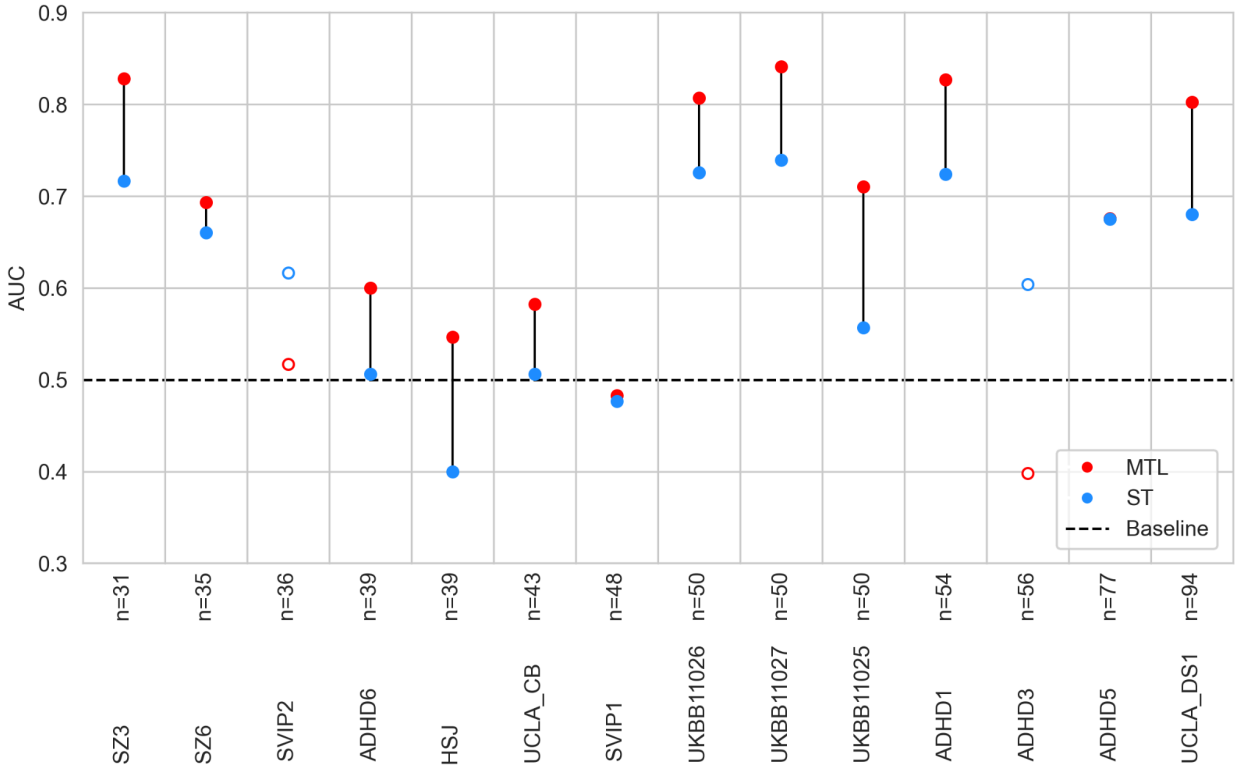

Figure 16 - Area Under the Receiver Operating Characteristic curve (AUC) of sex prediction using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents different data collection sites included as prediction tasks. Sites are ranked by sample size, with the largest to the right. The y axis shows the AUC, chance level of prediction is indicated by a black dashed line. For each task, the red point shows prediction using the MLPconn architecture in MTL, and the blue point shows prediction on the task trained independently using the MLPconn architecture. Where the red point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL prediction outperformed the ST, points were filled and connected by a line, and otherwise they were not.

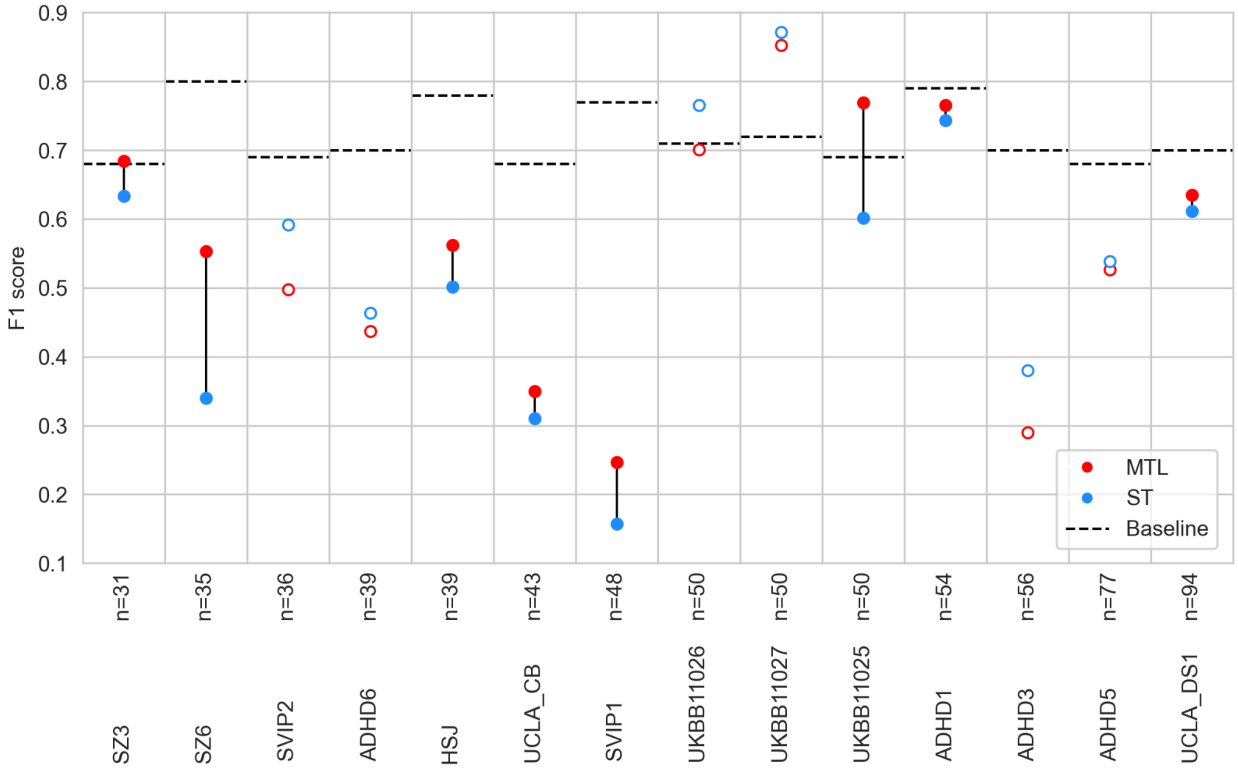

Figure 17 - F1 score of sex prediction using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents different data collection sites included as prediction tasks. Sites are ranked by sample size, with the largest to the right. The y axis shows the F1 score of prediction, chance level of prediction is indicated by a black dashed line. For each task, the red point shows prediction using the MLPconn architecture in MTL, and the blue point shows prediction on the task trained independently using the MLPconn architecture. Where the red point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL

prediction outperformed the ST, points were filled and connected by a line, and otherwise they were not.

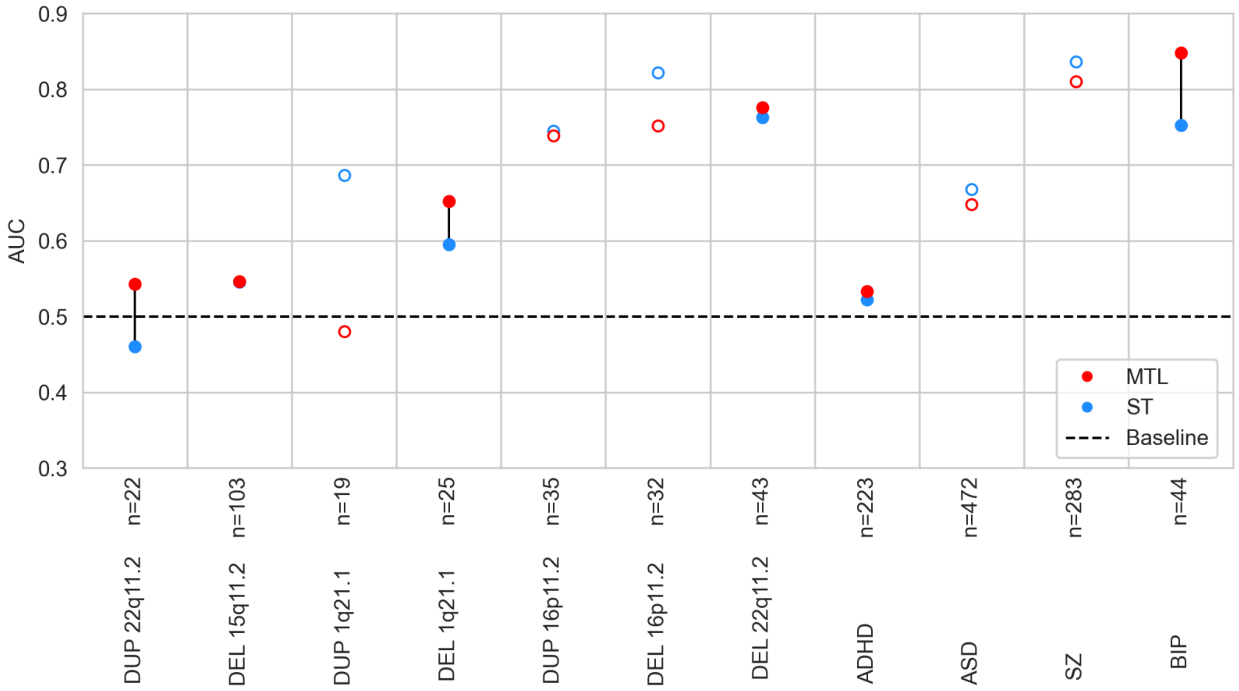

Figure 18 - Area Under the Receiver Operating Characteristic curve (AUC) of automated diagnosis using single (ST) vs multi-task learning (MTL). For each task, the red point shows prediction using the MLPconn architecture in MTL and the blue point shows prediction on the task trained independently using the MLPconn architecture. Where either the blue point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL prediction outperformed the ST, points are filled and connected by a line, and otherwise they are not. The x axis represents different conditions included as prediction tasks. The y axis shows the AUC of prediction, chance level of prediction is indicated by a black dashed line.

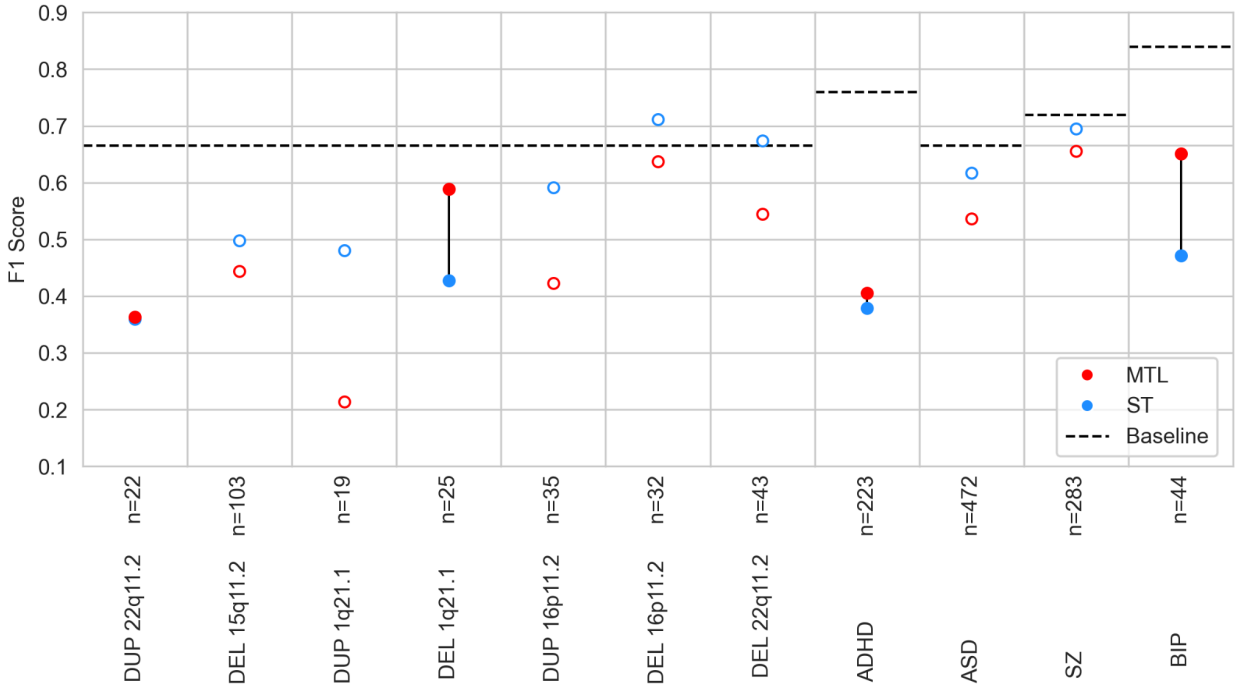

Figure 19 - F1 score of automated diagnosis using single (ST) vs multi-task learning (MTL). For each task, the red point shows prediction using the MLPconn architecture in MTL and the blue point shows prediction on the task trained independently using the MLPconn architecture. Where either the blue point appears missing, the accuracy values for the two models are so close that the points are overlapped. If the MTL prediction outperformed the ST, points are filled and connected by a line, and otherwise they are not. The x axis represents different conditions included as prediction tasks. The y axis shows the F1 score of prediction, chance level of prediction is indicated by a black dashed line.

10.6 - Effect Sizes as a Measure of Task Difficulty

Here we aimed to measure how difficult each prediction task is in order to contextualise the performance of MTL in different settings: predicting age or sex across sites of data collection, and performing automatic diagnosis across CNVs and psychiatric 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. We would like to emphasise that there is no theoretical reason for CWAS effect sizes to match accuracy with ML tools, as ML tools are a multivariate measure of effect size (akin to a statistical omnibus tests) rather than mass univariate like CWAS. In practice these two types of effect sizes do not necessarily align (Bzdok & Ioannidis, 2019; Lo et al., 2015; Shmueli, 2010). However, CWAS effect sizes are a common metric and provide intuitive guidance for interpretations.

Specifically, we implemented 13 CWAS using sex as a contrast for each site included in the sex prediction study, 18 CWAS using age groups as a contrast (younger half of subjects vs. older half) for each site included in age prediction study, and 11 CWAS for the following conditions: 7 CNVs and 4 psychiatric conditions. For the CWAS on conditions, 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. In order to have the best possible statistical power, we pooled all the control subjects we had access to (n = 31425, 16590 female subjects, age mean 62.31 and standard deviation 11.47, framewise displacement mean 0.18 and standard deviation 0.05, from a total of 53 sites of data collection). The results on CNVs and psychiatric conditions presented here were published in two studies: (Moreau et al., 2023) and (Moreau et al., 2022).

For each CWAS, 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. For the CWAS on sex at each site, models were head motion, age and global signal. For the CWAS on age at each site, models were head motion, sex and global signal. For the CWAS on conditions, 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 (J. 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 & Hochberg, 1995) at a threshold of $q < 0.05$. 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 (Moreau et al., 2023).

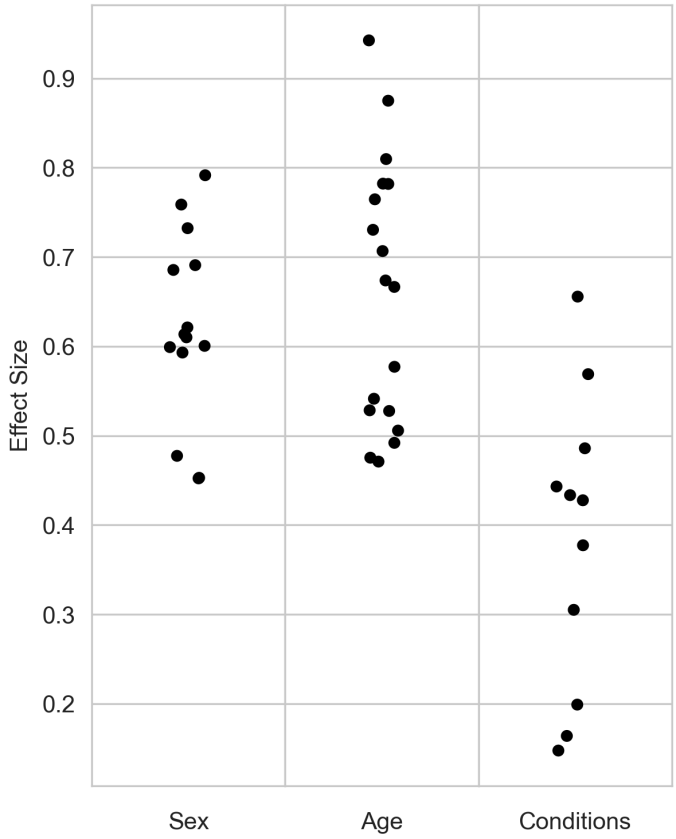

Figure 20 - Effect sizes on connectivity, defined as the mean of the top decile of the FC profile, for sex and age at each site of data collection and conditions (7 CNVs and 4 psychiatric conditions).

Effect sizes on connectivity for sex and age (at 14 and 18 scanning sites respectively) are much higher than for the 7 CNVs and 4 psychiatric conditions included in our dataset. This matches the observed behaviour of ML techniques in the case of predicting sex vs predicting conditions, as the accuracy of sex classification is higher in the full UK biobank sample (see section 3.1.1) than any of the automated diagnostic classifiers (see section 10.1). Predicting age is more difficult to directly compare with automatic diagnosis since it is a regression rather than a classification task. Overall, the higher effect sizes for sex and age relative to conditions makes them easier as prediction tasks in the ST setting and therefore more likely to benefit multi-task learning. Additionally, predicting sex or age across different scanning sites is intuitively better suited for MTL since the tasks have a common target and therefore clearly have shared information that can be exploited by a combined model. In the case of automatic diagnosis, while the conditions are related and have substantial shared information, they are all distinct conditions and therefore less easily combined by a single MTL model.

10.7 - Confound distributions by Site & Condition

Here we present plots of the distribution of the confounding variables for each dataset, first the single scanning site datasets (control subjects only) used for the age & sex prediction tasks (sections 3.1.3 and 3.1.4) followed by the multi-site datasets used to predict conditions in section 3.2 (matched number of cases & controls). These demonstrate the large dataset variability regarding confound distribution, which provides important context on the results from the previous studies, and is also important to interpret the ablation study (section 10.8) in which MTL prediction is repeated using all but one dataset in order to analyse the effect of each dataset on performance.

10.7.1 - Single Site Datasets

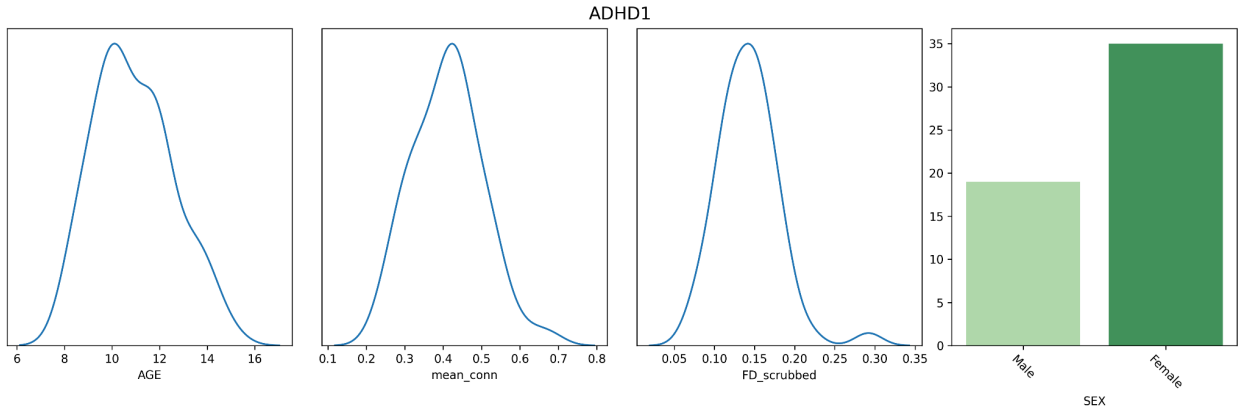

Figure 21 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the ADHD1 scanning site.

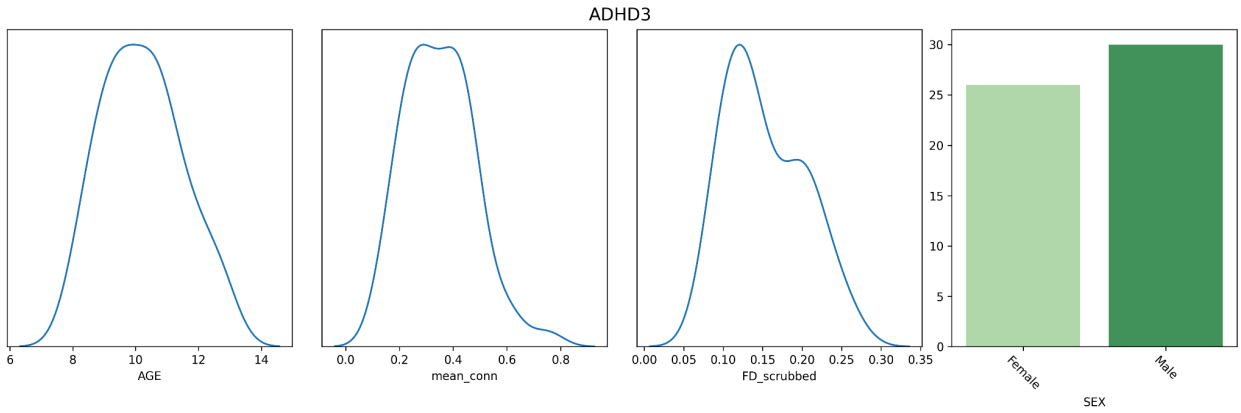

Figure 22 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the ADHD3 scanning site.

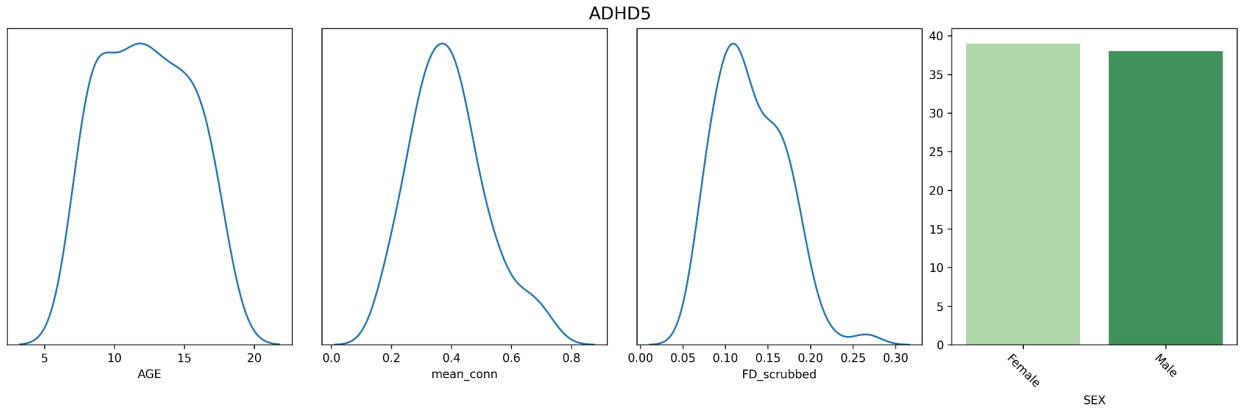

Figure 23 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the ADHD5 scanning site.

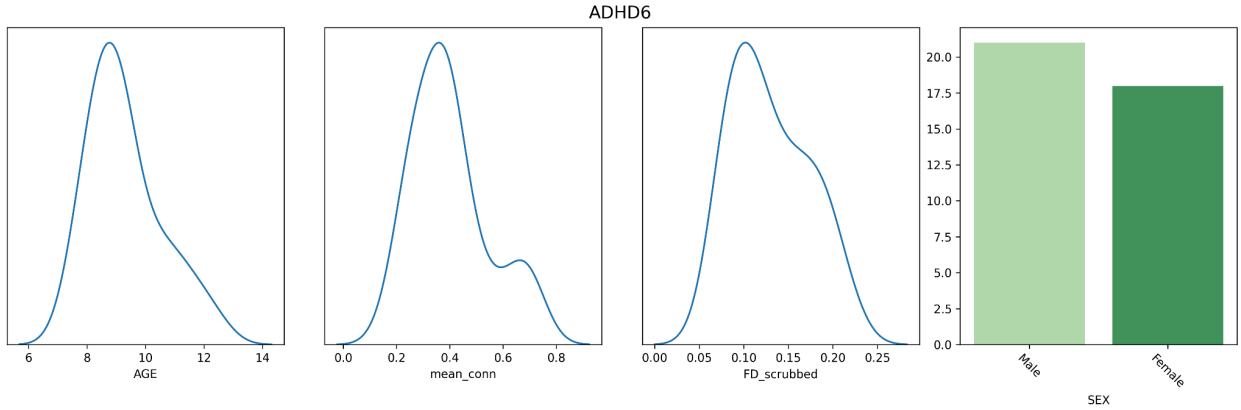

Figure 24 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the ADHD6 scanning site.

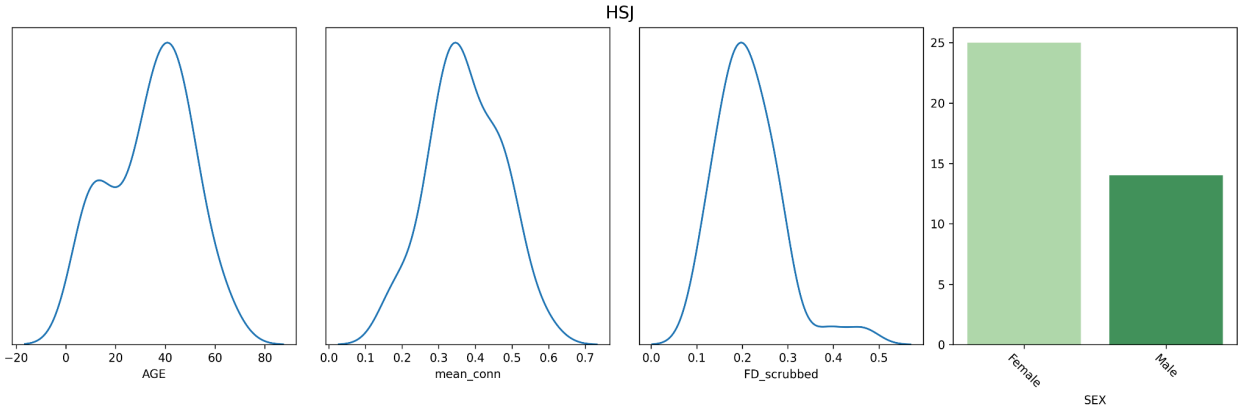

Figure 25 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the HSJ scanning site.

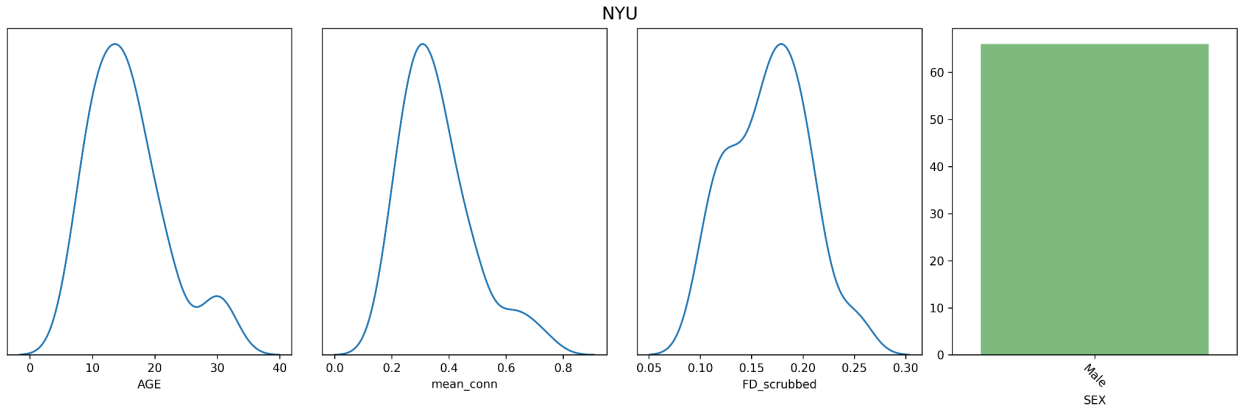

Figure 26 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the NYU scanning site.

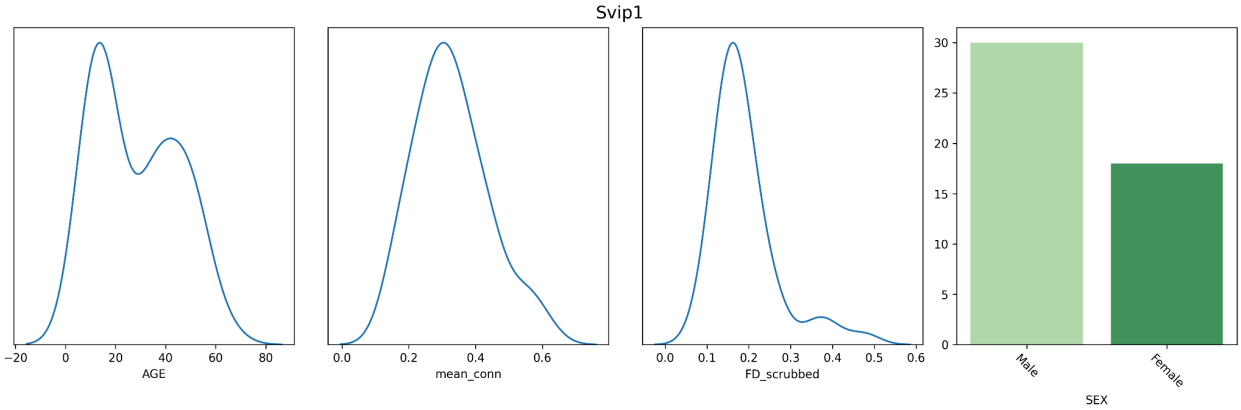

Figure 27 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the Svip1 scanning site.

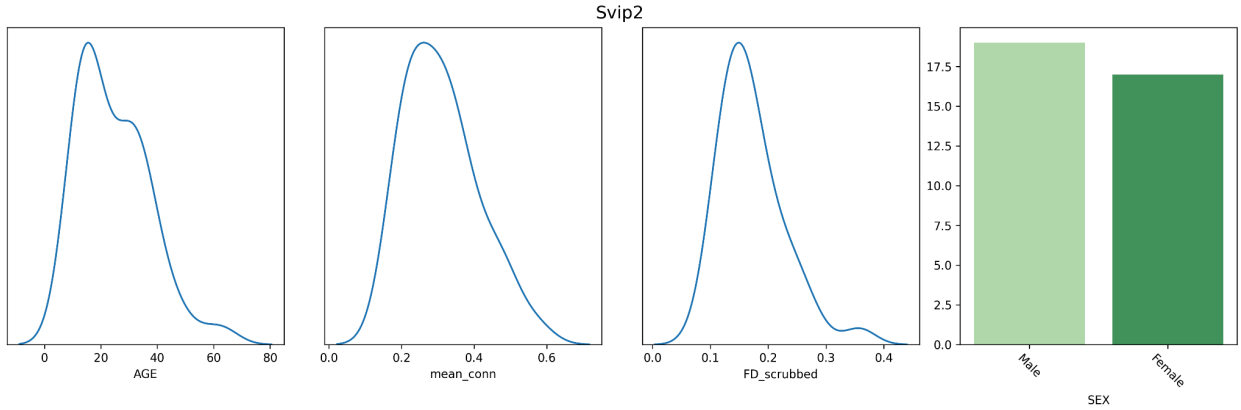

Figure 28 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the Svip2 scanning site.

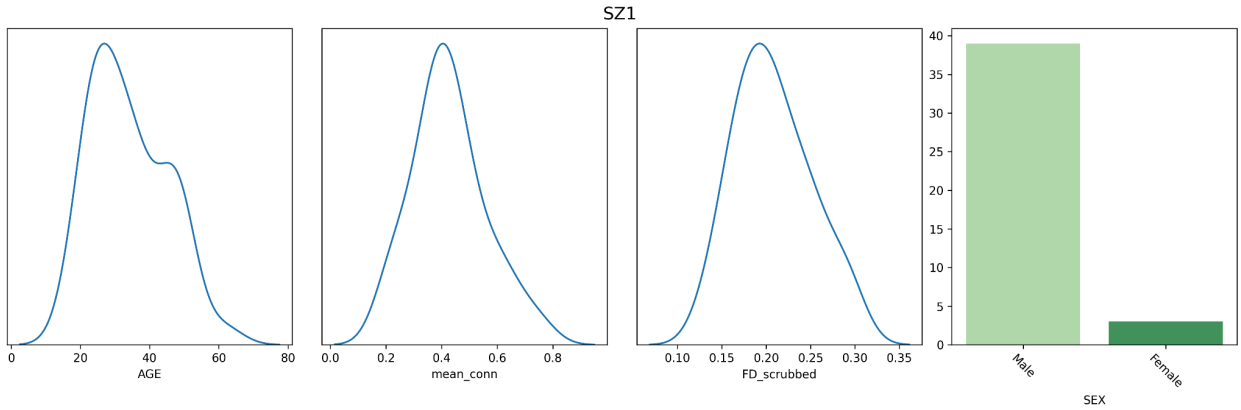

Figure 29 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the SZ1 scanning site.

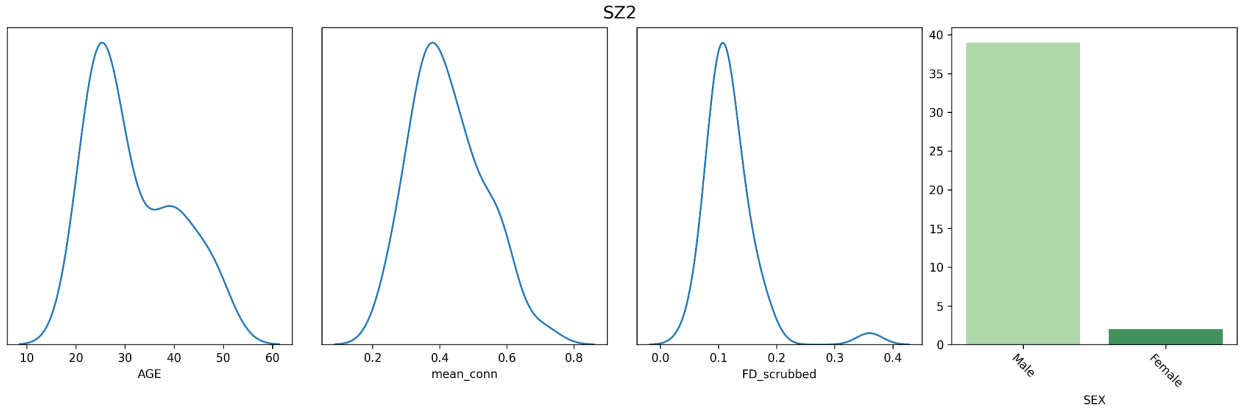

Figure 30 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the SZ2 scanning site.

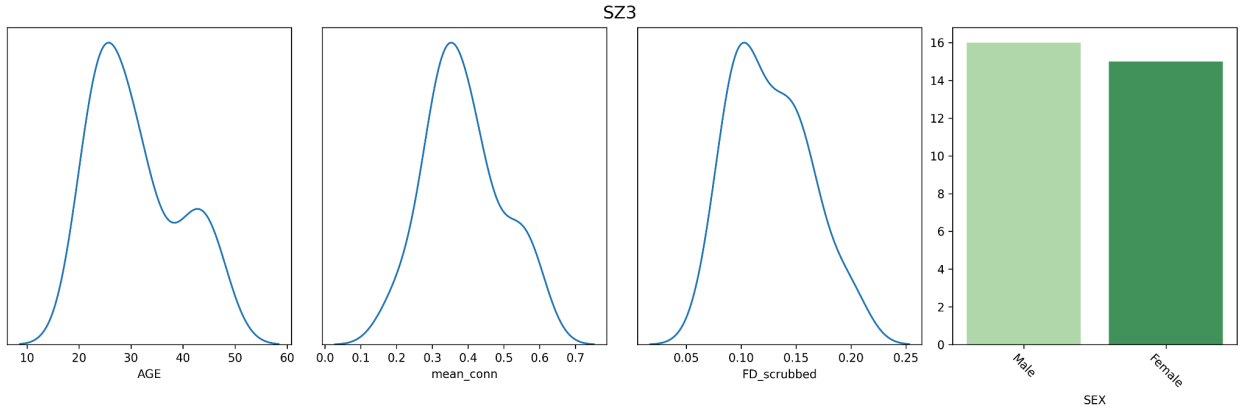

Figure 31 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the SZ3 scanning site.

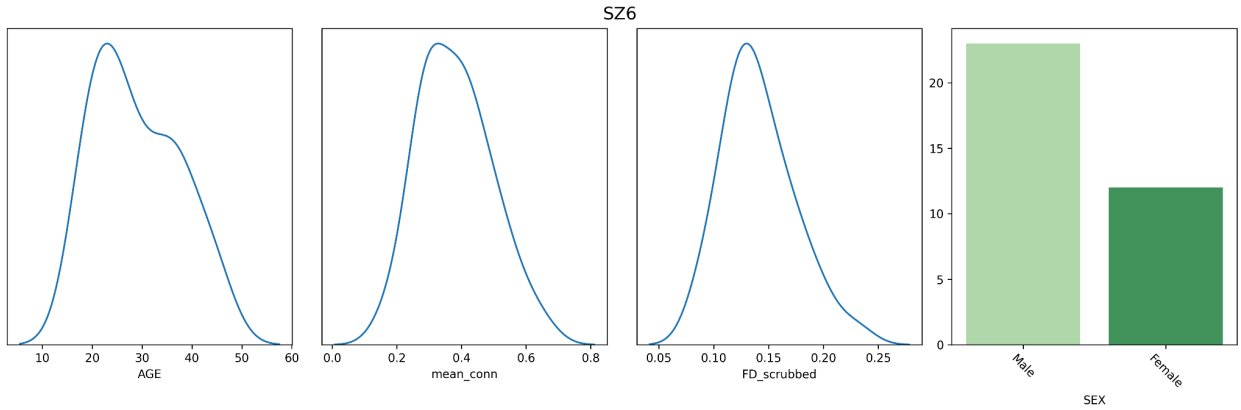

Figure 32 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the SZ6 scanning site.

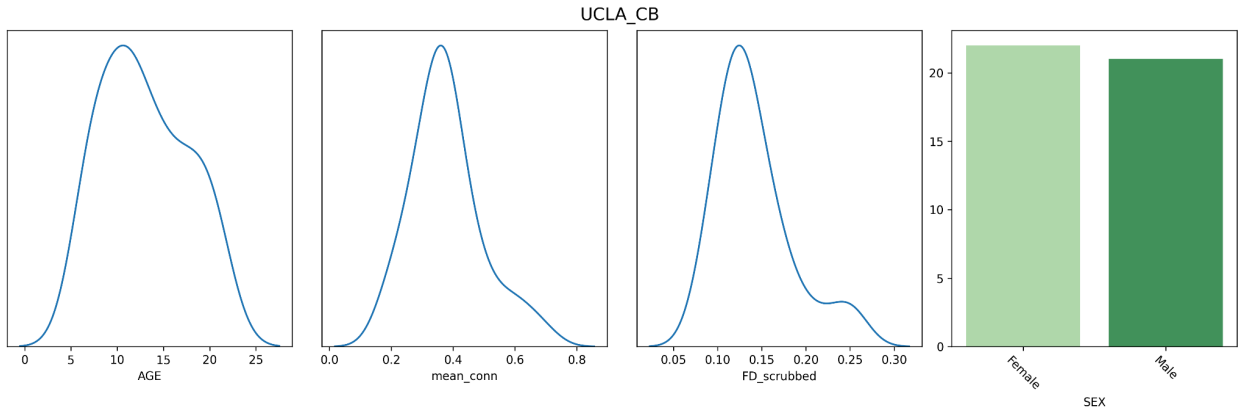

Figure 33 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the UCLA_CB scanning site.

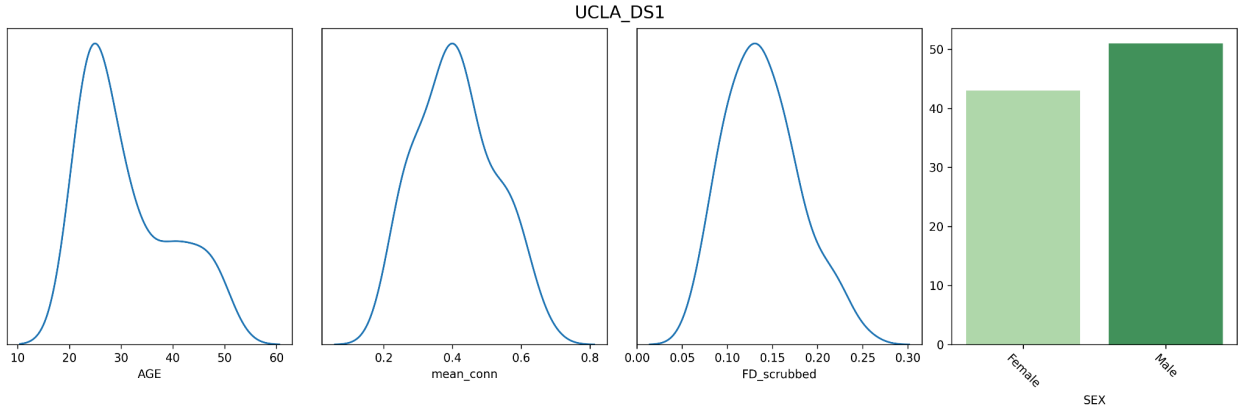

Figure 34 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the UCLA_DS1 scanning site.

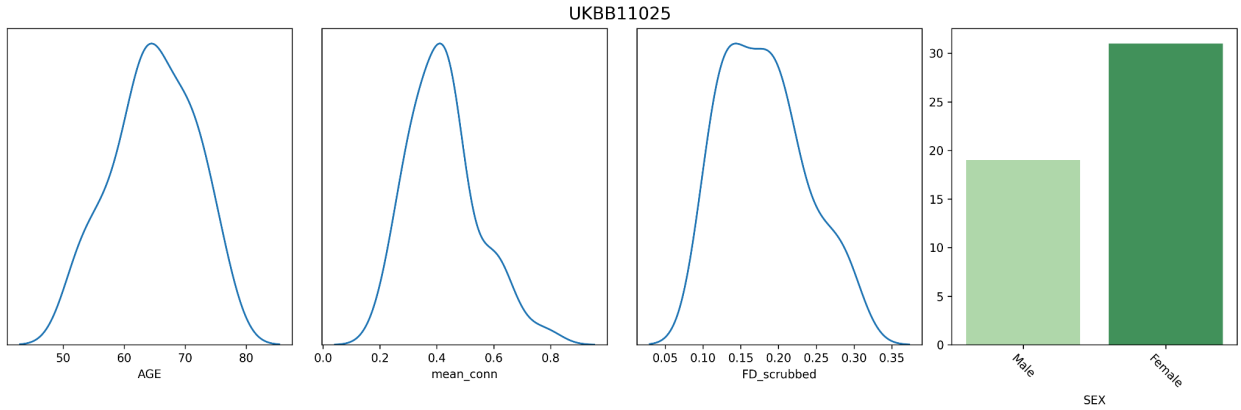

Figure 35 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among 50 subsampled control subjects at the UKBB11025 scanning site.

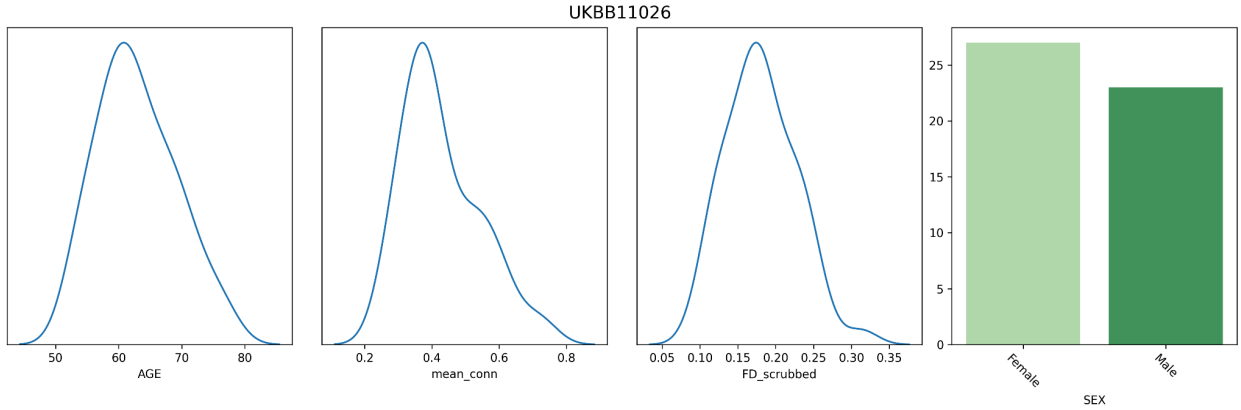

Figure 36 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among 50 subsampled control subjects at the UKBB11026 scanning site.

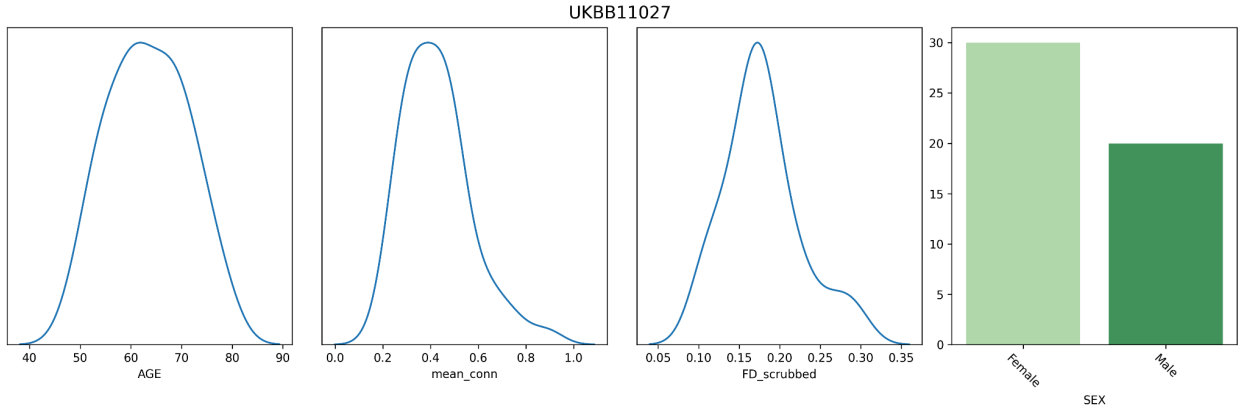

Figure 37 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among 50 subsampled control subjects at the UKBB11027 scanning site.

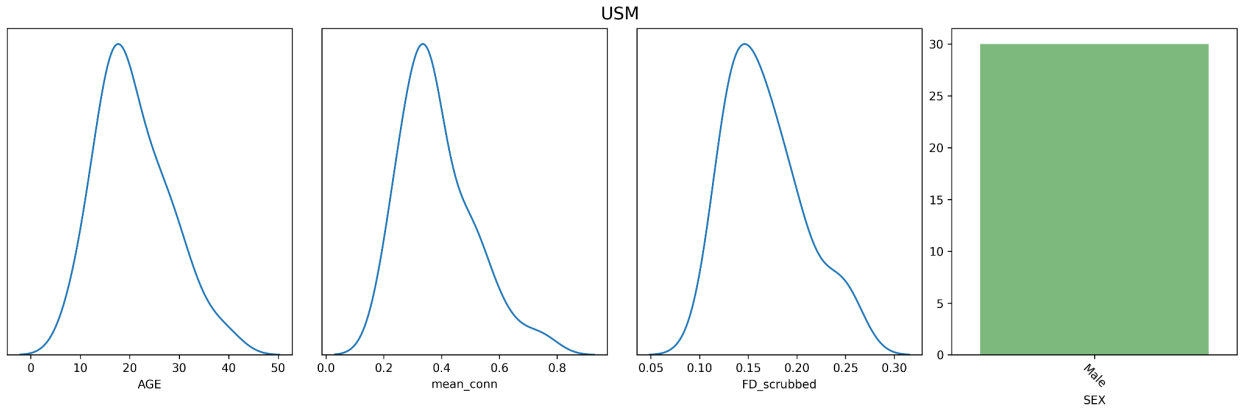

Figure 38 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) among control subjects at the USM scanning site.

10.7.2 - CNV & Psychiatric Condition Datasets

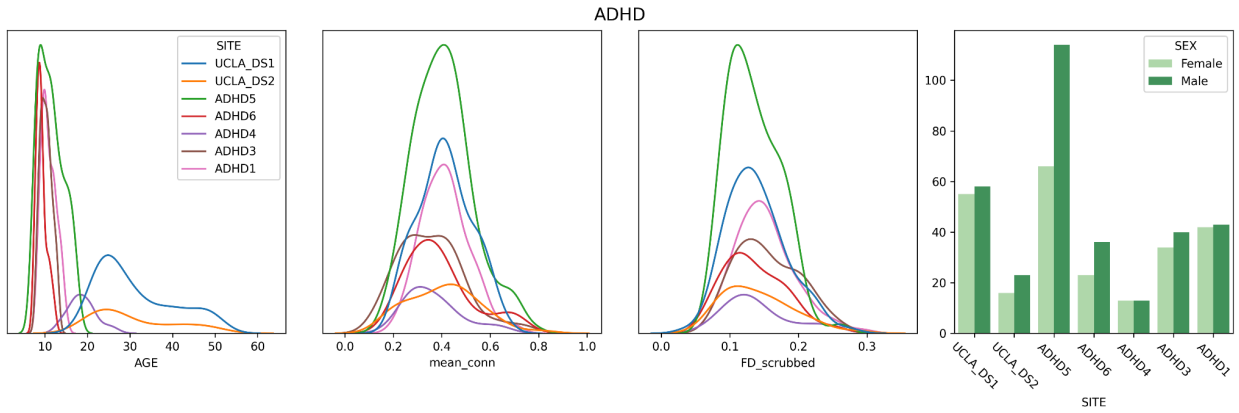

Figure 39 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the ADHD dataset.

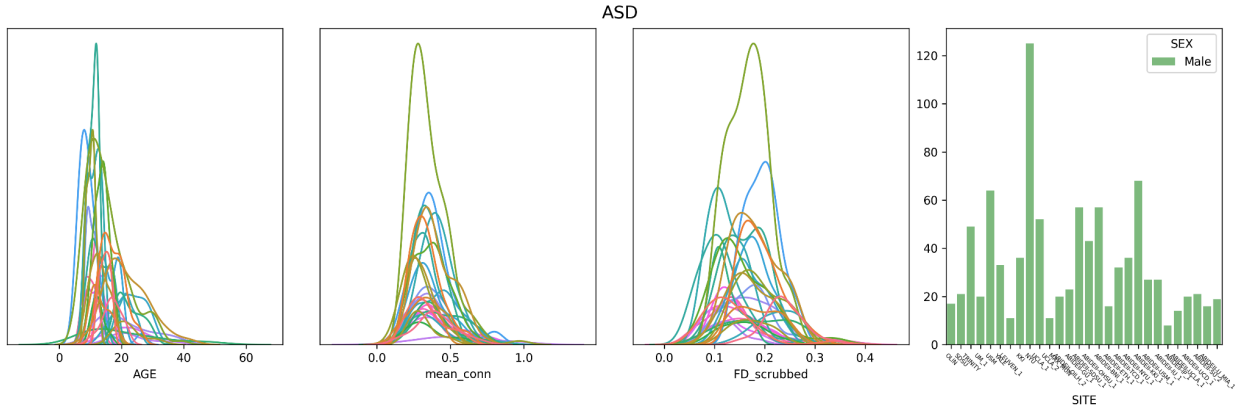

Figure 40 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the ASD dataset.

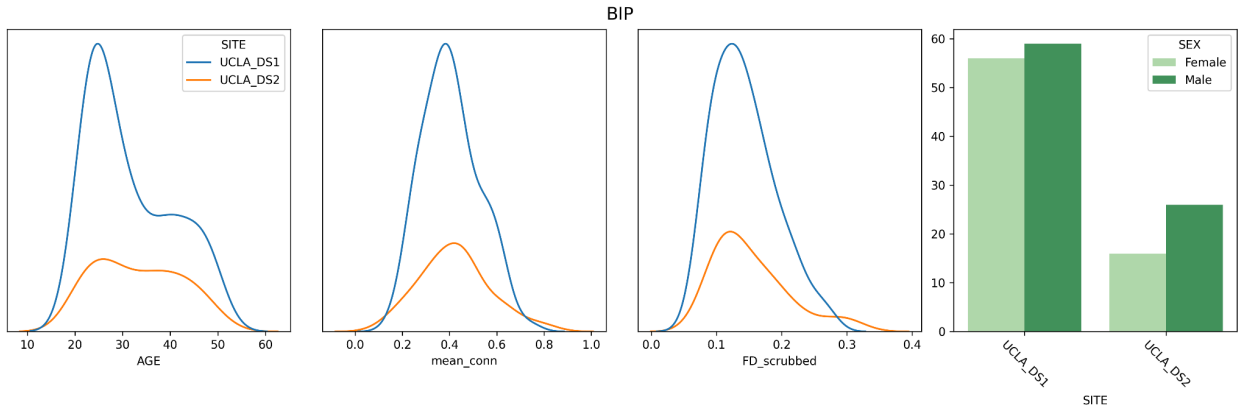

Figure 41 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the BIP dataset.

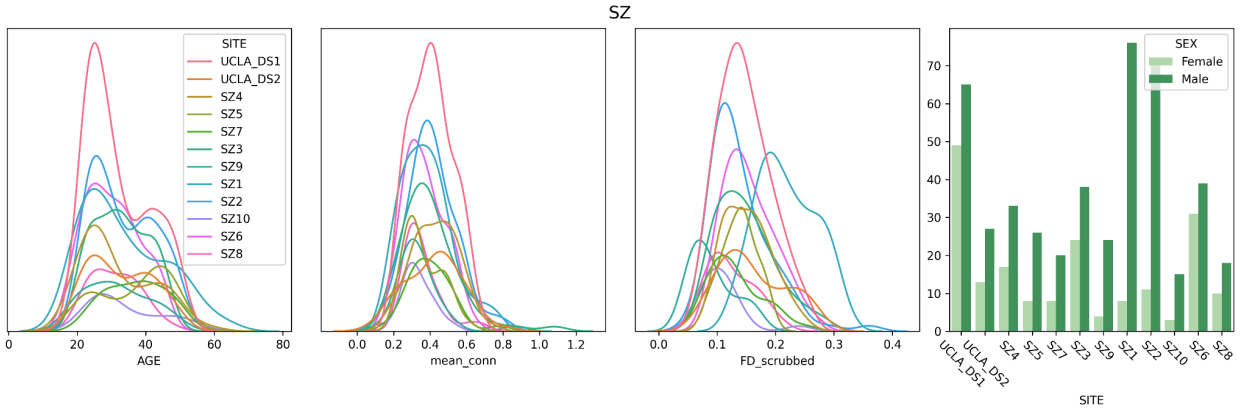

Figure 42 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the SZ dataset.

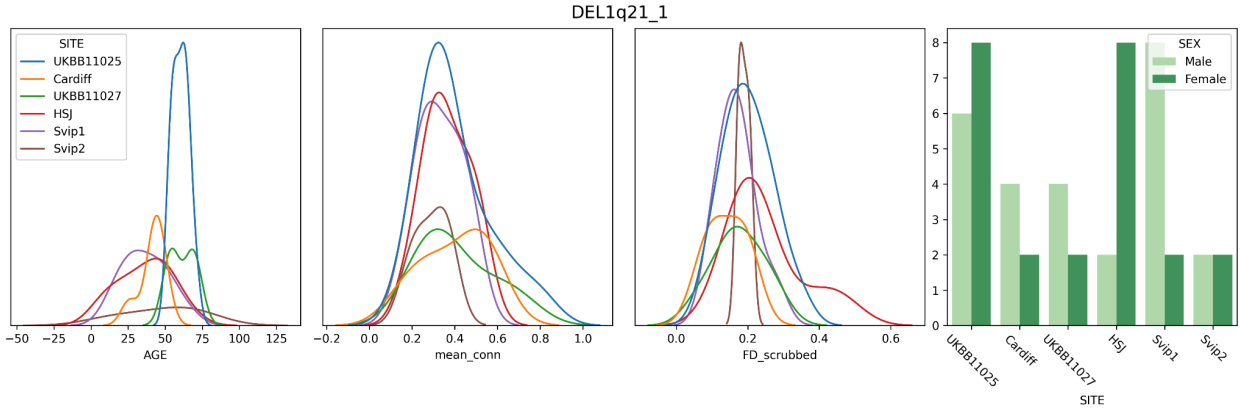

Figure 43 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DEL 1q21.1 dataset.

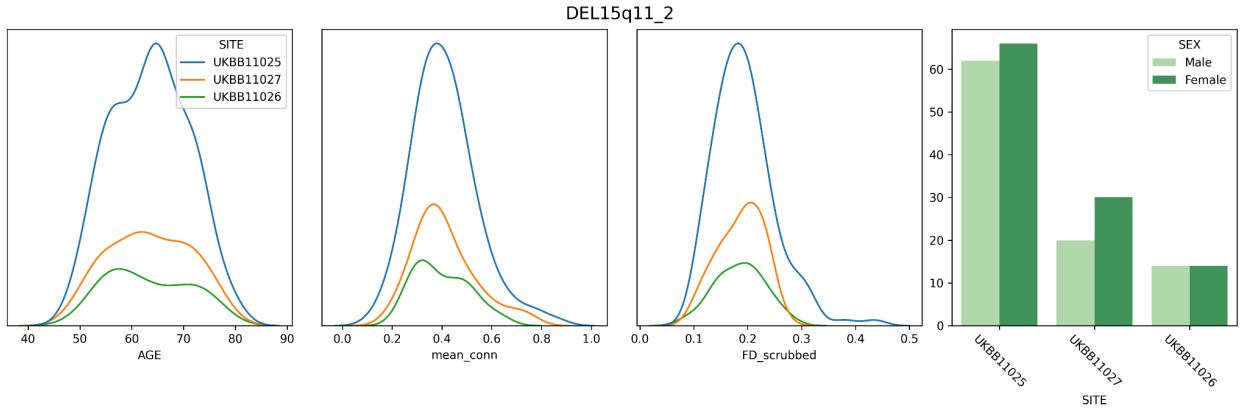

Figure 44 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DEL 15q11.2 dataset.

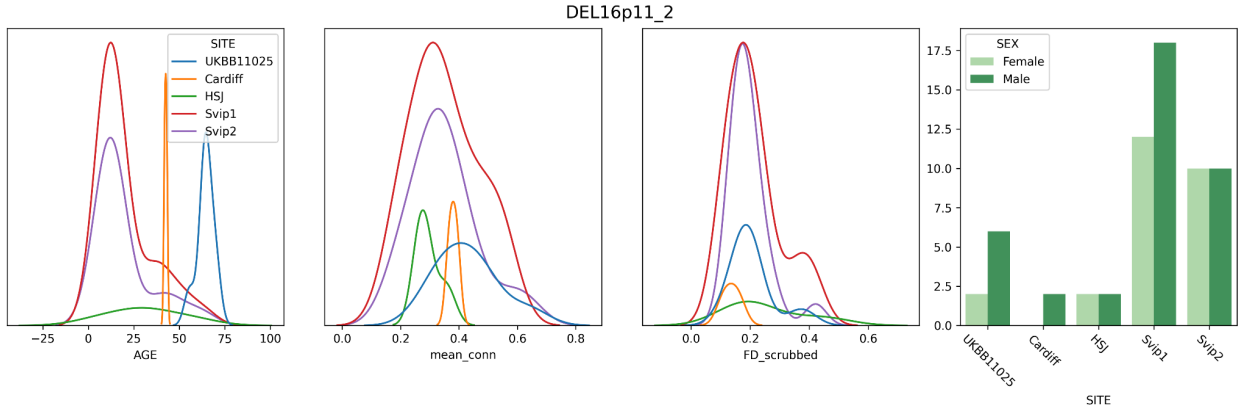

Figure 45 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DEL 16p11.2 dataset.

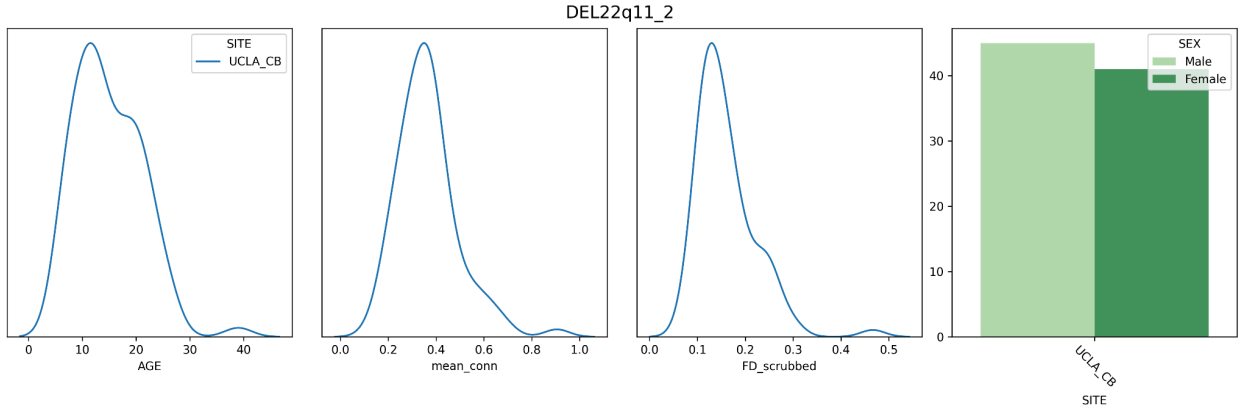

Figure 46 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DEL 22q11.2 dataset.

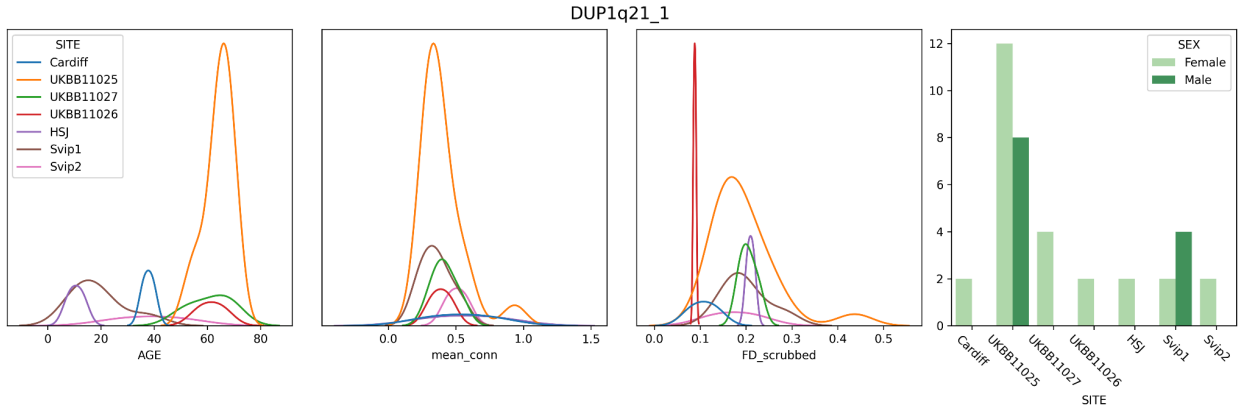

Figure 47 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DUP 1q21.1 dataset.

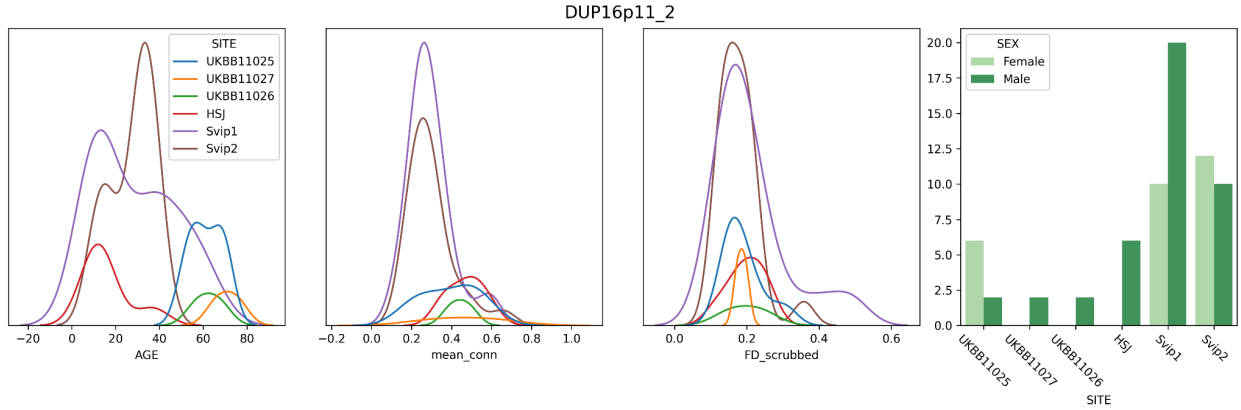

Figure 48 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DUP 16p11.2 dataset.

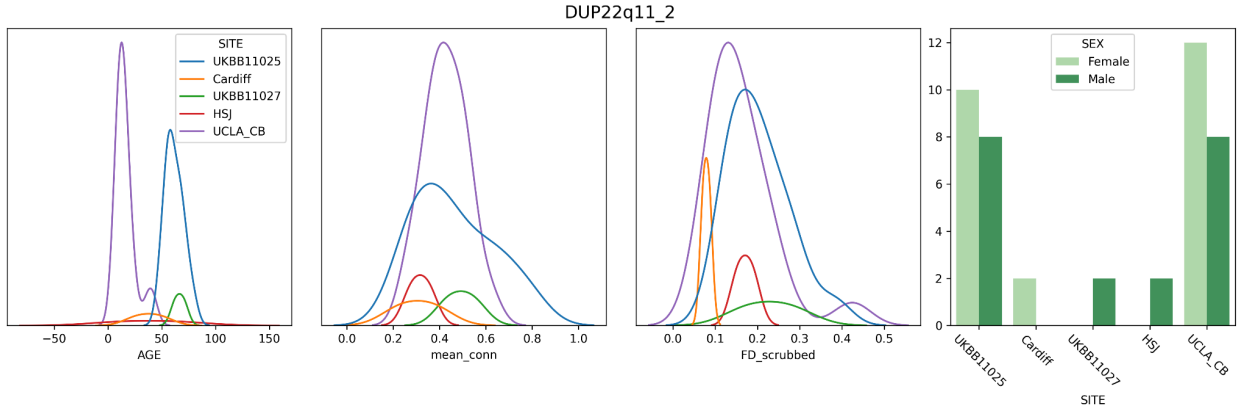

Figure 49 - Distribution of confounding variables (age, global signal (mean_conn), head motion (FD_scrubbed), and sex) by scanning site for the DUP 22q11.2 dataset.

10.8 - Ablation Study

In order to evaluate the impact of each dataset on the prediction performance, we performed an ablation study in which we iteratively dropped a single dataset from the set of tasks and repeated the MTL prediction experiments from sections 3.1.3, 3.1.4 and 3.2. Specifically, we conducted 14 experiments for sex prediction and 18 for age prediction (one for each site of data collection dropped). For automatic diagnosis we conducted 11 experiments (one for each condition dataset dropped). Training was performed as described in the methods (section 2.6). This analysis did not identify any dramatic effect of a single site, however excluding some sites did improve on the MTL accuracy, compared to ST (for example Svip2 for sex prediction), although we did not test the statistical significance of such improvements which would need to be adjusted for the very large number of ablation experiments performed here.

Figure 50 - Distribution of accuracy of sex prediction across tasks using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents the data collection site which is dropped from the set of prediction tasks, except for the first column which shows results using the full dataset. The y axis shows the accuracy of prediction. For each removed dataset, the red points show prediction accuracy distribution for the tasks using the MLPconn architecture in MTL, and the blue points show prediction accuracy distribution on the tasks trained using the MLPconn architecture in ST.

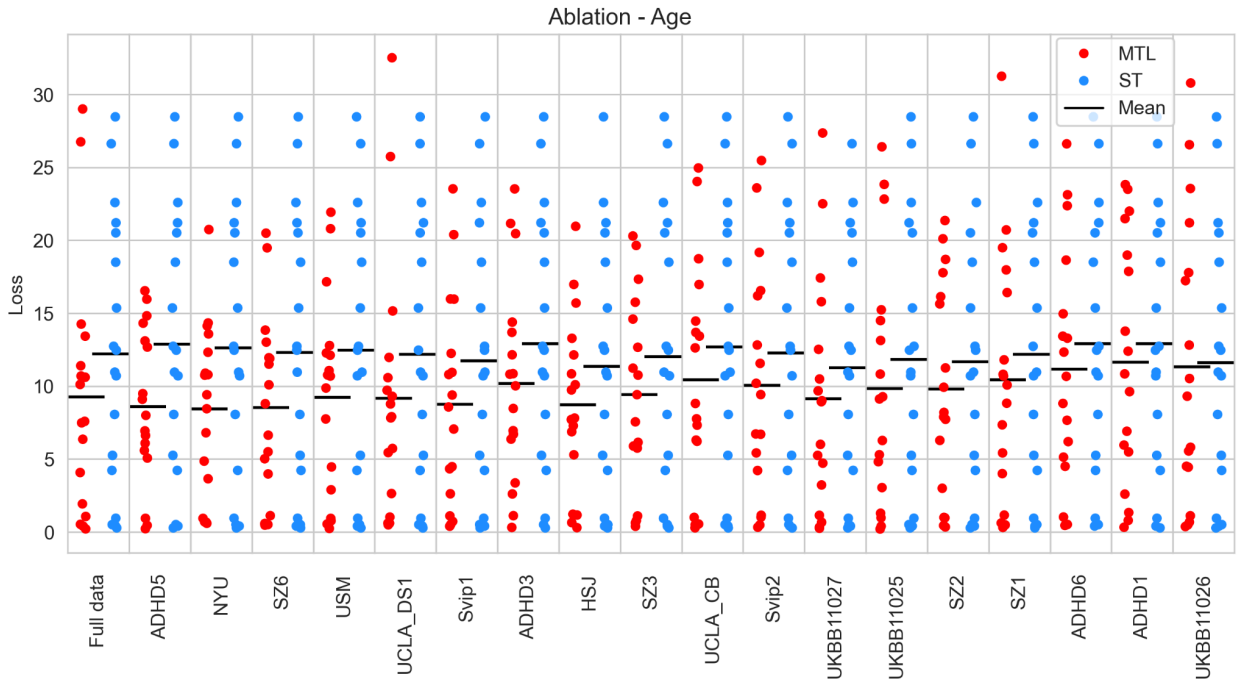

Figure 51 - Distribution of Mean Squared Error (MSE) of age prediction across tasks using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents the data collection site which is dropped from the set of prediction tasks, except for the first column which shows results using the full dataset. The y axis shows the MSE of prediction. For each removed dataset, the red points show prediction error distribution for the tasks using the MLPconn_reg architecture in MTL, and the blue points show prediction error distribution on the tasks trained using the MLPconn_reg architecture in ST.

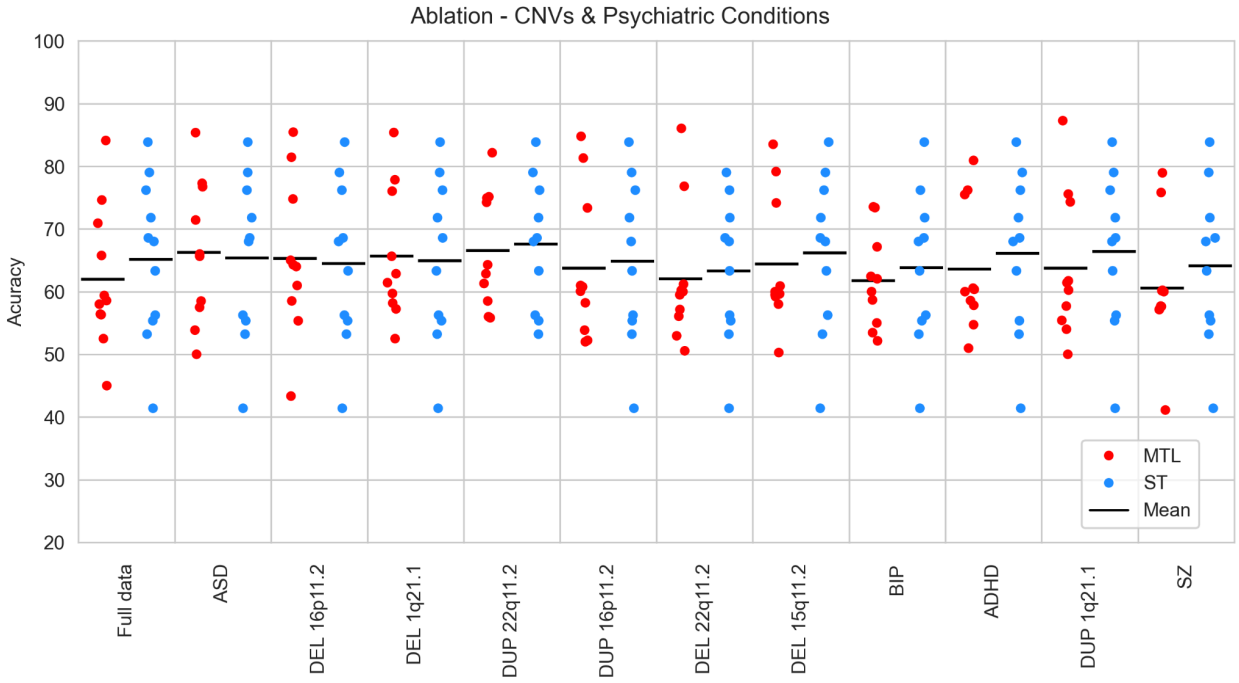

Figure 52 - Distribution of accuracy of automated diagnosis across tasks using single (ST) vs. multi-task learning (MTL) in a varied collection of sites. The x axis represents the condition which is dropped from the set of prediction

tasks, except for the first column which shows results using the full dataset. The y axis shows the accuracy of prediction. For each removed dataset, the red points show prediction accuracy distribution for the tasks using the MLPconn architecture in MTL, and the blue points show prediction accuracy distribution on the tasks trained using the MLPconn architecture in ST.

10.9 - Model Parameter Variations

Here we present the results of a sensitivity analysis in which we varied the parameters of our primary models (MLPconn for classification and MLPconn_reg for regression) in order to evaluate the impact on the performance of MTL in each setting (predicting age, sex and conditions). Training was performed as described in the methods (section 2.6).

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. For regression, the output layer is modified to have a single output so that the configuration becomes: 2080-256-64-64-64-1. The input to the model is the connectome vector.

The MLPconn_shorter model is a version of the MLPconn model with the two layers in the shared portion of the model replaced by a single layer with an intermediate width. The resulting configuration is 2080-128-2. For regression, the output layer is modified to have a single output so that the configuration becomes: 2080-128-1. The input to the model is the connectome vector.

The MLPconn_wider model is a version of the MLPconn model with layers that are double the width. The configuration is 2080-512-128-2. For regression, the output layer is modified to have a single output so that the configuration becomes: 2080-512-128-1. The input to the model is the connectome vector.

The MLPconn_thinner model is a version of the MLPconn model with layers that are half the width. The configuration is 2080-128-32-2. For regression, the output layer is modified to have a single output so that the configuration becomes: 2080-128-32-1. The input to the model is the connectome vector.

Regarding sex prediction across cohorts (Figure 53), we observed improved accuracy using MTL over ST, consistently across all variants of architecture. Regarding age prediction across cohorts (Figure 54), we observed improved accuracy (lower error) using MTL over ST for all but one architecture variant: MLPconn_deeper. This suggests that this highly parameterized model may be overfitting in the data regime where it is being trained. Finally, regarding diagnosis across psychiatric conditions and genetic variants (Figure 55), we observed decreased accuracy using MTL over ST for all but one variant: MLPconn_shorter, although the gains in this case are very marginal. This result suggests we may have over-parameterized our primary model MLPconn for this task, but still fails to demonstrate an advantage to MTL on this application. Overall, we found that the conclusions of our study are quite robust to the specific architectural choices we made for MLPconn.

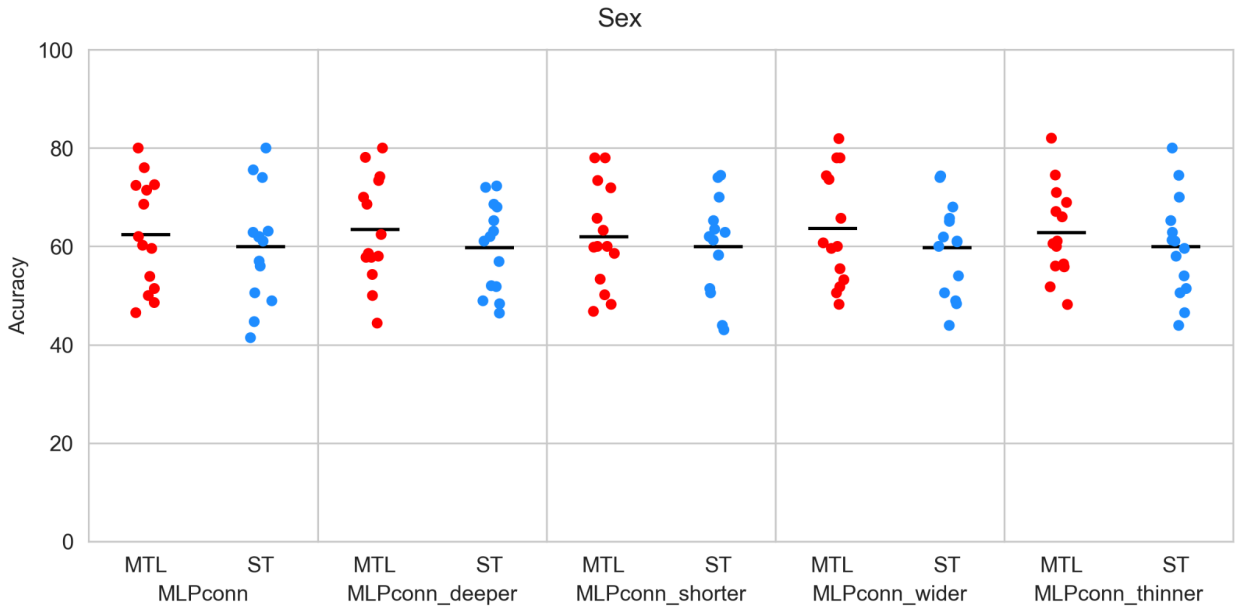

Figure 53 - Distribution of accuracy of sex prediction across model variations using single (ST) vs. multi-task learning (MTL). The x axis represents the model variations. The y axis shows the accuracy of prediction. For each model, the red points show prediction accuracy distribution for the tasks in MTL, and the blue points show prediction accuracy distribution on the tasks trained in ST.

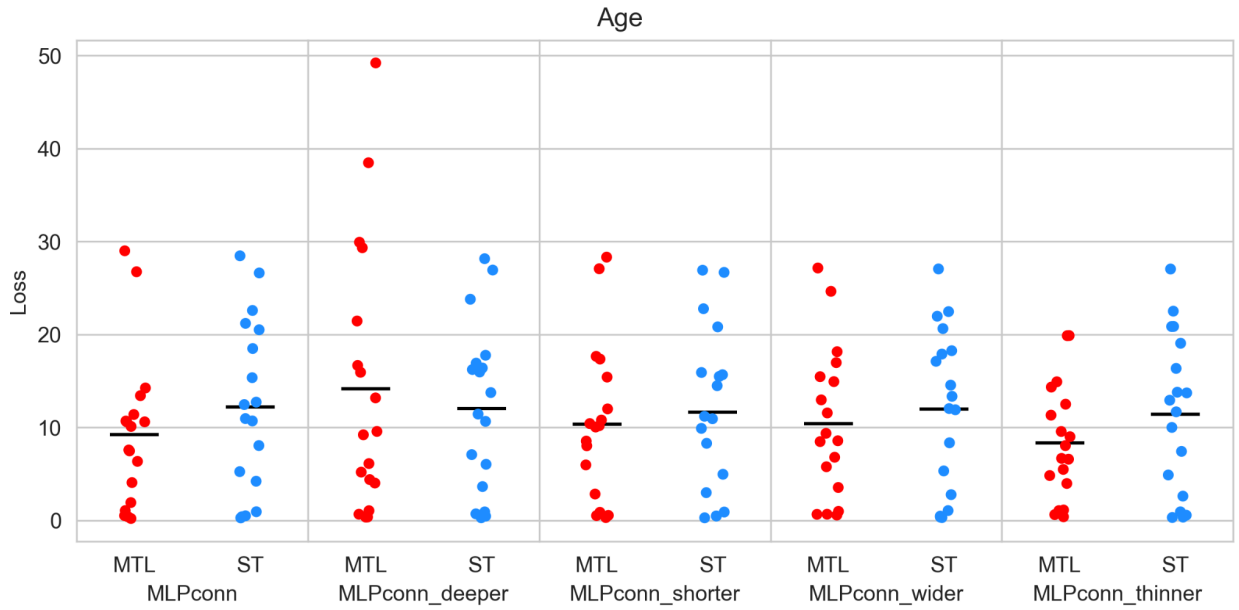

Figure 54 - Distribution of Mean Squared Error (MSE) of age prediction across model variations using single (ST) vs. multi-task learning (MTL). The x axis represents the model variations. The y axis shows the error of prediction. For each model, the red points show prediction accuracy distribution for the tasks in MTL, and the blue points show prediction accuracy distribution on the tasks trained in ST.

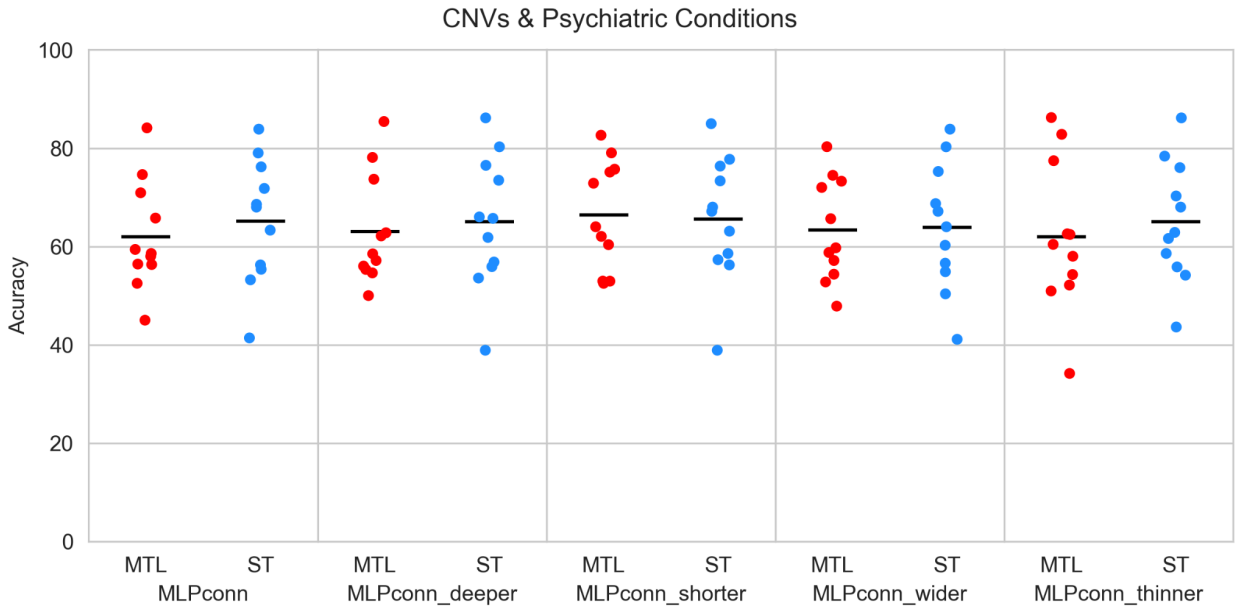

Figure 55 - Distribution of accuracy of automated diagnosis across model variations using single (ST) vs. multi-task learning (MTL). The x axis represents the model variations. The y axis shows the accuracy of prediction. For each model, the red points show prediction accuracy distribution for the tasks in MTL, and the blue points show prediction accuracy distribution on the tasks trained in ST.

10.10 - CNN Variations

Here we present the results of a sensitivity analysis in which we varied the parameters of our convolutional neural network model (CNN) in order to vary the format of the input to the model and evaluate the impact on the performance of MTL for automatic diagnosis. Our primary model uses a random permutation of the connectome as input, and therefore does not consider spatial information. The variations we present here (CNN_64 and CNN_clust, defined below) take as input the full connectome and a reordering of the full connectome determined by a functional clustering respectively. We chose the variants to test the impact of using formats that preserve more information about the spatial layout and functional similarity of the connectome. Training was performed as described in the methods (section 2.6).

The CNN_64 model is a convolutional neural network with a very similar architecture to the main CNN model (see methods section 2.5). Rather than taking the upper triangle of the symmetric connectome matrix (2080 values) randomly permuted and formatted into a 40 x 52 matrix as input, it takes as input the full 64x64 connectome matrix with regions as ordered in the original parcellation, which respects spatial groupings of the regions (S. Urchs et al., 2017). The model consists of a first convolution layer with 256 filters of shape 8 x 8, followed by two dense layers of 64 hidden units. The output layer has 2 units for binary classification. Batch normalisation (Ioffe & Szegedy, 2015) is applied after each layer.

The CNN_clust model has the same architecture as the CNN_64 model, but it takes as input the 64x64 connectome with regions grouped according to a hierarchical clustering performed using ward's criterion (Ward, 1963) over the mean connectome taken over all the subjects in our dataset.

We observed that MTL reached lower accuracy than ST for all choices, when applied to diagnosis across psychiatric conditions and genetic variants (Figure 56), consistent with the main results of our paper. We also observed that the model variants working directly on the 64x64 connectomes achieved similar performance to our main CNN architecture for STL, but performed much worse for MTL. We were not able to interpret that result. A possible culprit for the bad performance could be the inability to mix information from multiple networks in the convolutional layers, as neighbouring connections are by construction in similar networks. Ther fully connected layers also feature a much higher number of parameters in CNN_64 and CNN_clust, which may lead to overfitting as suggested by our experience of architecture variants (Figure 55).

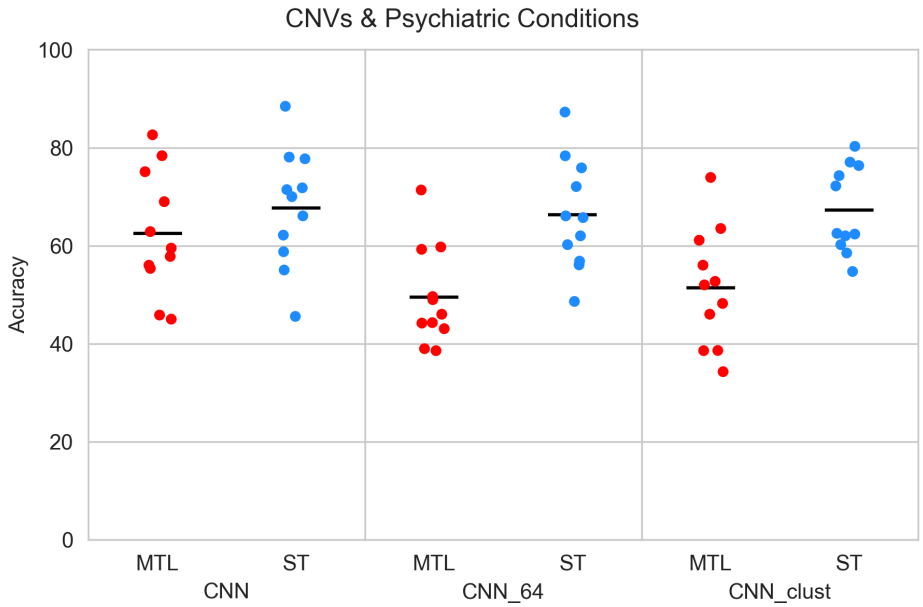

Figure 56 - Distribution of accuracy of automatic diagnosis across model variations using single (ST) vs. multi-task learning (MTL). The x axis represents the model variations. The y axis shows the accuracy of prediction. For each model, the red points show prediction accuracy distribution for the tasks in MTL, and the blue points show prediction accuracy distribution on the tasks trained in ST.

11 - References

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage, 147,* 736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045

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,* 62. https://doi.org/10.3389/fnsys.2012.00062

Attoe, D. E., & Climie, E. A. (2023). Miss. Diagnosis: A Systematic Review of ADHD in Adult Women. *Journal of Attention Disorders, 27*(7), 645–657. https://doi.org/10.1177/10870547231161533

Bassett, D. S., Nelson, B. G., Mueller, B. A., Camchong, J., & Lim, K. O. (2012). Altered resting state complexity in schizophrenia. *NeuroImage, 59*(3), 2196–2207. https://doi.org/10.1016/j.neuroimage.2011.10.002

Bellec, P., Benhajali, Y., Carbonell, F., Dansereau, C., Albouy, G., Pelland, M., Craddock, C., Collignon, O., Doyon, J., Stip, E., & Orban, P. (2015). Impact of the resolution of brain parcels on connectome-wide association studies in fMRI. *NeuroImage, 123,* 212–228. https://doi.org/10.1016/j.neuroimage.2015.07.071

Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J. P., Zijdenbos, A. P., & Evans, A. C. (2012). The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. *Frontiers in Neuroinformatics, 6,* 7. https://doi.org/10.3389/fninf.2012.00007

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and

powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bierer, B. E., Meloney, L. G., Ahmed, H. R., & White, S. A. (2022). Advancing the inclusion of underrepresented women in clinical research. *Cell Reports. Medicine, 3*(4), 100553. https://doi.org/10.1016/j.xcrm.2022.100553

Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences, 42*(4), 251–262. https://doi.org/10.1016/j.tins.2019.02.001

Caro, J. O., de O. Fonseca, A. H., Averill, C., Rizvi, S. A., Rosati, M., Cross, J. L., Mittal, P., Zappala, E., Levine, D., Dhodapkar, R. M., Abdallah, C. G., & van Dijk, D. (2023). BrainLM: A foundation model for brain activity recordings. In *bioRxiv* (p. 2023.09.12.557460). https://doi.org/10.1101/2023.09.12.557460

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science, 383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., & Ragoussis, J. (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), 2013–2025. https://doi.org/10.1093/nar/gkm076

Collins, R. L., Glessner, J. T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.-M., Ulirsch, J., Everett, S., Howrigan, D. P., Boone, P. M., Fu, J., Karczewski, K. J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., … Talkowski, M. E. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell, 185*(16),

3041–3055.e25. https://doi.org/10.1016/j.cell.2022.06.036

Crawford, K., Bracher-Smith, M., Owen, D., Kendall, K. M., Rees, E., Pardiñas, A. F., Einon, M., Escott-Price, V., Walters, J. T. R., O'Donovan, M. C., Owen, M. J., & Kirov, G. (2019). Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *Journal of Medical Genetics*, *56*(3), 131–138. https://doi.org/10.1136/jmedgenet-2018-105477

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., & Alzheimer's Disease Neuroimaging Initiative. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, *192*, 115–134. https://doi.org/10.1016/j.neuroimage.2019.02.062

Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K. J., Wassermann, D., Thirion, B., & Mensch, A. (2020). Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, *221*, 117126. https://doi.org/10.1016/j.neuroimage.2020.117126

Davies, R. W., Fiksinski, A. M., Breetvelt, E. J., Williams, N. M., Hooper, S. R., Monfeuga, T., Bassett, A. S., Owen, M. J., Gur, R. E., Morrow, B. E., McDonald-McGinn, D. M., Swillen, A., Chow, E. W. C., van den Bree, M., Emanuel, B. S., Vermeesch, J. R., van Amelsvoort, T., Arango, C., Armando, M., … Vorstman, J. A. S. (2020). Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nature Medicine*, *26*(12), 1912–1918. https://doi.org/10.1038/s41591-020-1103-1

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., … Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, *4*, 170010. https://doi.org/10.1038/sdata.2017.10

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf,

M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., … Milham, M. P. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Dong, Q., Zhang, J., Li, Q., Wang, J., Leporé, N., Thompson, P. M., Caselli, R. J., Ye, J., Wang, Y., & 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), 971–992. https://doi.org/10.3233/JAD-190973

Drakesmith, M., Parker, G. D., Smith, J., Linden, S. C., Rees, E., Williams, N., Owen, M. J., van den Bree, M., Hall, J., Jones, D. K., & Linden, D. E. J. (2019). Genetic risk for schizophrenia and developmental delay is associated with shape and microstructure of midline white-matter structures. *Translational Psychiatry*, *9*(1), 102. https://doi.org/10.1038/s41398-019-0440-7

El-Gazzar, A., Thomas, R. M., & van Wingen, G. (2023). Harmonization techniques for machine learning studies using multi-site functional MRI data. In *bioRxiv* (p. 2023.06.14.544758). https://doi.org/10.1101/2023.06.14.544758

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A. D., Joel, S., Pekar, J. J., Mostofsky, S. H., & Caffo, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, *6*, 61. https://doi.org/10.3389/fnsys.2012.00061

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L., & Brain Development Cooperative Group. (2011). Unbiased average age-appropriate atlases for

pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033

Giove, F., Gili, T., Iacovella, V., Macaluso, E., & Maraviglia, B. (2009). Images-based suppression of unwanted global signals in resting-state functional connectivity studies. *Magnetic Resonance Imaging*, *27*(8), 1058–1064. https://doi.org/10.1016/j.mri.2009.06.004

Hahn, S., Owens, M. M., Yuan, D., Juliano, A. C., Potter, A., Garavan, H., & Allgaier, N. (2022, April 3). *Performance Scaling for Structural MRI Surface Parcellations*. Performance Scaling for Structural MRI Surface Parcellations. https://sahahn.github.io/parc_scaling/

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., & Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage. Clinical*, *17*, 16–23. https://doi.org/10.1016/j.nicl.2017.08.017

He, L., Li, H., Wang, J., Chen, M., Gozdas, E., Dillman, J. R., & Parikh, N. A. (2020). A multi-task, multi-stage deep transfer learning model for early prediction of neurodevelopment in very preterm infants. *Scientific Reports*, *10*(1), 15072. https://doi.org/10.1038/s41598-020-71914-x

Huang, Z.-A., Liu, R., & Tan, K. C. (2020). Multi-Task Learning for Efficient Diagnosis of ASD and ADHD using Resting-State fMRI Data. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206852

Huang, Z.-A., Liu, R., Zhu, Z., & Tan, K. C. (2022). Multitask Learning for Joint Diagnosis of Multiple Mental Disorders in Resting-State fMRI. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. https://doi.org/10.1109/TNNLS.2022.3225179

Hu, D., & Zeng, L.-L. (2019). Multi-task Learning of Structural MRI for Multi-site Classification. In D. Hu & L.-L. Zeng (Eds.), *Pattern Analysis of the Human Connectome* (pp. 205–226). Springer Singapore. https://doi.org/10.1007/978-981-32-9523-0_11

Huguet, G., Schramm, C., Douard, E., Tamer, P., Main, A., Monin, P., England, J., Jizi, K.,

Renne, T., Poirier, M., Nowak, S., Martin, C.-O., Younis, N., Knoth, I. S., Jean-Louis, M., Saci, Z., Auger, M., Tihy, F., Mathonnet, G., … Jacquemont, S. (2021). Genome-wide analysis of gene dosage in 24,092 individuals estimates that 10,000 genes modulate cognitive ability. *Molecular Psychiatry*, *26*(6), 2663–2676. https://doi.org/10.1038/s41380-020-00985-z

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1502.03167

Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., & Pant, S. (2023). A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare (Basel, Switzerland)*, *11*(3). https://doi.org/10.3390/healthcare11030285

Jalbrzikowski, M., Lin, A., Vajdi, A., Grigoryan, V., Kushan, L., Ching, C. R. K., Schleifer, C., Hayes, R. A., Chu, S. A., Sugar, C. A., Forsyth, J. K., & Bearden, C. E. (2022). Longitudinal trajectories of cortical development in 22q11.2 copy number variants and typically developing controls. *Molecular Psychiatry*, *27*(10), 4181–4190. https://doi.org/10.1038/s41380-022-01681-w

Jonas, R. K., Montojo, C. A., & Bearden, C. E. (2014). The 22q11.2 deletion syndrome as a window into complex neuropsychiatric disorders over the lifespan. *Biological Psychiatry*, *75*(5), 351–360. https://doi.org/10.1016/j.biopsych.2013.07.019

Jønch, A. E., Douard, E., Moreau, C., Van Dijck, A., Passeggeri, M., Kooy, F., Puechberty, J., Campbell, C., Sanlaville, D., Lefroy, H., Richetin, S., Pain, A., Geneviève, D., Kini, U., Le Caignec, C., Lespinasse, J., Skytte, A.-B., Isidor, B., Zweier, C., … 15q11.2 Working Group. (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), 701–710. https://doi.org/10.1136/jmedgenet-2018-105879

Katzman, M. A., Bilkey, T. S., Chokka, P. R., Fallu, A., & Klassen, L. J. (2017). Adult ADHD and

comorbid disorders: clinical implications of a dimensional approach. *BMC Psychiatry*, *17*(1), 302. https://doi.org/10.1186/s12888-017-1463-3

Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. (2018). Ensemble learning with 3D convolutional neural networks for connectome-based prediction. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1809.06219

Kim, J., Calhoun, V. D., Shim, E., & Lee, J.-H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*, *124*(Pt A), 127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Leming, M., Górriz, J. M., & Suckling, J. (2020). Ensemble Deep Learning on Large, Mixed-Site fMRI Datasets in Autism and Other Tasks. *International Journal of Neural Systems*, *30*(7), 2050012. https://doi.org/10.1142/S0129065720500124

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

Liang, W., Zhang, K., Cao, P., Liu, X., Yang, J., & Zaiane, O. (2021). Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network. *Computers in Biology and Medicine*, *138*, 104935. https://doi.org/10.1016/j.compbiomed.2021.104935

Li, J., Joshi, A. A., & Leahy, R. M. (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*, 544–548. https://doi.org/10.1109/isbi45749.2020.9098584

Li, J., Kong, R., Liégeois, R., Orban, C., Tan, Y., Sun, N., Holmes, A. J., Sabuncu, M. R., Ge, T., & Yeo, B. T. T. (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage*, *196*, 126–141. https://doi.org/10.1016/j.neuroimage.2019.04.016

Lin, A., Ching, C. R. K., Vajdi, A., Sun, D., Jonas, R. K., Jalbrzikowski, M., Kushan-Wells, L., Pacheco Hansen, L., Krikorian, E., Gutman, B., Dokoru, D., Helleman, G., Thompson, P. M., & Bearden, C. E. (2017). Mapping 22q11.2 Gene Dosage Effects on Brain Morphometry. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(26), 6183–6199. https://doi.org/10.1523/JNEUROSCI.3759-16.2017

Linn, K. A., Gaonkar, B., Doshi, J., Davatzikos, C., & Shinohara, R. T. (2016). Addressing Confounding in Predictive Models with an Application to Neuroimaging. *The International Journal of Biostatistics*, *12*(1), 31–44. https://doi.org/10.1515/ijb-2015-0030

Li, X., Zhou, W., & Yi, Z. (2022). A glimpse of gender differences in schizophrenia. *General Psychiatry*, *35*(4), e100823. https://doi.org/10.1136/gpsych-2022-100823

Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, *112*(45), 13892–13897. https://doi.org/10.1073/pnas.1518285112

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child and Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., & Nichols, T. E. (2006). Non-white noise in fMRI: does modelling have an impact? *NeuroImage*, *29*(1), 54–66. https://doi.org/10.1016/j.neuroimage.2005.07.005

Mahamud, F., Emon, A. S., Nahar, N., Imam, M. H., Hossain, M. S., & Andersson, K. (2023). Transfer Learning Based Method for Classification of Schizophrenia Using MobileNet. *Intelligent Computing & Optimization*, 210–220. https://doi.org/10.1007/978-3-031-19958-5_20

Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., & Chi, E. H. (2018). Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1930–1939. https://doi.org/10.1145/3219819.3220007

Ma, Q., Zhang, T., Zanetti, M. V., Shen, H., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Fan, Y., Hu, D., Busatto, G. F., & Davatzikos, C. (2018). Classification of multi-site MR images in the presence of heterogeneity using multi-task learning. *NeuroImage. Clinical*, *19*, 476–486. https://doi.org/10.1016/j.nicl.2018.04.037

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

Marquand, A. F., Brammer, M., Williams, S. C. R., & Doyle, O. M. (2014). Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage*, *92*(100), 298–311. https://doi.org/10.1016/j.neuroimage.2014.02.008

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S.,

Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajarado, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., … CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, *49*(1), 27–35. https://doi.org/10.1038/ng.3725

{Martineau, J. L. M., Main, A., & Jacquemont, S. J. (n.d.). *Python based parallel CNV calling prioritizing mpi4py usage and memory optimization*. https://doi.org/10.5281/zenodo.3497400

Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, *42*(2), 275–293. https://doi.org/10.1007/s10462-012-9338-y

McElroy, S. L. (2004). Diagnosing and treating comorbid (complicated) bipolar disorder. *The Journal of Clinical Psychiatry*, *65 Suppl 15*, 35–44. https://www.ncbi.nlm.nih.gov/pubmed/15554795

Mellema, C. J., Nguyen, K. P., Treacher, A., & Montillo, A. (2022). Reproducible neuroimaging features for diagnosis of autism spectrum disorder with machine learning. *Scientific Reports*, *12*(1), 3057. https://doi.org/10.1038/s41598-022-06459-2

Merikangas, K. R., Jin, R., He, J.-P., Kessler, R. C., Lee, S., Sampson, N. A., Viana, M. C., Andrade, L. H., Hu, C., Karam, E. G., Ladea, M., Medina-Mora, M. E., Ono, Y., Posada-Villa, J., Sagar, R., Wells, J. E., & Zarkov, Z. (2011). Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of General Psychiatry*, *68*(3), 241–251. https://doi.org/10.1001/archgenpsychiatry.2011.12

Modenato, C., Kumar, K., Moreau, C., Martin-Brevet, S., Huguet, G., Schramm, C., Jean-Louis, M., Martin, C.-O., Younis, N., Tamer, P., Douard, E., Thébault-Dagher, F., Côté, V., Charlebois, A.-R., Deguire, F., Maillard, A. M., Rodriguez-Herreros, B., Pain, A., Richetin, S., … Jacquemont. (2021). Effects of eight neuropsychiatric copy number variants on

human brain structure. *Translational Psychiatry*, *11*(1), 399. https://doi.org/10.1038/s41398-021-01490-9

Moreau, C. A., Ching, C. R., Kumar, K., Jacquemont, S., & Bearden, C. E. (2021). Structural and functional brain alterations revealed by neuroimaging in CNV carriers. *Current Opinion in Genetics & Development*, *68*, 88–98. https://doi.org/10.1016/j.gde.2021.03.002

Moreau, C. A., Harvey, A., Kumar, K., Huguet, G., Urchs, S. G. W., Douard, E. A., Schultz, L. M., Sharmarke, H., Jizi, K., Martin, C.-O., Younis, N., Tamer, P., Rolland, T., Martineau, J.-L., Orban, P., Silva, A. I., Hall, J., van den Bree, M. B. M., Owen, M. J., … Jacquemont, S. (2023). Genetic Heterogeneity Shapes Brain Connectivity in Psychiatry. *Biological Psychiatry*, *93*(1), 45–58. https://doi.org/10.1016/j.biopsych.2022.08.024

Moreau, C. A., Kumar, K., Harvey, A., Huguet, G., Urchs, S., Schultz, L. M., Sharmarke, H., Jizi, K., Martin, C. O., Younis, N., Tamer, P., Martineau, J. L., Orban, P., Silva, A. I., Hall, J., van den Bree, M. B. M., Owen, M. J., Linden, D. E. J., Lippé, S., … Jacquemont, S. (2022). Brain functional connectivity mirrors genetic pleiotropy in psychiatric conditions. *Brain: A Journal of Neurology*. https://doi.org/10.1093/brain/awac315

Moreau, C. A., Raznahan, A., Bellec, P., Chakravarty, M., Thompson, P. M., & Jacquemont, S. (2021). Dissecting autism and schizophrenia through neuroimaging genomics. *Brain: A Journal of Neurology*, *144*(7), 1943–1957. https://doi.org/10.1093/brain/awab096

Moreau, C. A., Urchs, S. G. W., Kuldeep, K., Orban, P., Schramm, C., Dumas, G., Labbe, A., Huguet, G., Douard, E., Quirion, P.-O., Lin, A., Kushan, L., Grot, S., Luck, D., Mendrek, A., Potvin, S., Stip, E., Bourgeron, T., Evans, A. C., … Jacquemont, S. (2020). Mutations associated with neuropsychiatric conditions delineate functional brain connectivity dimensions contributing to autism and schizophrenia. *Nature Communications*, *11*(1), 5272. https://doi.org/10.1038/s41467-020-18997-2

Moreno-De-Luca, D., Sanders, S. J., Willsey, A. J., Mulle, J. G., Lowe, J. K., Geschwind, D. H., State, M. W., Martin, C. L., & Ledbetter, D. H. (2013). Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Molecular Psychiatry*, *18*(10), 1090–1095. https://doi.org/10.1038/mp.2012.138

Moreno-Küstner, B., Martín, C., & Pastor, L. (2018). Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PloS One*, *13*(4), e0195687. https://doi.org/10.1371/journal.pone.0195687

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36. https://doi.org/10.4097/kja.21209

Ngo, D.-K., Tran, M.-T., Kim, S.-H., Yang, H.-J., & Lee, G.-S. (2020). Multi-Task Learning for Small Brain Tumor Segmentation from MRI. *NATO Advanced Science Institutes Series E: Applied Sciences*, *10*(21), 7790. https://doi.org/10.3390/app10217790

Niarchou, M., Chawner, S. J. R. A., Doherty, J. L., Maillard, A. M., Jacquemont, S., Chung, W. K., Green-Snyder, L., Bernier, R. A., Goin-Kochel, R. P., Hanson, E., Linden, D. E. J., Linden, S. C., Raymond, F. L., Skuse, D., Hall, J., Owen, M. J., & van den Bree, M. B. M. (2019). Psychiatric disorders in children with 16p11.2 deletion and duplication. *Translational Psychiatry*, *9*(1), 8. https://doi.org/10.1038/s41398-018-0339-8

Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., Lainhart, J. E., & Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. *Frontiers in Human Neuroscience*, *7*, 599. https://doi.org/10.3389/fnhum.2013.00599

Orban, P., Dansereau, C., Desbois, L., Mongeau-Pérusse, V., Giguère, C.-É., Nguyen, H., Mendrek, A., Stip, E., & Bellec, P. (2018). Multisite generalizability of schizophrenia diagnosis classification based on functional brain connectivity. *Schizophrenia Research*, *192*,

167–171. https://doi.org/10.1016/j.schres.2017.05.027

Orban, P., Desseilles, M., Mendrek, A., Bourque, J., Bellec, P., & Stip, E. (2017). Altered brain connectivity in patients with schizophrenia is consistent across cognitive contexts. *Journal of Psychiatry & Neuroscience: JPN*, *42*(1), 17–26. https://doi.org/10.1503/jpn.150247

Pacheco, J., Garvey, M. A., Sarampote, C. S., Cohen, E. D., Murphy, E. R., & Friedman-Hill, S. R. (2022). Annual Research Review: The contributions of the RDoC research framework on understanding the neurodevelopmental origins, progression and treatment of mental illnesses. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *63*(4), 360–376. https://doi.org/10.1111/jcpp.13543

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *arXiv [cs.LG]*. arXiv. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Pedersen, S. L., Lindstrom, R., Powe, P. M., Louie, K., & Escobar-Viera, C. (2022). Lack of Representation in Psychiatric Research: A Data-Driven Example From Scientific Articles Published in 2019 and 2020 in the American Journal of Psychiatry. *The American Journal of Psychiatry*, *179*(5), 388–392. https://doi.org/10.1176/appi.ajp.21070758

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K. H., Mumford, J. A., Sabb, F. W., Freimer, N. B., London, E. D., Cannon, T. D., & Bilder, R. M. (2016). A phenome-wide examination of neural and cognitive function. *Scientific Data, 3*, 160110. https://doi.org/10.1038/sdata.2016.110

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage, 59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Raghav, A., Anand, A., Sharma, R., Singh, N., & RamiReddy, C. V. (2023). Autism Spectrum Disorder Detection in Children Using Transfer Learning Techniques. *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, 550–555. https://doi.org/10.1109/ICECAA58104.2023.10212257

Rahim, M., Thirion, B., Bzdok, D., Buvat, I., & Varoquaux, G. (2017). Joint prediction of multiple scores captures better individual traits from brain images. *NeuroImage, 158*, 145–154. https://doi.org/10.1016/j.neuroimage.2017.06.072

Rao, N., Cox, C., Nowak, R., & Rogers, T. (2013). Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1311.5422

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., & Calhoun, V. D. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *NeuroImage, 134*, 645–657. https://doi.org/10.1016/j.neuroimage.2016.04.051

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, 11*(3), 269–282. https://doi.org/10.14778/3157794.3157797

Rees, E., & Kirov, G. (2021). Copy number variation and neuropsychiatric illness. *Current Opinion in Genetics & Development*, *68*, 57–63. https://doi.org/10.1016/j.gde.2021.02.014

Roffet, F., Delrieux, C., & Patow, G. (2022). Assessing Multi-Site rs-fMRI-Based Connectomic Harmonization Using Information Theory. *Brain Sciences*, *12*(9). https://doi.org/10.3390/brainsci12091219

Romero, C., Werme, J., Jansen, P. R., Gelernter, J., Stein, M. B., Levey, D., Polimanti, R., de Leeuw, C., Nagel, M., & van der Sluis, S. (2022). Exploring the genetic overlap between twelve psychiatric disorders. *Nature Genetics*, *54*(12), 1795–1802. https://doi.org/10.1038/s41588-022-01245-2

Rösler, M., Casas, M., Konofal, E., & Buitelaar, J. (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), 684–698. https://doi.org/10.3109/15622975.2010.483249

Sajatovic, M. (2005). Bipolar disorder: disease burden. *The American Journal of Managed Care*, *11*(3 Suppl), S80–S84. https://www.ncbi.nlm.nih.gov/pubmed/16097718

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., Murtha, M. T., Bal, V. H., Bishop, S. L., Dong, S., Goldberg, A. P., Jinlu, C., Keaney, J. F., 3rd, Klei, L., Mandell, J. D., Moreno-De-Luca, D., Poultney, C. S., Robinson, E. B., Smith, L., … State, M. W. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, *87*(6), 1215–1233. https://doi.org/10.1016/j.neuron.2015.09.016

Sanders, S. J., Sahin, M., Hostyk, J., Thurm, A., Jacquemont, S., Avillach, P., Douard, E., Martin, C. L., Modi, M. E., Moreno-De-Luca, A., Raznahan, A., Anticevic, A., Dolmetsch, R., Feng, G., Geschwind, D. H., Glahn, D. C., Goldstein, D. B., Ledbetter, D. H., Mulle, J. G., … Bearden, C. E. (2019). A framework for the investigation of rare genetic disorders in

neuropsychiatry. *Nature Medicine*, *25*(10), 1477–1487. https://doi.org/10.1038/s41591-019-0581-5

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568–584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Schleifer, C. H., O'Hora, K. P., Jalbrzikowski, M., Bondy, E., Kushan-Wells, L., Lin, A., Uddin, L. Q., & Bearden, C. E. (2023). Longitudinal development of thalamocortical functional connectivity in 22q11.2 deletion syndrome. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*. https://doi.org/10.1016/j.bpsc.2023.09.001

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), 4238. https://doi.org/10.1038/s41467-020-18037-z

Shmueli, G. (2010). To Explain or to Predict? *Schweizerische Monatsschrift Fur Zahnheilkunde = Revue Mensuelle Suisse D'odonto-Stomatologie / SSO*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., & Baird, G. (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), 921–929. https://doi.org/10.1097/CHI.0b013e318179964f

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), 1063–1067. https://doi.org/10.1016/j.neuron.2012.02.014

Simon, V., Czobor, P., Bálint, S., Mészáros, A., & Bitter, I. (2009). Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. *The British Journal of Psychiatry: The Journal of Mental Science*, *194*(3), 204–211. https://doi.org/10.1192/bjp.bp.107.048827

Sønderby, I. E., Ching, C. R. K., Thomopoulos, S. I., van der Meer, D., Sun, D., Villalon-Reina, J. E., Agartz, I., Amunts, K., Arango, C., Armstrong, N. J., Ayesa-Arriola, R., Bakker, G., Bassett, A. S., Boomsma, D. I., Bülow, R., Butcher, N. J., Calhoun, V. D., Caspers, S., Chow, E. W. C., … ENIGMA 22q11.2 Deletion Syndrome Working Group. (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. https://doi.org/10.1002/hbm.25354

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: A simple way to prevent neural networks from overfitting*. https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer,

Standley, T., Zamir, A. R., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2019). Which Tasks Should Be Learned Together in Multi-task Learning? In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1905.07553

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (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. https://doi.org/10.1371/journal.pmed.1001779

Tabarestani, S., Eslami, M., Cabrerizo, M., Curiel, R. E., Barreto, A., Rishe, N., Vaillancourt, D., DeKosky, S. T., Loewenstein, D. A., Duara, R., & Adjouadi, M. (2022). A Tensorized Multitask Deep Learning Network for Progression Prediction of Alzheimer's Disease. *Frontiers in Aging Neuroscience*, *14*, 810873. https://doi.org/10.3389/fnagi.2022.810873

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, *15*, 29. https://doi.org/10.1186/s12880-015-0068-x

Traut, N., Heuer, K., Lemaître, G., Beggiato, A., Germanaud, D., Elmaleh, M., Bethegnies, A., Bonnasse-Gahot, L., Cai, W., Chambon, S., Cliquet, F., Ghriss, A., Guigui, N., de Pierrefeu, A., Wang, M., Zantedeschi, V., Boucaud, A., van den Bossche, J., Kegl, B., … Varoquaux, G. (2022). Insights from an autism imaging biomarker challenge: Promises and threats to biomarker discovery. *NeuroImage*, *255*, 119171. https://doi.org/10.1016/j.neuroimage.2022.119171

Tsai, J., & Rosenheck, R. A. (2013). Psychiatric comorbidity among adults with schizophrenia: a latent class analysis. *Psychiatry Research*, *210*(1), 16–20. https://doi.org/10.1016/j.psychres.2013.05.013

Urchs, S., Armoza, J., Benhajali, Y., St-Aubin, J., Orban, P., & Bellec, P. (2017). MIST: A multi-resolution parcellation of functional brain networks. *MNI Open Research*, *1*, 3. https://doi.org/10.12688/mniopenres.12767.1

Urchs, S. G. W., Nguyen, H. D., Moreau, C., Dansereau, C., Tam, A., Evans, A. C., & Bellec, P. (2020). Reproducible functional connectivity endophenotype confers high risk of ASD diagnosis in a subset of individuals. In *bioRxiv* (p. 2020.06.01.127688). https://doi.org/10.1101/2020.06.01.127688

Vanes, L. D., & Dolan, R. J. (2021). Transdiagnostic neuroimaging markers of psychiatric risk: A narrative review. *NeuroImage. Clinical, 30*, 102634. https://doi.org/10.1016/j.nicl.2021.102634

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage, 180*(Pt A), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Venkataraman, A., Whitford, T. J., Westin, C.-F., Golland, P., & Kubicki, M. (2012). Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophrenia Research, 139*(1-3), 7–12. https://doi.org/10.1016/j.schres.2012.04.021

Wain, L. V., Shrine, N., Miller, S., Jackson, V. E., Ntalla, I., Soler Artigas, M., Billington, C. K., Kheirallah, A. K., Allen, R., Cook, J. P., Probert, K., Obeidat, M. 'en, Bossé, Y., Hao, K., Postma, D. S., Paré, P. D., Ramasamy, A., UK Brain Expression Consortium (UKBEC), Mägi, R., … Hall, I. P. (2015). Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet. Respiratory Medicine, 3*(10), 769–781. https://doi.org/10.1016/S2213-2600(15)00283-0

Wang, H., Zhu, R., Tian, S., Shao, J., Dai, Z., Xue, L., Sun, Y., Chen, Z., Yao, Z., & Lu, Q. (2022). Classification of bipolar disorders using the multilayer modularity in dynamic minimum spanning tree from resting state fMRI. *Cognitive Neurodynamics*. https://doi.org/10.1007/s11571-022-09907-x

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., & Bucan, M. (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), 1665–1674. https://doi.org/10.1101/gr.6861907

Wang, X., Zhang, T., Chaim, T. M., Zanetti, M. V., & Davatzikos, C. (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*, 125–132. https://doi.org/10.1007/978-3-319-24553-9_16

Wang, Y.-W., Chen, X., & Yan, C.-G. (2023). Comprehensive evaluation of harmonization on functional brain imaging for multisite data-fusion. *NeuroImage*, *274*, 120089. https://doi.org/10.1016/j.neuroimage.2023.120089

Wang, Z., Zhou, X., Gui, Y., Liu, M., & Lu, H. (2023). Multiple measurement analysis of resting-state fMRI for ADHD classification in adolescent brain from the ABCD study. *Translational Psychiatry*, *13*(1), 45. https://doi.org/10.1038/s41398-023-02309-5

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Watanabe, T., Kessler, D., Scott, C., & Sripada, C. (2014). *Multisite disease classification with functional connectomes via multitask structured sparse SVM*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=da3752254a91621ecbfa8bfe06d109bdffd45549

Willsey, H. R., Willsey, A. J., Wang, B., & State, M. W. (2022). Genomics, convergent neuroscience and progress in understanding autism spectrum disorder. *Nature Reviews. Neuroscience*, *23*(6), 323–341. https://doi.org/10.1038/s41583-022-00576-7

Xiao, L., Stephen, J. M., Wilson, T. W., Calhoun, V. D., & Wang, Y.-P. (2020). A Manifold Regularized Multi-Task Learning Model for IQ Prediction From Two fMRI Paradigms. *IEEE Transactions on Biomedical Engineering*, *67*(3), 796–806. https://doi.org/10.1109/TBME.2019.2921207

Xie, C., Xiang, S., Shen, C., Peng, X., Kang, J., Li, Y., Cheng, W., He, S., Bobou, M., Broulidakis,

M. J., van Noort, B. M., Zhang, Z., Robinson, L., Vaidya, N., Winterer, J., Zhang, Y., King, S., Banaschewski, T., Barker, G. J., … ZIB Consortium. (2023). A shared neural basis underlying psychiatric comorbidity. *Nature Medicine, 29*(5), 1232–1242. https://doi.org/10.1038/s41591-023-02317-4

Yan, C.-G., Craddock, R. C., Zuo, X.-N., Zang, Y.-F., & Milham, M. P. (2013). Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. *NeuroImage, 80*, 246–262. https://doi.org/10.1016/j.neuroimage.2013.04.081

Yu, C., Cui, D., Shang, M., Zhang, S., Guo, L., Han, J., Du, L., & Alzheimer's Disease Neuroimaging Initiative. (2021). A Multi-task Deep Feature Selection Method for Brain Imaging Genetics. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/2107.00388

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., & Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism Research: Official Journal of the International Society for Autism Research, 15*(5), 778–790. https://doi.org/10.1002/aur.2696

Zhang, D., Shen, D., & 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. https://doi.org/10.1016/j.neuroimage.2011.09.069

Zhou, J., Liu, J., Narayan, V. A., Ye, J., & Alzheimer's Disease Neuroimaging Initiative. (2013). Modeling disease progression via multi-task learning. *NeuroImage, 78*, 233–248. https://doi.org/10.1016/j.neuroimage.2013.03.073