Aus dem Zentralinstitut für Seelische Gesundheit
der Medizinischen Fakultät Mannheim
(Direktor: Prof. Dr. med. Andreas Meyer-Lindenberg)

Dimensional reconstruction of psychotic disorders through multi-task learning

Inauguraldissertation
zur Erlangung des Doktor scientiarum humanarum (Dr.sc.hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Han Cao

aus
HENAN, China
2021

Dekan: Prof. Dr. med. Sergij Goerdt
Referent: Dr. Emanuel Schwarz

To my family

TABLE OF CONTENTS

ABBREVIATIONS **1**

1 INTRODUCTION **3**

1.1 Background 3

1.1.1 Clinical features of schizophrenia 3

1.1.2 Genetics and pleiotropy of schizophrenia 4

1.2 Multi-modal data integration using machine learning 6

1.2.1 Genomics 6

1.2.2 Transcriptomics 7

1.2.3 Neuroimaging 9

1.2.4 Dimensional reconstruction of mental illness using machine learning 9

1.3 Regularization based machine learning 10

1.3.1 Penalization on coefficients 11

1.3.2 Penalization on the difference of coefficients 12

1.4 Cross-task regularization based multi-task learning 14

1.4.1 MTL with structural regularization 15

1.4.2 MTL incorporating pairwise task similarity 16

1.4.3 Federated MTL 16

2 STUDY 1: RMTL: AN R LIBRARY FOR MULTI-TASK LEARNING **18**

2.1 Abstract 18

2.2 Introduction 18

2.3 Methods 18

2.4 Results 19

2.4.1 Model interpretability 19

2.4.2 Predictive performance 19

2.5 Conclusion 20

2.6 Supplements 20

2.6.1 Supplementary methods 20

2.6.2 Supplementary Figures 22

3 STUDY 2: COMPARATIVE EVALUATION OF MACHINE LEARNING STRATEGIES FOR ANALYZING BIG DATA IN PSYCHIATRY 24

3.1 Abstract 24

3.2 Introduction 24

3.3 Materials and Methods 25

3.3.1 Datasets 25

3.3.2 Preprocessing 26

3.3.3 Machine learning approaches 26

3.4 Results 29

3.4.1 Accuracy comparison between MTL and STL 29

3.4.2 Dependency of classification performance on the number of training datasets 30

3.4.3 Consistency and stability of trained models 31

3.5 Discussion 33

3.6 Supplements 34

3.6.1 Supplementary Methods 34

3.6.2 Supplementary Figures 35

4 STUDY 3: DSMTL - A COMPUTATIONAL FRAMEWORK FOR PRIVACY-PRESERVING, DISTRIBUTED MULTI-TASK MACHINE LEARNING 38

4.1 Abstract 38

4.2 Introduction 38

4.3 Results 39

4.4 Discussion 41

4.5 Methods 42

4.6 Supplements 47

4.6.1 Supplementary methods 47

4.6.2 Supplementary Results 59

5 DISCUSSION 61

5.1 Shared molecular alterations between schizophrenia and T2D 62

5.2 Standalone and distributed MTL 62

5.3 Heterogeneity in brain expression data of individuals with schizophrenia 62

5.4 Limitations 63
5.5 Outlook 63

6 SUMMARY 66

7 REFERENCES 67

8 PUBLICATIONS 86

9 CURRICULUM VITAE 88

10 ACKNOWLEDGEMENT 89

ABBREVIATIONS

AD	Alzheimer' disease
ADHD	Attention-deficit/hyperactivity disorder
BMI	Body mass index
CNV	Copy Number Variants
CV	Cross-validation
DC	Difference of two convex functions
DLPFC	Dorsolateral prefrontal cortex
DNA	Deoxyribonucleic acid
dsMTL	Federated multi-task learning on DataSHIELD
dsMTL_iNMF	Federated MTL with integrative NMF
dsMTL_L21	Federated MTL with joint feature selection
dsMTL_net	Federated MTL with network incorporation
dsMTL_trace	Federated MTL with trace-norm regularization
dsLasso	Federated Lasso
eQTL	Expression quantitative trait loci
FDR	False Discovery Rate
FeMTL	Federated MTL
fMRI	Functional magnetic resonance imaging
GEO	Gene Expression Omnibus
GPU	Graphics processing unit
GO	Gene Ontology
GWAS	Genome-wide association study
HBCC	Human Brain Collection Core
HIV	Human immunodeficiency viruses
HMM	Hidden Markov Model
iNMF	integrative non-negative matrix factorization
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage disequilibrium
LR	Logistic regression with regularization
ML	Machine learning
MTL	Multi-task learning
MTL_CMTL	MTL with clustering structure
MTL_Graph	MTL with graph incorporation
MTL_L21	MTL with joint feature selection
MTL_Lasso	MTL with Lasso
MTL_NET	MTL with network constraints
MTL_SNET	MTL with sparse network constraints
MTL_Trace	MTL with trace-norm constraints
NB	Negative binomial distribution
NGS	Next-generation sequencing
PPI	Protein-protein interactions
PRS	Polygenic risk score
RIN	RNA Integrity Number
RF	Random forest
RMA	Robust Multiarray Averaging
RMTL	Regularized multi-task learning
RNA-Seq	Ribonucleic acid sequencing
sMRI	Structural Magnetic Resonance Imaging
SNP	Single nucleotide polymorphisms

SSL	Secure Sockets Layer
STL	Single task learning
SVM	Support vector machine

1 INTRODUCTION

1.1 Background

Mental disorders are highly prevalent conditions that cause enormous suffering, and a massive clinical and socioeconomic burden. 38.2% (164.8 million) of the adult European population has been affected by at least one mental disorder (Wittchen, Jacobi et al. 2011). The top three most frequent disorders are anxiety disorders (14.0%), insomnia (7.0%) and major depression (6.9%). Patients with mental disorders have a reduced life expectancy of up to 20 years (De Hert, Correll et al. 2011). This is mirrored in the fact that 1.5m people try to commit suicide every year in Europe alone (De Hert, Correll et al. 2011). During the recent pandemic of covid-19 (2019~present), a significantly increased prevalence of mental disorders has been observed in Germany across all dimensions (Bauerle, Teufel et al. 2020), including anxiety (44.9%) and depression (14.3%).

Schizophrenia is a severe mental disorder affecting 20 million individuals worldwide (James, Abate et al. 2018). Patients with schizophrenia experience a broad spectrum of clinical symptoms, frequently show reduced educational and occupational performance, have a 2 to 3 fold elevated risk for early death due to comorbid somatic diseases such as cardiovascular disease (Laursen, Nordentoft et al. 2014), and commonly experience stigma, and discrimination. The onset of schizophrenia typically occurs between late adolescence and early adulthood, with males frequently showing an approximately 5 years earlier age-of-onset and a sharper age-related incidence peak (DeLisi 1992). The extensive efforts made to advance biological and clinical research have thus far not resulted in substantial improvements of the illness' clinical management, in particular due to our still incomplete understanding of its underlying biology. However, the field is now advancing to a stage where such progress appears within reach, due to the strongly intensifying collaborative integration of research efforts, the increasing availability of deeply phenotyped, transdiagnostic patient populations, and the rapidly advancing progress in data science. Leveraging this multidisciplinary expertise will advance our understanding of molecular mechanisms relevant at the individual patient level, and provide the basis for the development of novel approaches for diagnosis, at-risk identification, therapy selection, and the development of mechanistically-informed treatment approaches.

1.1.1 Clinical features of schizophrenia

The definition of schizophrenia by Emil Kraepelin dates back to over 100 years ago (Lehmann and Ban 1997). Since then, the concept and boundary of schizophrenia has kept evolving with affected patients typically displaying a mixture of positive, negative, cognitive, mood and motor symptoms (Tandon, Nasrallah et al. 2009). Positive symptoms are related to the impaired perception of reality and include delusions and hallucinations. Negative symptoms refer to a reduction or loss of affective and cognitive functions, leading to e.g. avolition and apathy. The severity, composition and course of such symptoms varies widely between schizophrenia patients, causing extensive clinical heterogeneity at the patient group level. For example, the age of onset range is relatively broad and chronically hospitalized patients with an earlier age of onset have been reported to be more strongly affected by cognitive impairment and negative symptoms (Johnstone, Owens et al. 1989). As further detailed below, this clinical heterogeneity goes hand-in-hand with a pronounced biological heterogeneity where illness-related changes frequently cross diagnostic boundaries. This substantially complicates the

identification of illness-specific biological signatures and the characterization of the illness' underlying biology.

Numerous studies have explored the possibility to identify diagnostic biomarker signatures for schizophrenia, covering a broad spectrum of data modalities, e.g. including neurocognitive dysfunction and brain morphology (Jablensky 2010). However, these signatures have thus far not been sufficiently sensitive and specific to be clinically useful as a diagnostic test for schizophrenia. The current diagnostic systems DSM-5 (Statistical Manual of Mental Disorders) and ICD-10 (the International Classification of Diseases 10) are thus still largely based on the evaluation of clinical symptoms and patient history.

A similar lack of objective biological tools exists in the context of therapy selection. The major molecular targets of antipsychotic therapy are the dopamine D2 receptors and 5-HT2 receptors of the Central Nervous System (CNS). The antipsychotic drugs aim at blocking these receptors (Gaebel and Zielasek 2015) and are typically effective in reducing positive (e.g., delusions), but not negative symptoms or cognitive impairment (Seeman 2004). The so-called "atypical" antipsychotic drugs are not associated with the extrapyramidal side effects that frequently occur after treatment with first generation, "typical" antipsychotics, but often induce a spectrum of metabolic adverse effects (e.g., increase of plasma glucose). Due to the difficulties of predicting inter-individual differences of the susceptibility to such side-effects, or the likelihood of response to a given therapy, the clinical management of schizophrenia is unfortunately still characterized be repeated try-outs of treatment approaches.

Therefore, although the current diagnostic system and approach to treatment have heuristic clinical utility, it is widely accepted that the clinical management, as well as the development of novel therapies could be substantially improved by a more in-depth understanding of illness biology, the development of mechanistically-informed biomarkers, and by advancing these insights to a more personalized approach to diagnosis and treatment.

1.1.2 Genetics and pleiotropy of schizophrenia

While the exact causes of schizophrenia have not been clearly understood, it is thought to arise from a complex interplay of genetic predisposition and exposure to environmental risk factors. Environmental risk factors, such as pregnancy and birth complications, childhood trauma, migration, social isolation, substance abuse, have been associated with an increased susceptibility for schizophrenia (Stilo and Murray 2019). Individual environmental factor typically demonstrate a modest effect size (~2 fold increase in risk), and none is specific to schizophrenia. However, the cumulative effect of individual factors has been associated with relevant clinical effects. For example, patients exposed to 4 environmental risk factors demonstrate an earlier age of onset compared to those exposed to 3 environmental risk factors (Stepniak, Papiol et al. 2014). , An interesting approach to disentangle this complex risk pattern has been to identify biological alterations that may modulate the effects of environmental risk on schizophrenia symptoms. For example, the elevation of dopamine synthesis has been shown to be associated with migration (Egerton, Howes et al. 2017), childhood abuse (Oswald, Wand et al. 2014) and low parental care (Pruessner, Champagne et al. 2004). Exploring biological interactions with environmental predisposition may lead to the identification of modifiable factors that could aid in the improvement of the clinical management of schizophrenia, or the development of novel preventative approaches.

Genetics

The early genetic studies of schizophrenia, dating back to the middle of the 20th century, were primarily family history and twin studies (Kallmann 1946). These studies investigated the association of phenotypes from a set of blood-related samples and indirectly inferred the properties of the phenotype-associated genetic risk factors. Notably, the accumulated evidence from twin studies supports an 81% heritability of schizophrenia (Sullivan, Kendler et al. 2003). In the 90s, advanced technology allowed the determination of DNA sequences. A milestone in this period was the completion of the human genome project in 2003, which successfully mapped DNA sequences to 20,500 human genes, and started a new era for understanding nature from a molecular level. For genetics research of schizophrenia, genome-wide association studies (GWAS) became a central resource for providing the downstream investigations with evidence on risk-associated regions on the genome. GWAS determine the risk association of a given variant by regressing the diagnosis on the genotype of the variant in a large group of patients and controls. With the ever-increasing sample sizes, GWAS has profoundly impacted the understanding of schizophrenia genetics. In the course of two decades, an increasing number of genome-wide significant SNPs, and a more accurate estimation of genetic effects were determined. For example, a recent work (Ripke, Walters et al. 2020) identified 329 genome-wide significant SNPs and 23% heritability from a large sample comprising 69,369 schizophrenia patients and 236,642 controls.

Despite these successes, the genome-wide significant SNPs explain individually, as well as in combination, only a small proportion of schizophrenia's heritable variance. This might be due to the polygenic inheritance of schizophrenia (International Schizophrenia, Purcell et al. 2009), meaning that the heritability is attributable to a large number of weakly associated SNPs. The Polygenic Risk Scores (PRS) integrating effects over a large number of common variants currently predict the diagnosis with a modest accuracy (AUC=0.71) (Ripke, Walters et al. 2020).

Pleiotropy

Pleiotropy refers to a phenomenon that an identical genetic factor influences multiple traits. In the human genome, over 4.6% of SNPs and 17% of genes (Sivakumaran, Agakov et al. 2011) demonstrate a pleiotropic effect. In the context of mental illness, this effect appears to be even more substantial, where e.g., a 40% genetic correlation has been found between major depressive and bipolar disorder (Cross-Disorder Group of the Psychiatric Genomics, Lee et al. 2013, Lee, Ripke et al. 2013). For schizophrenia, pleiotropic effects have also been found across numerous mental and somatic illnesses. For example, schizophrenia shows a significant genetic correlation with bipolar disorder (68%) (Lee, Ripke et al. 2013), major depressive disorder (40%) (Bulik-Sullivan, Finucane et al. 2015) and amyotrophic lateral sclerosis (14.3%) (McLaughlin, Schijven et al. 2017). A detailed description of schizophrenia pleiotropy can be found elsewhere (Bulik-Sullivan, Finucane et al. 2015, Zheng, Erzurumluoglu et al. 2017, Watanabe, Stringer et al. 2019).

The accurate analysis of pleiotropy relies on large-scale GWAS data and sophisticated analysis methodologies. Linear mixed models (Lee, Yang et al. 2012) and cross-trait LD score regression (Bulik-Sullivan, Loh et al. 2015) are two major tools for this analyses that require different types of input data. Linear mixed models require the individual genotype data as input to estimate the genetic correlation. In contrast, cross-trait LD score regression only require the summary statistics data and are robust with respect to shared confounders. Due to data privacy protection concerns related to the sharing of genotype data and the ubiquitous presence of shared confounders, these benefits lead to the wide use of cross-trait LD score regression. However, these benefits also come at the cost of a high variance of the LD score regression model, requiring more samples to achieve the desired precision. A complete

review of methodologies for analyzing the pleiotropy of mental illnesses can be found in (Cao and Schwarz 2019).

1.2 Multi-modal data integration using machine learning

Biological research of schizophrenia is increasingly moving towards the integrative exploration of multiple data modalities, in order to integrate the diverse, individually weak biological changes and characterize more comprehensively the affected biological systems. Due to the strong heritability of schizophrenia, a critical element of such integrative efforts is the consideration of genetic predisposition. Large-scale GWAS studies have identified over 300 independent, genome-wide significant loci associated with the illness (Ripke, Walters et al. 2020). Functional genomics analyses have linked these genetic susceptibility effects encoded in common genetic variants to alterations in synaptic function, as well as histone and immune-system related effects(Fromer, Roussos et al. 2016) (Schwarz, Izmailov et al. 2016) (O'Dushlaine, Rossin et al. 2015). The integrative analysis of such genetic susceptibility effects with other data modalities, including neuroimaging, is a promising avenue to obtain deeper insight into their functional consequences. This thesis explores such integrative analysis using advanced multi-task machine learning, focusing on the genetic association, gene expression, as well as neuroimaging data. The following sections provide a high-level overview of these data modalities in the context of schizophrenia, and are followed by an in-depth description of the machine learning approaches developed and deployed as part of this thesis.

1.2.1 Genomics

Genomics studies explore an organism's DNA sequence by characterizing its structure, function, evolution, mapping and editing. DNA is described by an ordered sequence of nucleic acids. The first generation of sequencing technology was the Sanger method (Sanger, Nicklen et al. 1977) which utilized the so-called “chain termination method” to trace the molecules. In 1987, the Sanger method (Hood, Hunkapiller et al. 1987) was automated, indicating the maturation of the first-generation sequencing technique. The currently used technique is the so-called next-generation sequencing (NGS) technology. The key difference is the high-throughput sequencing volume, which allows hundreds of millions of DNA molecules to be measured simultaneously, due to the massive parallelization of a large number of reactions. The most important milestone of NGS was the success of the human genome project, which produced the first draft of the human genome (Lander, Linton et al. 2001). With the development of NGS and related technologies, the economic cost for the whole-genome sequencing of an individual kept decreasing rapidly over the years, leading to an increase in the availability of large number of human genomic datasets.

These datasets, allow exploring the relationship between the genetic predisposition and clinical phenotypes, and locating the specific loci that contribute to this effect. This promoted the emergence of GWAS. GWAS is a powerful method to evaluate the association between a given genotyped marker and the phenotype. In a recent decade, GWAS achieved great success in exploring human genetics, with around 5,000 GWAS results covering over 400 unique studies and 3,000 unique traits (Watanabe, Stringer et al. 2019). Two types of phenotype-associated variants exist: rare variants of large effect and common variants of small effects. Characterizing risk architectures using GWAS is particularly effective for the former type, whereas the interpretation of findings related to common variants is challenging due to the complex polygenic structure of disorders such as schizophrenia. For both types, the power of GWAS is influenced by the heritability of the phenotype, which is commonly quantified as the

phenotypic variance explained by the genetic markers in a population. Technically, obtaining a deeper understanding of the explained variance depends on two factors: 1) the frequency of the risk variant in the population and 2) the effect size of the risk variant compared to that of the alternative variant. One drawback of GWAS is the inability to account for the linkage disequilibrium (LD) structure. Causative markers can be strongly associated with the many non-causative markers in the same LD block, which commonly induces significant, but mechanistically-uninformative GWAS associations. Another drawback is the frequently occurring missingness of data and low quality of measurements of the SNP array. Excluding either individuals or SNPs with high missingness can reduce the power of GWAS. However, this can be mitigated by data imputation techniques (Li, Willer et al. 2009), which complete the missing genotype according to the available haplotypes of other individuals in a sample. The underlying mathematical model facilitating this imputation is a Hidden Markov Model (HMM).

Although it is difficult to directly identify the significant risk loci for common-variants-associations, the individual's polygenetic risk can be determined by summarizing the weak risk contributions over a large number of common variants based on the GWAS results, as performed in polygenic risk score (PRS) analysis (Choi, Mak et al. 2020). While several methods exist to determined PRS, the standard approach is the so-called "C+T" (clumping + thresholding) (Choi, Mak et al. 2020). Clumping is used to control the biases caused by the LD effect because the SNP-SNP correlations are not uniformly distributed across the genome. Clumping selects independent SNPs using a statistical correlation metric as well as the pairwise physical distance on the genome (Prive, Vilhjalmsson et al. 2019). P-value thresholding aims at retaining the high-risk SNPs from the clumped SNP set. For this, a maximum P-value is set as the threshold to remove low-risk SNPs. However, this threshold is difficult to select and specific for a phenotype and its genetic architecture. Validation on an independent cohort would be commonly utilized to select the optimal threshold. The final score is obtained by multiplying the effect size of these high-risk SNPs and the genotypes of a given individual, and building the sum of these values. PRS analysis is particularly important for schizophrenia research due to the polygenic nature of the illness. A milestone work was the first study determining PRS on schizophrenia (International Schizophrenia, Purcell et al. 2009), which explained over 3% of the heritable variance. More importantly, the score was specific to schizophrenia compared to other psychotic and non-psychotic disorders. The work demonstrated the potential utility of PRS in translational schizophrenia research. With the increasing accumulation of genetic data, the most recent schizophrenia PRS explains a substantially higher portion of variance (Ripke, Walters et al. 2020).

A particularly interesting line of research that originated from GWAS was eQTL (expression quantitative trait loci) analysis. eQTL refers to loci that explain a significant amount of the variation of gene expression in a specific tissue. This analysis, instead of locating the risk loci that impact on the final phenotype (e.g., schizophrenia), aims at identifying variants that affect gene expression and regulation. The corresponding association test is performed between the expression level of a given gene and the genotype of each SNP, followed by multiple hypothesis testing correction. In schizophrenia research, eQTL analysis was able to shed light on the functional effect of the identified susceptibility variants. A study (Bhalala, Nath et al. 2018) identified over 2000 cis-eQTL related to 40 genes, including 11 non-coding RNAs. Interestingly, these eQTLs were overrepresented in brain tissue compared to blood, pointing to a brain-specific effect of genetic susceptibility on gene regulation.

1.2.2 Transcriptomics

Transcriptomics studies aim at characterizing a given organism's entire transcriptome – a snapshot of all RNA transcripts in a cell. Transcriptomics studies started in the early 1990s, and two key

technologies have been developed for analysis at a genome-wide scale: microarray and RNA-seq. The early microarray technology utilizes a “library” of transcripts, against which the transcripts in a given sample are matched to facilitate quantitation. One of the drawbacks of this method is that RNA molecules that have no match in the library cannot be quantified. In contrast, the more recently developed RNA-seq allows the complete sequencing of an entire transcriptome. Compared to microarray, RNA-seq can measure more transcriptome information, including information on splice variants and non-coding transcripts (Rao, Van Vleet et al. 2018).

In the context of genetic risk, it is notable that schizophrenia-associated common variants have been found enriched in genes expressed in the brain (Schizophrenia Working Group of the Psychiatric Genomics 2014) that aggregated into pathways related to synaptic functions, histone and immune systems (O'Dushlaine, Rossin et al. 2015). Since most risk loci are located outside of coding exons, it is assumed that the susceptibility of schizophrenia-associated variants is mediated via the regulation of gene expression. A large number of studies have explored gene expression differences in patients with schizophrenia, most frequently using case-control study designs. Compared to the genetic association studies, differences in gene expression typically show larger effect sizes but are more easily confounded by a large number of potential factors, including measurement batch and medication effects. This makes the validation of identified gene expression differences mandatory, in particular when algorithms integrate a large number of changes observed for different genes. Several studies have performed such independent validation [e.g. (Chen, Cao et al. 2020)] and have also been successful in identifying comparable signatures in other conditions, such as type 2 diabetes (Cao, Chen et al. 2017). In a more complex, multi-modal data analysis, transcriptomic analysis is able to play a particularly critical role as the intermediate phenotype to connect other data modalities. A notable example is the integrative analysis of genetic association, gene expression, and ontological annotation data in deep neural networks that substantially improved the ability to predict schizophrenia diagnosis compared to the conventional approach (Wang, Liu et al. 2018).

Transcriptomic analyses are frequently interpreted in the context of the biological processes within they take place. Assigning genes to biological processes relies on gene ontological information that reflects a hierarchical representation of a biological system at a molecular level. Two of the most well-known databases are the gene ontology (GO) and the kyoto encyclopedia of genes and genomes (KEGG). Both assign functionally-related genes into sets. The difference between the two databases is that GO aims at building an entire tree ranging from gene categories to the organism, whereas KEGG characterizes the dependency between genes within a given pathway, including their activity and the respective functions of the participating genes.

These gene ontology databases support diverse analyses in molecular studies. First, the functional set enrichment analysis (e.g., gene set enrichment analysis) is commonly used to interpret the biological function of identified transcriptomic changes (e.g., affected biological pathways) (Wu, Hu et al. 2021). A widely used computational tool for this task is further described in (Wu, Hu et al. 2021). These databases are furthermore increasingly used to support machine learning analysis. Pathway-annotation allows stratifying high-dimensional data into biologically meaningful, smaller datasets, making the training of machine learning models easier. Previous work has used this approach on genome-wide DNA methylation, as well as GWAS data, and then aggregated pathway-specific algorithms into a systems-level classifier, in order to test associations with brain function (Chen, Zang et al. 2020).

1.2.3 Neuroimaging

Advances in imaging acquisition techniques have allowed neuroscientists to observe the structure and function of the brain in living individuals. The most widely-used techniques for neuroimaging are sMRI (structural magnetic resonance imaging) and fMRI (functional magnetic resonance imaging) due to their low invasiveness and the lack of radiation exposure (Xue, Chen et al. 2010). sMRI has been used to quantify the brain's anatomy through images with high contrast between gray and white matter based on differences of water content in the respective tissues. To analyze sMRI scans, researchers commonly use voxel-based morphometry as a computational tool to determine localized differences at the voxel level. sMRI supports several analyses, including the volumetric comparison of brain tissue, the assessment of the degree of cortical folding, and the exploration of the cortical gyrification pattern (Gifford, McCutcheon et al. 2020). fMRI is used to measure neural activity by identifying the changes in blood oxygenation because an increased activity in a given location is associated with increased energy consumption. The underlying principle is the differential magnetic properties of oxygenated and deoxygenated blood. This technique is commonly used to test the neural responses of subjects during a set of well-designed tasks or in a resting state.

Neuroimaging has played a critical role for the characterization of functionally-relevant biological mechanisms of schizophrenia (Abi-Dargham and Horga 2016). With a complex genetic architecture, genetic susceptibility for schizophrenia is carried by a large number of risk variants with small effect sizes. This implies that no individual genes (or environmental factors) are fundamental to the disease process for most schizophrenia patients. Neuroimaging has provided a tool for determination of “intermediate phenotypes” to study schizophrenia, that are thought to be of fundamental relevance to the clinical phenotype, but closer to the underlying biology that the clinical manifestation (Meyer-Lindenberg and Weinberger 2006). As a consequence, the effect sizes of, e.g. associations between risk variants and schizophrenia-relevant brain function, are expected to be larger. There is an extensive literature describing brain-structural, -functional and molecular differences in schizophrenia, e.g. dopamine hyperactivity (Hietala and Syvälahti 1996), N-Methyl-D-aspartate receptor alterations(Olney and Farber 1995), hippocampal hyperactivity (Lieberman, Girgis et al. 2018) and immune dysregulation (Dalmau, Gleichman et al. 2008). These and other findings gave rise to the “imaging genetics” field, aiming to explore the genetic underpinning of the effects.

A large number of studies using conventional, “univariate” statistical analysis were performed for the analysis of such data. For example, the ENIGMA consortium has investigated the genetic basis for structural alterations (Medland, Grasby et al. 2020), and characterized the genetic pleiotropy of mental disorders in relation to subcortical brain volumes (Campbell, Jahanshad et al. 2021), or the effects of CNVs on cerebral and cognitive alterations (Sonderby, van der Meer et al. 2021). Meanwhile, to integrate the risk of genetic features as well as omics-derived data, machine learning is increasingly used in the psychiatric field. For example, the epigenetic signature of schizophrenia was learnt using machine learning and associated with the schizophrenia-relevant brain function (Chen, Zang et al. 2020).

1.2.4 Dimensional reconstruction of mental illness using machine learning

Biological research has demonstrated that alterations found in patients with schizophrenia most frequently show small effect sizes and are widely distributed across different data modalities. Therefore, in order to integrate such changes and obtain deeper insight into biological mechanisms affected in schizophrenia, advanced computational tools are needed that can extract illness-associated signals in high-dimensional, multimodal data.

Machine Learning (ML) offers a well-established framework for integrating numerous phenotype-associated alterations in a given high-dimensional data modality, in order to maximize the ability to predict a given phenotype. A large body of literature exists that has used machine learning for such tasks in biomedicine and more specifically, in biological psychiatry, and this has been reviewed recently in (Cao and Schwarz 2020). While ML has overwhelmingly focused on learning biological signatures from individual data modalities, the simultaneous capture of biological signatures from multiple data modalities may be advantageous for characterizing complex disease mechanisms. A particularly promising approach for such integrative analysis is the so-called ”multi-task learning” (MTL). MTL is an advanced ML technique for the simultaneous learning from multiple related datasets. It has been successfully used in numerous data-intensive fields, including biomedical informatics (Li, Wang et al. 2016), natural language processing (Li, Liu et al. 2020), image processing and computer vision (Zhang, Luo et al. 2014), as well as web-based applications (Chapelle, Shivaswamy et al. 2010). In psychiatry, MTL has been applied to integrate the heterogeneous gene expression cohorts of schizophrenia to identify a predictable, stable and consistent signature (Cao, Meyer-Lindenberg et al. 2018).
For molecular studies of psychiatric illnesses, MTL demonstrated utility as a dimensional approach for disentangling shared and specific biological alterations in multi-modal data cohorts(Han Cao and Schwarz).

1.3 Regularization based machine learning

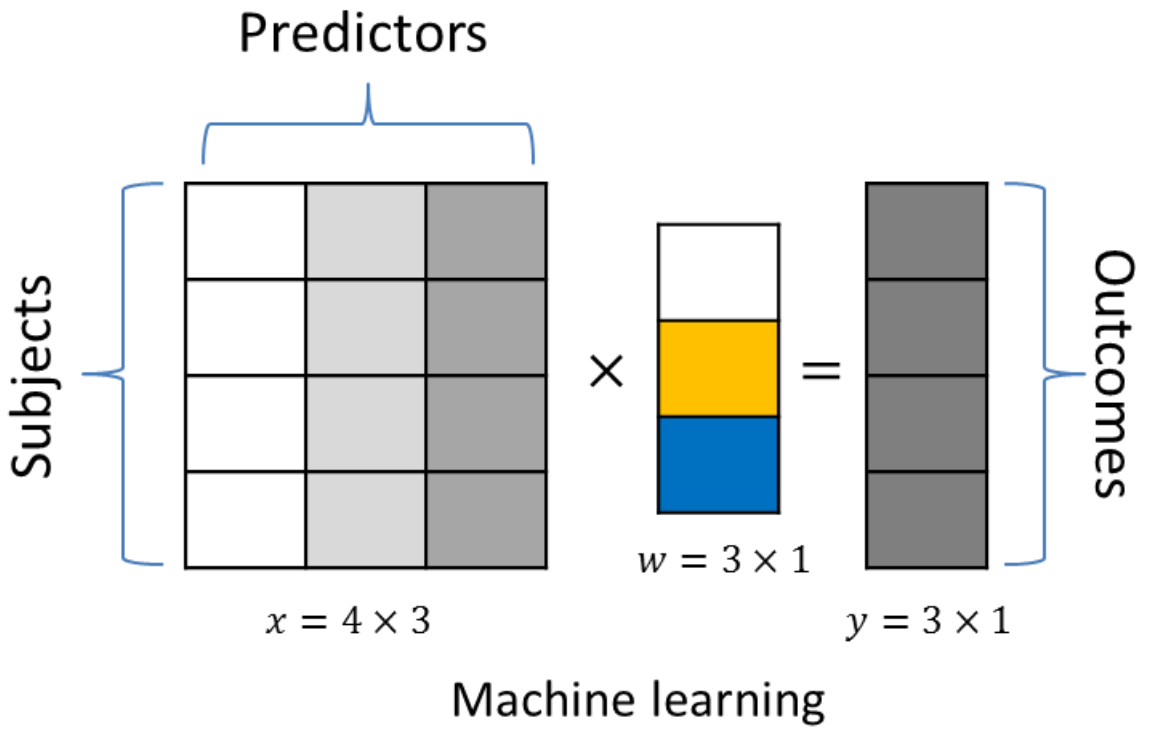


Figure 1. Example of a linear ML model. The outcome y is predicted by the multiplication of feature matrix (x) and model (or coefficients) w. The learning procedure was to identify w.

High dimensionality is a common characteristic of most molecular and neuroimaging data modalities, and constitutes a significant challenge for ML algorithms. The so-called “curse of dimensionality”

describes the typical scenario that the predictive accuracy of machine learning algorithms when tested on unseen datasets decreases as the dimensionality of the training data increases. This is because ML algorithms easily overfit the model on high-dimensional training data and do not generalize well on unseen data. The bias-variance tradeoff theory (Hastie, Tibshirani et al. 2009) showed that the model complexity is proportional to the chance of overfitting. Therefore compressing the model complexity has been a widely adopted strategy in high-dimensional data problems, e.g., in the form of a linear model with the utilization of a regularization technique. The bias-variance tradeoff theory shows that linear models have the lowest bias compared to more complex models (i.e., random forest). Figure 1 shows a linear ML model for outcome prediction. A coefficient vector w is learnt for predicting the outcome and interpreted given the context of the application. However, the linear model has a high variance leading to a limited predictive power in an actual analysis. This situation becomes more severe in a high-dimensional setting due to the increased chance of overfitting. To mitigate this issue, regularization techniques were developed, which greatly reduced the model variance by introducing constructive prior information (a slight bias), in order to guide the optimization. During the evolution of regularization techniques over several decades, it has been explored as a tool of domain-knowledge integration by incorporating a well-defined prior information.

$$\min_{\mathrm{w}} \mathcal{L}(\mathrm{w}|\mathrm{x},\mathrm{y}) + \lambda\Omega(\mathrm{w}) \quad (1)$$

As shown in formulation (1), $\mathcal{L}(\mathrm{w})$ is a data fitting term (also called the "loss function"), a major factor influencing the determination of the coefficient solution. The machine learning model (i.e., coefficients w) is obtained by minimizing this loss function given the data. A specific loss function is associated with a specific prediction task (i.e., the least-square loss is commonly used in regression tasks).
The function $\Omega(\mathrm{w})$ is a regularization term and frequently called the "penalty". This function aids in identifying a generalizable and interpretable solution by penalizing the unwanted characteristics of the coefficients. From the perspective of the penalty, regularization methods can be classified into two categories: I) penalization on the coefficients and II) penalization on the difference between the coefficients. Here, we explain each category in the context of biomedical studies.

1.3.1 Penalization on coefficients

In this category, the magnitudes of coefficients are penalized, leading to a sparse (or near-sparse) model – many coefficients are 0 (or near to 0). Examples for this class of methods are the Lasso ($\Omega(\mathrm{w}) = \lambda||\mathrm{w}||_1$), ridge regression ($\Omega(\mathrm{w}) = \lambda||\mathrm{w}||_2^2$) and elastic net ($\Omega(\mathrm{w}) = \lambda(\alpha||\mathrm{w}||_1 + \frac{(1-\alpha)}{2}||\mathrm{w}||_2^2$). Lasso assumes a sparse structure of the coefficients (i.e., many coefficients are 0) and applies the l_1-norm to achieve this aim. This method works well for high-dimensional data applications because outcome-irrelevant features commonly existed in such data. One study (Kohannim, Hibar et al. 2012) applied the Lasso to explore the genetic underpinnings of brain structure. However, as pointed out in one study (Zou and Hastie 2005), when there is a strong correlation structure among features, Lasso may select among such correlated features at random, leading to a potential loss of information and difficulties in interpreting the identified biological patterns. This situation is quite common in molecular studies where e.g. genetic co-expression and linkage disequilibrium (LD) cause a strong correlation structure between features. To capture this correlation structure in ML applications, the 'elastic net' was introduced based on the Lasso by adding an extra penalty (l_2-norm) term, in order to select sets of correlated predictors. For the ridge regression, only the l_2-norm is applied to penalize the coefficients and unimportant coefficients are shrunken towards 0.

Besides the penalization of individual features, another research line explored penalizing groups of features. This approach is meaningful in biological applications since biological features can be frequently grouped according to ontological annotations (e.g., genes are grouped into pathways), which assists in biological interpretation when changes in higher-level biological function can be associated with a given outcome, compared to those in individual genes. One regularization approach for this aim is called the "group Lasso" with the form $\Omega(w) = \lambda \sum_{g=1}^{G} ||w_{I_g}||_2$ where G represents a set of groups. This method has been used for tumor classification (Huo, Xin et al. 2020).
An extension to the group lasso assumes the presents of sparsity within a given group, facilitating feature selection at the feature- as well as the group-level. This assumption is justified in numerous biological applications, for example when only some genes of a gene group (e.g. pathway) are associated with a given outcome. The regularization derived from this assumption is called the "sparse group Lasso" with penalty term $\Omega(w) = \lambda_1||w||_1 + \lambda_2 \sum_{g=1}^{G} ||w_{I_g}||_2$ (Simon, Friedman et al. 2013). One successful application of this approach was the unsupervised analysis across omics modalities (Lin, Zhang et al. 2013).

1.3.2 Penalization on the difference of coefficients

Instead of the penalization on the coefficients, another research line has explored penalizing the difference between coefficients. This class of regularization techniques is suitable for modeling the relationship between features within an ML framework, e.g., the sequential order or network structure of features. In molecular studies, such feature relationships are frequently encountered, e.g., in the form of LD structure and genetic co-expression networks.
The 'fused Lasso' (Tibshirani, Saunders et al. 2005) ($\Omega(\mathrm{w}) = \lambda_1||\mathrm{w}||_1 + \lambda_2 \sum_{\mathrm{i}=1}^{\mathrm{p}} |\mathrm{w_i} - \mathrm{w_{i=1}}|$) forces the sequential order of the features and encourages the smoothness over the sequence of coefficients. This strategy has been applied in GWAS studies to account for LD effects (Liu, Wang et al. 2013, Yang, Liu et al. 2016). Alternatively, the 'network-based regularization' (i.e. $\Omega(\mathrm{w}) = \lambda_1||\mathrm{w}||_1 + \lambda_2 \mathrm{w^T L w}$, where L is the graph Laplacian) incorporates a network over features into the ML framework by encouraging coefficient similarity of features connected in the network. Such methods have been repeatedly applied in molecular biology, for example to incorporate PPI network (Wu, Wang et al. 2015), or co-expression network (Li and Li 2008).
The above methods naively assume that the coefficients connected via the network have the same signs (or association direction), which is unlikely in real biomedical applications. For example, it is quite common that two genes are inversely regulated by a third factor such that the expression values of these two genes are negatively associated. To address this, it can be useful to estimate the signs of the features prior to the ML stage. The studies (Li and Li 2008, Avey, Mohanty et al. 2017) estimated such signs as the association direction between the features and the outcome. It has been found that using this method to incorporate multiple biological networks can significantly improved pathway analysis (Avey, Mohanty et al. 2017). Another research line for tackling this issue focused on designing a more flexible penalty. One study(Yang, Liu et al. 2016) proposed a penalty ($\Omega(\mathrm{w}) = \lambda_1||\mathrm{w}||_1 + \lambda_2 \sum_{\mathrm{i}=1}^{\mathrm{p}} ||\mathrm{w_i}| - |\mathrm{w_{i-1}}||$) to penalize the difference between the absolute values of the coefficients instead of the actual coefficients, reducing the impact of the signs of individual coefficients. However, this algorithm was not easy to solve due to its non-convex nature. A common approach was reformulating the penalty into the "DC" (difference of convex functions) form and solving by DC programming. Another strategy adopted a convex alternative formulation (Bondell and Reich 2008, Yang, Yuan et al. 2012), $\Omega(\mathrm{w}) = \lambda_1||\mathrm{w}||_1 + \lambda_2 \sum_{(\mathrm{i,j})\in \mathrm{E}} \max_{\mathrm{w_i}, \mathrm{w_j}} \{|\mathrm{w_i}|, |\mathrm{w_j}|\}$. This min-max term penalizes

the larger absolute coefficient within a given coefficient pair such that the pair's coefficients are penalized to 0 simultaneously if the features are both poor predictors.

Regularization Name	Type	Subtype	Math Form	Effect	Ref
Lasso	Machine Learning	Penalization on coefficients	$\Omega(w) = \lambda\|\|w\|\|_1$	Remove irrelevant predictors	(Tibshirani 1996)
Ridge Regression			$\Omega(w) = \lambda\|\|w\|\|_2^2$	Prevent overfitting	(Hoerl and Kennard 1970)
Elastic net			$\Omega(w) = \lambda(\alpha\|\|w\|\|_1 + \frac{(1-\alpha)}{2}\|\|w\|\|_2^2$	Select important predictors with grouping effect	(Zou and Hastie 2005)
Group Lasso			$\Omega(w) = \lambda\sum_{g=1}^{G}\|\|w_{I_g}\|\|_2$	Account for group structure over predictors	(Meier, Van De Geer et al. 2008)
Sparse Group Lasso			$\Omega(w) = \lambda_1\|\|w\|\|_1 + \lambda_2\sum_{g=1}^{G}\|\|w_{I_g}\|\|_2$	Account for sparse group structure over predictors	(Simon, Friedman et al. 2013)
Fused Lasso		Penalization on the difference between coefficients	$\Omega(w) = \lambda_1\|\|w\|\|_1 + \lambda_2\sum_{i=1}^{p}\|w_i - w_{i-1}\|$	Account for ordered predictors	(Tibshirani, Saunders et al. 2005)
Absolute Fused Lasso			$\Omega(w) = \lambda_1\|\|w\|\|_1 + \lambda_2\sum_{i=1}^{p}\|\|w_i\| - \|w_{i-1}\|\|$	Account for ordered predictors (only magnitudes matter)	(Yang, Liu et al. 2016)
Network based Regularization			$\Omega(w) = \lambda_1\|\|w\|\|_1 + \lambda_2 w^T L w$ $\Omega(w) = \lambda_1\|\|w\|\|_1 + \lambda_2\sum_{(i,j)\in E}\max_{w_i,w_j}\{\|w_i\|,\|w_j\|\}$	Incorporate network structure over predictors	(Bondell and Reich 2008, Yang, Yuan et al. 2012)
Joint Feature Selection	Multi-Task Learning	MTL with structural regularization	$\Omega(W) = \lambda\|\|W\|\|_{2,1}$	Select predictors important to all tasks simultaneously	(Argyriou, Evgeniou et al. 2007)
Trace-norm model			$\Omega(W) = \lambda\|\|W\|\|_*$	Find the low-rank structure of the models	(Ji and Ye 2009)
Mean-regularized model		MTL incorporating pairwise task similarity	$\Omega(w_1, \dots w_i, \dots w_t) = \lambda\sum_{i=1}^{t}\|\|w_i - \frac{1}{t}\sum_{j=1}^{t} w_j\|\|_2^2$	Identify the mean model as the latent model behind all tasks	(Evgeniou and Pontil 2004)
Temporal Smoothness Prior			$\Omega(w_1, \dots w_i, \dots w_t) = \lambda_1\sum_{i=1}^{t-1}\|\|w_i - w_{i+1}\|\|_2^2 + \lambda_2\|\|W\|\|_2^2$	Incorporate temporal order among tasks to	(Zhou, Yuan et al. 2011, Zhou, Liu

			$\Omega(w_1, \dots w_i, \dots w_t) = \lambda_1 \sum_{i=1}^{t-1} \|\|w_i - w_{i+1}\|\|_2^2 + \lambda_2 \|\|W\|\|_{2,1}$ $\Omega(w_1, \dots w_i, \dots w_t) = \lambda_1 \sum_{i=1}^{t-1} \|\|w_i - w_{i+1}\|\|_1 + \lambda_2 \|\|W\|\|_{2,1} + \lambda_3 \|\|W\|\|_1$	predict the disease progression	et al. 2012)
Multi-task relationship learning			$\Omega(W, S) = \frac{\lambda_1}{2} tr(WW^T) + \frac{\lambda_2}{2} tr(WS^{-1}W^T)$ s.t. $\{S \succcurlyeq 0, tr(S) = 1\}$	Learn task-relationship	(Zhang and Yeung 2012, Liu, Pan et al. 2017)
Convex Clustered MTL			$\Omega(W, M) = \lambda\alpha\eta(1+\eta) tr(W(\eta I + M)^{-1}W^T)$ s.t. $\{tr(M) = k, M \preccurlyeq 1, M \in S_+^m\}$	Incorporate clustered structure between tasks	(Jacob, Vert et al. 2008, Zhou, Chen et al. 2011)
MTL with dirty model		Accounting for biological variations	$\Omega(W) = \lambda_1 \|\|P\|\|_{\infty,1} + \lambda_2 \|\|Q\|\|_1$ s.t. $W = P + Q$	Identify the shared and task-specific predictors simultaneously	(Jalali, Sanghavi et al. 2010)
Robust MTL			$\Omega(W) = \|\|P\|\|_{2,1} + \|\|Q^T\|\|_{2,1}$ s.t. $W = P + Q$	Detect the outlier tasks	(Gong, Ye et al. 2012)
Multilinear MTL	High-order MTL		$\Omega(\mathcal{W}) = \lambda \|\|\mathcal{W}\|\|_*, \mathcal{W} = p \times t_1 \times t_2$	Model the complex task-relationships	(Romera-Paredes, Aung et al. 2013)
Multi-task predictor interaction learning			$\Omega(\mathcal{Q}) = \lambda \sum_{i=1}^{p} \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{t} (\mathcal{Q}_{ijk}^2 + \mathcal{Q}_{jk}^2)}$ $\Omega(\mathcal{Q}) = \lambda \sum_{i,j} \sqrt{\sum_{k=1}^{t} \mathcal{Q}_{ijk}^2}$ $\Omega(\mathcal{Q}) = \lambda \|\|\mathcal{Q}\|\|_*, \mathcal{Q} = p \times p \times t$	Identify the simple and representative structure of predictor interaction pattern across tasks	(Lin, Xu et al. 2016)

Table 1. Algorithms for regularization-based ML and cross-task regularization-based MTL .

1.4 Cross-task regularization based multi-task learning

MTL is an ML paradigm that simultaneously learns from multiple datasets while utilizing task-relatedness to improve the model's generalizability. It has numerous interesting applications in molecular biology, e.g., illness comorbidity analysis, multi-omics analysis, and multiple outcome prediction. Different MTL algorithms adopt various strategies to transfer information among tasks, e.g., multi-task Gaussian process transfers information via the covariance structure; multi-task deep network shares the hidden layers directly among tasks. In high-dimensional data problems, a common approach is knowledge transfer via cross-task regularization.

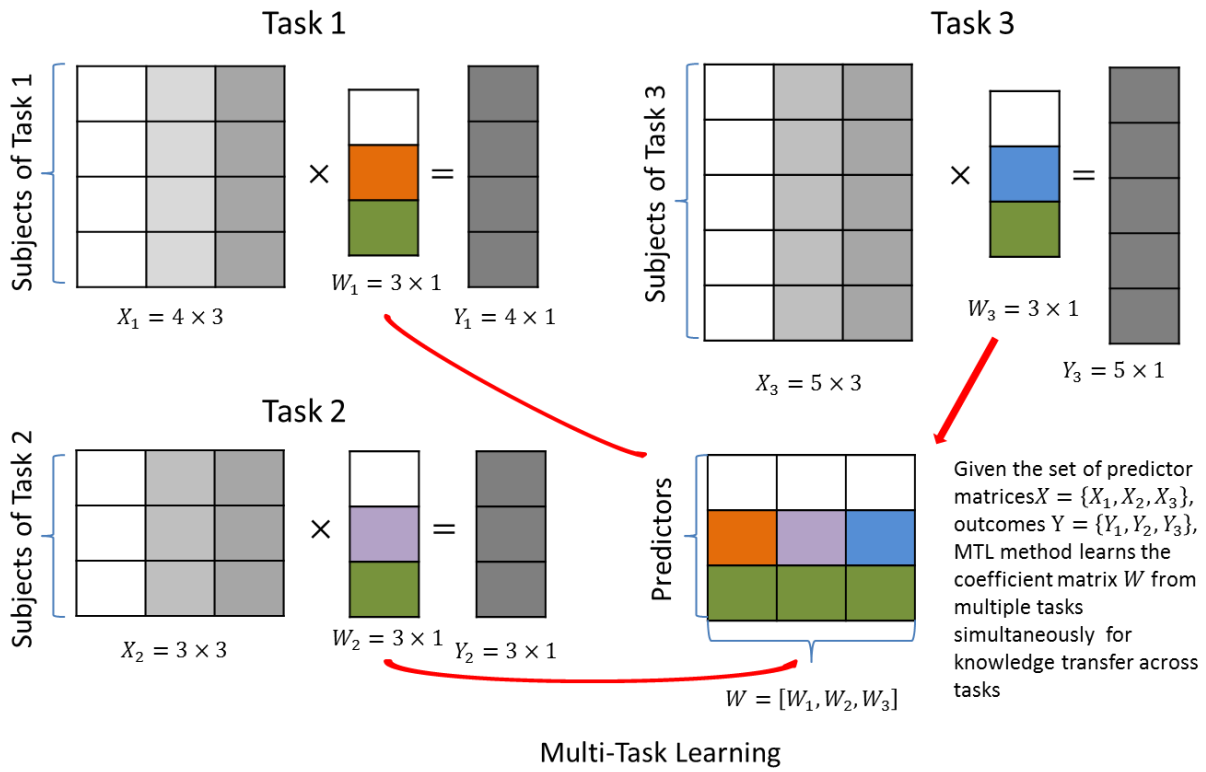


Figure 2. Graphical illustration of multi-task learning with joint predictor selection. The identical predictor set is shared among three different tasks. The aim is to identify a predictable model with the selected shared predictors.

$$\min_{w_1,\dots w_i,\dots w_t} \sum_{i=1}^{t} \mathcal{L}(w_i | X_i, Y_i) + \lambda \Omega(w_1, \dots w_i, \dots w_t) \quad (2)$$

The framework of cross-task regularization-based MTL can be represented as formulation (2). The regularization function $\Omega(w_1, \dots w_i, \dots w_t)$ takes the coefficient vectors of all tasks as input and outputs a score describing the departure of the task-relatedness (of the current model) to that of the assumed one. And $\sum_{i=1}^{t} \mathcal{L}(w_i | X_i, Y_i)$ was the sum of loss functions across all tasks, describing the model's fitness to the training data. Therefore, minimizing this composite objective leading to an interpretable and predictable model.

$\Omega(w_1, \dots w_i, \dots w_t)$ is commonly abbreviated as $\Omega(W)$, where $W = [w_1, \dots w_i, \dots w_t]$. The feature and task spaces are represented as W's row-wise and column-wise elements (see Figure 2). This simplified form illustrates an essential class of regularization approaches, which aim at identifying a simple representation of a matrix W (i.e., the rank of W is low). This kind of "unsupervised approach" doesn't assume a specific form of task-relatedness but learns from the data. Alternatively, another class of approaches explicitly assumes task-relatedness as a pair-wise similarity matrix. This method class aims at utilizing or learning this similarity matrix, in order to incorporate the task-relatedness.

1.4.1 MTL with structural regularization

Two commonly used "simplified" matrix forms within MTL are sparse or low-rank. For sparsity, a highly-cited work has described as "MTL with joint feature selection" (Argyriou, Evgeniou et al. 2007, Liu, Ji et al. 2009) ($\Omega(W) = \lambda \sum_i ||W_{i,}||_2 = \lambda ||W||_{2,1}$). In this formulation, features unimportant to all tasks are simultaneously filtered out. This approach has been used in cancer genetics to identify a gene pattern from multiple cancer treatments (Xu, Xue et al. 2011). The low-rank model constrains the

model searching in a low-dimensional space. A representative approach uses the trace-norm penalty $\Omega(\mathrm{W}) = \sum_{\mathrm{i}} |\lambda_{\mathrm{i}}^{(\mathrm{W})}| = \lambda||\mathrm{W}||_*$ which is a convex-relaxation of low-rank model with penalization on the l_1-norm of the singular values of W. Such method has been applied for predicting multiple drug responses (Yuan, Paskov et al. 2016). The results showed that the drug mechanism were reflective of the task-relatedness.

1.4.2 MTL incorporating pairwise task similarity

An early work in this class was the mean-regularized MTL (Evgeniou and Pontil 2004) ($\Omega(\mathrm{w}_1, \dots \mathrm{w}_{\mathrm{i}}, \dots \mathrm{w}_{\mathrm{t}}) = \lambda \sum_{\mathrm{i}=1}^{\mathrm{t}} ||\mathrm{w}_{\mathrm{i}} - \frac{1}{\mathrm{t}} \sum_{\mathrm{j}=1}^{\mathrm{t}} \mathrm{w}_{\mathrm{j}} ||_2^2$). This MTL algorithm assumes all tasks' models are derived from a single model, combined with the presence of a task-specific bias. Therefore, the regularization penalizes the difference between each model and the mean model. In psychiatry, this method has already been applied for identifying gene expression signatures of schizophrenia in multiple heterogeneous cohorts (Cao, Meyer-Lindenberg et al. 2018). This showed that the identified signature was more generalizable, robust, and consistent than those derived from other ML and MTL algorithms. Another interesting work was the prediction of Altzheimer's disease progression by incorporating the temporal smoothness as the task-relatedness in MTL (Zhou, Yuan et al. 2011, Zhou, Liu et al. 2013). The regularization took the form $\Omega(\mathrm{w}_1, \dots \mathrm{w}_{\mathrm{i}}, \dots \mathrm{w}_{\mathrm{t}}) = \lambda_1 \sum_{\mathrm{i}=1}^{\mathrm{t}-1} ||\mathrm{w}_{\mathrm{i}} - \mathrm{w}_{\mathrm{i}+1}||_2^2 + \lambda_2 ||\mathrm{W}||_2^2$, where the difference between two sequential models was penalized.
Instead of engineering a similarity matrix, some investigations have estimated it from the data. One study (Zhang and Yeung 2012) proposed a convex formulation for this aim: $\Omega(\mathrm{W}, \mathrm{S}) = \frac{\lambda_1}{2} \mathrm{tr}(\mathrm{W}\mathrm{W}^{\mathrm{T}}) + \frac{\lambda_2}{2} \mathrm{tr}(\mathrm{W}\mathrm{S}^{-1}\mathrm{W}^{\mathrm{T}})$, s.t. $\{\mathrm{S} \succcurlyeq 0, \mathrm{tr}(\mathrm{S}) = 1\}$, where S is the similarity matrix (Zhang and Yeung 2012). Minimizing $\frac{\lambda_2}{2} \mathrm{tr}(\mathrm{W}\mathrm{S}^{-1}\mathrm{W}^{\mathrm{T}})$ leads to a learned rank-1 similarity matrix. Another similar approach assumed a clustering structure between models (i.e., the columns of W). It combined the clustering and MTL loss to incorporate a clustering structure given the number of clusters. For example, one work took the regularization form (Zhou, Chen et al. 2011) ($\Omega(\mathrm{W}, \mathrm{M}) = \alpha\eta(1+\eta)\mathrm{tr}(\mathrm{W}(\eta\mathrm{I} + \mathrm{M})^{-1}\mathrm{W}^{\mathrm{T}})$, s.t. $\{\mathrm{tr}(\mathrm{M}) = \mathrm{k}, \mathrm{M} \preccurlyeq 1, \mathrm{M} \in \mathrm{S}_+^{\mathrm{m}}\}$, where $\mathrm{M} = \mathrm{t} \times \mathrm{t}$ is the similarity matrix). This formulation was derived from a convex relaxation form of k-means.

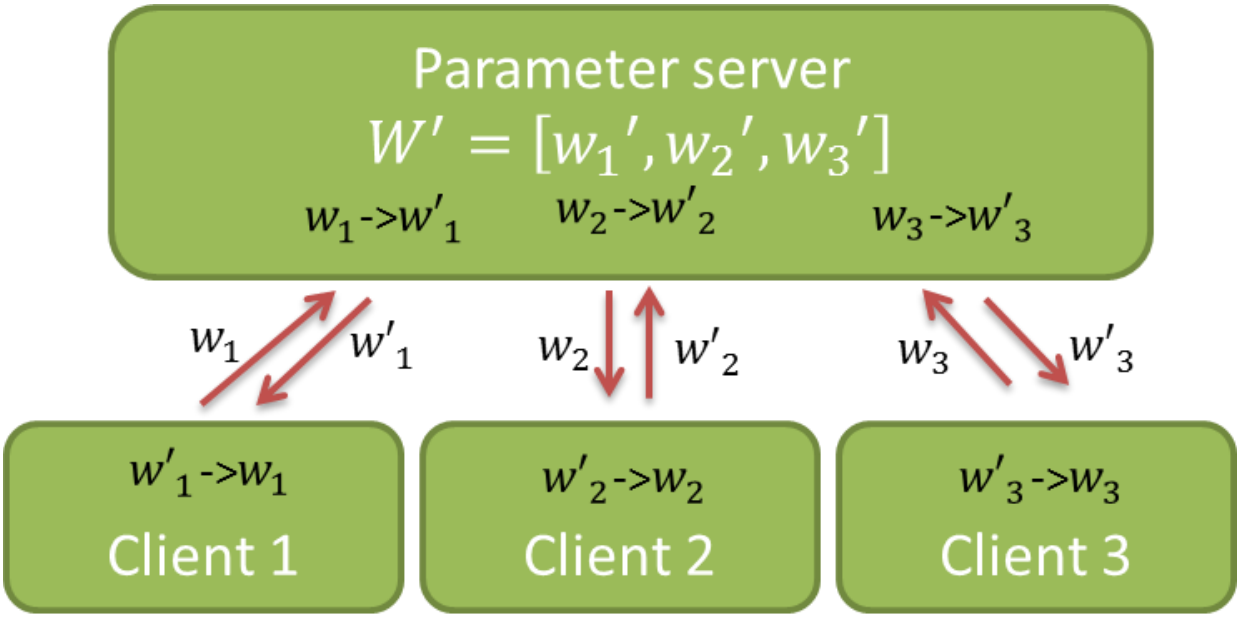
1.4.3 Federated MTL

Machine learning techniques have the potential to revolutionize biomedicine. For example, in the medical imaging field, deep neural networks have achieved significant pattern recognition success. The success of ML models relies on the availability of biomedical datasets at a large scale, but the aggregation of such data resources is challenging due to legal and logistic reasons limiting the ability to combine data stored at different locations into a single storage system. Therefore, there is an increasing need for computational solutions that facilitate the "federated" analysis of such resources without the need for physically combining them. Towards this, federated learning (Konečný, McMahan et al. 2015) was introduced in biomedicine to learn from geo-distributed data cohorts and protect the personally identifiable information. Compared to the traditional distributed learning system, the federated learning approach emphasized two points: 1) communication cost was the bottleneck to the algorithmic efficiency and 2) ML-oriented privacy protection was crucial and challenging. Several communication-efficient federated optimization methods have been developed to address the former issue (Xie, Baytas et al. 2017, Smith, Forte et al. 2018). Researchers were attempting to solve the latter issue in part from the perspective of statistics and in part from that of cryptography. For example,

differential privacy (Dwork 2006) was introduced as a statistical method quantifying the level of privacy leakage and controlling the leakage by adding noise. The method guaranteed a presumed attacker's prior and posterior view of an individual's data was not "significantly different". This method can prevent the leakage of sensitive from the final model but not the leakage from the individual device, e.g., the gradient information calculated from the mobile device of an individual can reveal sensitive information even with added noise.

It is straightforward to transform a standalone MTL algorithm into a federated MTL application. Because for most MTL methods, the calculation on sensitive data (e.g., the calculation of the gradient) can be separated from the algorithmic coordination. As shown in Figure 3, the federated learning optimization can be seen as an iterative method containing the calculation on the server, the clients and the message passing over the internet. First, the operations on sensitive data are performed on the client to improve model fitness. Second, the model is sent from the clients to a server over the internet. Third, the server performs the coordination for the knowledge transfer (e.g., cross-task regularization).

From a methodological perspective, compared to federated ML, federated MTL captures the cohort-level heterogeneity among geo-distributed cohorts (Smith, Chiang et al. 2017). This is essential in biomedicine because geo-distributed biomedical datasets are commonly heterogeneous due to the various procedures of sample recruitment, data preprocessing, and numerous other confounding effects that might bias the result.



$$\text{First operation: } w_i = w_i' - \frac{\partial \mathcal{L}(w_i')}{\partial w_i'}, \text{ performed on client machine}$$

$$\text{Second operation: } W_{(j)}' = \boldsymbol{sign}(W_{(j)}) \left(\left(1 - \frac{\lambda}{\|W_{(j)}\|_2} \right) |W_{(j)}| \right)_+,$$

performed on sverver, where (j) denotes the jth row, and
$W = [w_1, w_2, w_3]$

Figure 3. Schematic overview of federated MTL. The figure shows the computations on the server and the client, and the message passing over the internet. The model is MTL with joint feature selection (see Table 1 for details). This algorithm would keep running until the model converges.

2 STUDY 1: RMTL: AN R LIBRARY FOR MULTI-TASK LEARNING

2.1 Abstract

Motivation: Multi-task learning (MTL) is a machine learning technique for simultaneous learning of multiple related classification or regression tasks. Despite its increasing popularity, MTL algorithms are currently not easily available, creating a bottleneck for their application in biomedical research.
Results: We developed an efficient, easy-to-use R library for MTL (www.r-project.org) comprising 10 algorithms applicable for regression, classification, joint feature selection, task clustering, low-rank learning and incorporation of biological networks. We demonstrate the utility of the algorithms using simulated data.
Availability: The RMTL package is an open source R package and is freely available at https://github.com/transbioZI/RMTL. RMTL will also be available on cran.r-project.org

2.2 Introduction

Multi-task learning (MTL) is a machine learning technique that explores and exploits the relatedness across a set of different learning tasks. Since its inception (Caruana 1998), MTL has been used in numerous data-intensive research areas, including biomedical informatics (Xu, Pan et al. 2011, Widmer and Ratsch 2012, Zhou, Liu et al. 2013, Feriante 2015, Li, Wang et al. 2016, Yuan, Paskov et al. 2016), speech and natural language processing [i.e. (Wu, Valentini-Botinhao et al. 2015)], image processing and computer vision [i.e. (Wang, Zhang et al. 2009)], as well as web based applications [i.e. (Chapelle, Shivaswamy et al. 2010)].
A strong motivation to develop biomedical MTL applications stems from the necessity to integrate diverse data sources to explore the biological underpinning of complex illnesses, such as schizophrenia. Previous research has already shown that for such illnesses, integrative multi-omics open a new avenue for identification of etiological mechanisms, for example by taking into account genetic, expression and methylation data simultaneously [i.e. (Lin, Zhang et al. 2014)]. For such applications, multi-task learning offers the possibility to directly explore illness-related biological profiles that are linked across data modalities and therefore a new route toward the identification of biomarker signatures.
Previous implementations of MTL have focused on knowledge transfer via regularization(Zhou, Chen et al. 2011), Bayesian methods(Greenlaw, Szefer et al. 2017) or deep architectures(Yang and Hospedales 2016). Here, we developed the first R library for MTL, offering a comprehensive machine learning pipeline that covers several types of MLT algorithms and can be easily applied to high-dimensional data.

2.3 Methods

This package provides an automated, simple-to-use implementation of MTL, comprising 5 classification and 5 regression algorithms, which share knowledge across tasks according to different priors via regularization. All algorithms aim to minimize the same objective:

$$\min_{W} \sum_{i}^{t} \frac{1}{n_i} L(W_i|X_i, Y_i) + \Omega(W)$$

where $L(\circ)$ is the loss function (logistic loss for classification or least square loss for regression). X, Y are feature matrices and the corresponding responses, W is the coefficient matrix, and t is the number

of tasks. Accordingly, X_i, Y_i W_i and n_i refer to the data matrix, responses, model parameter vector and the number of subjects of task i, respectively. Note that W_i is the ith column of W. Knowledge transfer among tasks is achieved via a convex term $\Omega(W)$ that jointly modulates models according to specific functionalities. In this package, five common regularization techniques are implemented to suit different applications, i.e. sparse structure, joint feature selection, low-rank structure, network constraint for task relatedness and task clustering. Here, we refer to the above regularization strategies as MTL_Lasso, MTL_L21, MTL_Trace, MTL_Graph and MTL_CMTL, in the same sequence. These strategies can be broadly categorized into two classes: strategies for predictor selection (MTL_Lasso and MTL_L21) and strategies for task relatedness exploration (MTL_Graph, MTL_Trace and MTL_CMTL). While the former class explores sparse patterns are explored over the predictor space, the latter class exploits task relatedness based on additional assumptions. For all algorithms, we implemented a solver based on the accelerated gradient descent method(Nesterov 2012). To solve the non-smooth and convex regularization, the proximal operator(Parikh and Boyd 2014) was applied. Overall, the solver achieves a complexity of $O(1/k^2)$, which is optimal among first-order gradient methods. Further methodological details are shown in the **Supplementary Methods**.

2.4 Results

Predictive performance and model interpretability of the implemented algorithms were explored using simulated data. The simulated datasets were constructed by the ground truth model W, which is specified for a given prior (**Supplementary Figure 1)**. We compared the ground truth and the learnt model as an indicator of model interpretability. For predictive comparison, the primary baseline method was the conventional lasso, which reflects single task learning performance. We further applied MTL with lasso (MTL_Lasso), to explore the effect of inappropriate prior choice as a second baseline method.

2.4.1 Model interpretability

Supplementary Figure 1a shows the coefficient matrix of MTL_Lasso and MTL_L21 and demonstrates that the number of predictors identified by MTL_Lasso was approximately half the number of ground truth predictors. This may be due to the fact that highly correlated predictors exist in the high-dimensional space(Zou and Hastie 2005). As a consequence and similar to conventional Lasso, MTL_Lasso tended to select one among several correlated predictors. Despite this, 75% (precision) of selected predictors were ground truth predictors. For MTC_L21, the ground truth was highly sparse: only 40 out of 400 predictors were active predictors for all tasks. The simulation demonstrates that 39 of the predictors were successfully identified (sensitivity: 97.5%), with a precision of 72%. These results indicate that MTL algorithms could successfully identify ground truth predictors.
The relatedness of tasks was represented by pairwise correlation between models. **Supplementary Figure 1b** shows that all methods were able to capture correctly the pairwise relatedness compared to the ground truths. Particularly, MTL_Graph incorporated a strong network prior such that the "in-group" differences became zero. This may be because the network prior provided the most complete information about task relatedness among all priors.

2.4.2 Predictive performance

Supplementary Figure 2 indicates that conventional Lasso failed to yield accurate predictions on all simulated datasets except when using the l_{21} prior. Compared to this baseline, the MTL models

improved the accuracy by 18.7% on average. The MTL_Lasso incorporating an inappropriate prior achieved an average accuracy of 67% and was substantially inferior to MTL models with appropriate priors (average accuracy: 79.2%).

2.5 Conclusion

In this study, we developed an R library for multi-task learning comprising 10 algorithms incorporating 5 different priors. MTL models outperformed two baseline methods when applied on simulated data. High model-interpretability was observed in terms of predictor selection and task-relatedness compared to the respective ground truths.

2.6 Supplements

2.6.1 Supplementary methods

Multi-task learning implementation

This package provides an automated, simple-to-use implementation of MTL (classification and regression). The pipeline contains two core stages. First, a set of datasets is fed to (stratified) k-fold cross-validation (cv) for selection of model parameters. The output of this stage consists of the selected parameters, the minimum cv error and the corresponding plot. The training data and selected model parameters are then sent to the 'training stage'. The output of this second stage contains the trained model and a statistical summary. Using this model, predictions can be performed on independent datasets. To train sparse models, the warm-start technique(O'Brien 2016) is used generate the entire solution path along the parameter sequence.

As part of the library, we implemented 10 MTL algorithms (5 for classification and 5 for regression), which share knowledge across tasks according to different priors via regularization. All algorithms aim to minimize the same objective:

$$\min_{W} \sum_{i}^{t} \frac{1}{n_i} L(W_i | X_i, Y_i) + \Omega(W)$$

where $L(\circ)$ is the loss function (logistic loss for classification or least square loss for regression). X, Y are feature matrixes and the corresponding responses, W is the coefficient matrix, and t is the number of tasks. Accordingly, X_i, Y_i W_i and n_i refer to the data matrix, responses, model parameter vector and the number of subjects of task i, respectively. Note that W_i is the ith column of W.

Knowledge transfer among tasks is achieved via a convex term $\Omega(W)$ that jointly modulates models according to specific functionalities. In this package, five common regularization techniques are implemented to suit different applications, i.e. sparse structure ($\Omega(W) = ||W||_1$) (Tibshirani 2011), joint feature selection ($\Omega(W) = ||W||_{2,1}$) (Liu, Ji et al. 2009, Liu and Ye 2009), low-rank structure ($\Omega(W) = ||W||_*$)(Pong, Tseng et al. 2010), network constraint for task relatedness ($\Omega(W) = ||WG||_F^2$)(Widmer, Kloft et al. 2012) and task clustering (($\Omega(W) = \lambda_1 \eta(1+\eta) tr(W(\eta I + M)^{-1} W^T)$) (Jacob, Vert et al. 2008, Zhou, Chen et al. 2011). Here, we refer to the above regularization strategies as MTL_Lasso, MTL_L21, MTL_Trace, MTL_Graph and MTL_CMTL, in the same sequence.

These strategies can be broadly categorized into two classes: strategies for predictor selection (MTL_Lasso and MTL_L21) and strategies for task relatedness exploration (MTL_Graph, MTL_Trace and MTL_CMTL). For the former class, sparse patterns are explored over the predictor space and different types of information, i.e. the strength of penalization (MTL_Lasso) or the predictive pattern (MTL_L21), are shared between tasks. For the latter class, task relatedness is explored based on additional assumptions. MTL_Trace assumes that all models are spanned in a low-rank space, thus highly

correlated models can be obtained via strong penalization. MTL_Graph assumes that all models are smooth over a given network. Such network encodes the task-task relatedness using edge information. Therefore, the task-task relatedness is incorporated in the model by penalizing the non-smoothness over network. MTL_CMTL assumes that a cluster structure exists in the models. The algorithm then attempts to optimize data fitting and cluster effects simultaneously. Conventional lasso used as baseline for predictive performance comparison was trained using the R library *glmnet* (Friedman, Hastie et al. 2010).

For all algorithms, we implemented a solver based on the accelerated gradient descent method(Nesterov 2012), which takes advantage of information from the previous two iterations to calculate the current gradient and thus achieves a better convergent rate. To solve the non-smooth and convex regularization, the proximal operator(Parikh and Boyd 2014) was applied. Moreover, backward line search was used to determine the appropriate step-size for each iteration. Overall, the solver achieves a complexity of $O(1/k^2)$, which is optimal among first-order gradient methods.

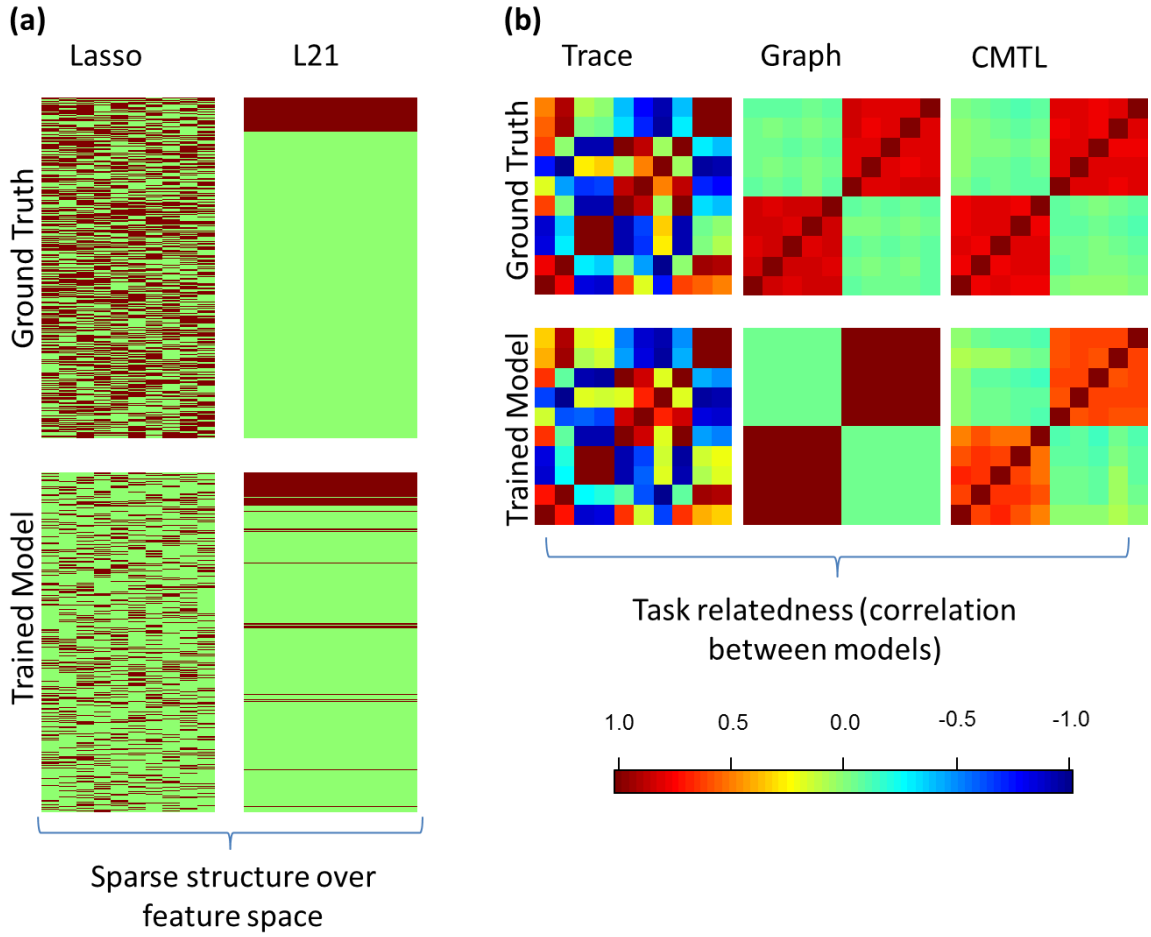
Construction of simulated data

Simulated datasets were constructed as follows. Suppose $t = 10$, $n = 200$ and $p = 400$, then each subject $X_{ij} \sim N(0,1)$, where $i \in \{1, \dots, t\}, j \in \{1, \dots, n\}\ and\ X_{ij} \in R^p$. And the responses $Y_{ij} = sign(X_{ij} \times W_i + 0.5\sigma)$, where $\sigma \sim N(0,1)$ was random noise.

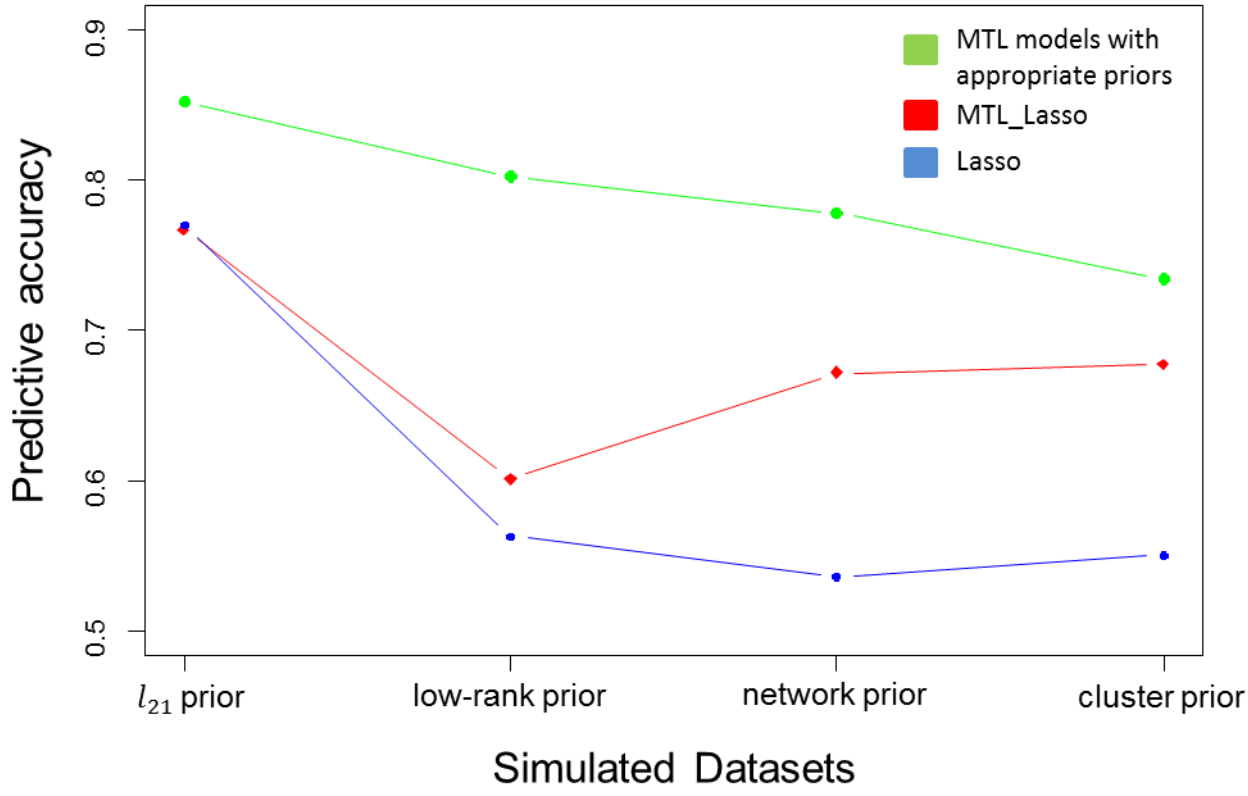
To construct the ground truth model W, we sampled from $W \sim N(0,1)$, and then made modifications depending on the given prior. For example:

- l_1 prior: half of all entries were randomly selected and set to 0
- l_{21} prior: 90% of features across tasks were set to 0
- Low-rank prior: all eigenvalues were set to 0 except for the top 2.
- Network prior and cluster prior: task 1~5 form a group α and 6~10 forms the group β. Then $W_m = \begin{cases} W_\alpha + 0.5\sigma, & m < 5 \\ W_\beta + 0.5\sigma, & m > 5 \end{cases}$

2.6.2 Supplementary Figures



Supplementary Figure 1. Visualization of model interpretability. a) the coefficient matrices of MTC_Lasso and MTC_L21 are shown. The selected predictors are marked in red. **b)** pairwise correlation matrixes capturing between-model (MTC_Trace, MTC_Graph, MTC_CMTL) correlations indicate the relatedness of tasks. For all algorithms, the learnt models are compared to their respective ground truths.



Supplementary Figure 2. Comparison of predictive performance between algorithms. MTL models were compared against the baseline models "conventional lasso" (single task learning) and MTC_Lasso (incorporating an inappropriate prior). The simulated datasets were created according to different priors.

3 STUDY 2: COMPARATIVE EVALUATION OF MACHINE LEARNING STRATEGIES FOR ANALYZING BIG DATA IN PSYCHIATRY

3.1 Abstract

The requirement of innovative big data analytics has become a critical success factor for research in biological psychiatry. Integrative analyses across distributed data resources are considered essential for untangling the biological complexity of mental illnesses. However, little is known about algorithm properties for such integrative machine learning. Here, we performed a comparative analysis of eight machine learning algorithms for identification of reproducible biological fingerprints across data sources, using five transcriptome-wide expression datasets of schizophrenia patients and controls as a use case. We found that multi-task learning (MTL) with network structure (MTL_NET) showed superior accuracy compared to other MTL formulations as well as single task learning, and tied performance with support vector machines (SVM). Compared to SVM, MTL_NET showed significant benefits regarding the variability of accuracy estimates, as well as its robustness to cross-dataset and sampling variability. These results support the utility of this algorithm as a flexible tool for integrative machine learning in psychiatry.

3.2 Introduction

Biological research on psychiatric illnesses has highlighted the scale of investigations required to identify reproducible hallmarks of illness(Sullivan 2010, Passos, Mwangi et al. 2016). In schizophrenia, collaborative analysis of common genetic variants has exceeded 150,000 subjects(Schizophrenia Working Group of the Psychiatric Genomics 2014), demonstrating the challenges tied to low-effect sizes of individual variants, large biological and clinical heterogeneity, and genetic complexity. Not surprisingly, these challenges are also found in other mental illnesses(Major Depressive Disorder Working Group of the Psychiatric, Ripke et al. 2013) and do not seem to be modality specific, as analysis of neuroimaging data, for example, faces similar problems(Wolfers, Buitelaar et al. 2015, Franke, Stein et al. 2016).

The combined “mega-analysis” of data across cohorts and modalities has advantages compared to the more traditional meta-analysis(Major Depressive Disorder Working Group of the Psychiatric, Ripke et al. 2013, de Wit, Alonso et al. 2014), as it makes data amenable for a broader spectrum of computational analyses and allows consideration of confounders across studies. There is growing consensus that advanced computational strategies are required to extract biologically meaningful patterns from these data sources. Beyond functional analysis, a particular focus is on machine learning, which in other areas has shown substantial success in integrating weak signals into accurate classifiers(Jordan and Mitchell 2015). In addition to potential clinical use of such classifiers, the discovery of robust biological patterns may uncover new insights into etiological processes. However, the increasing scale and complexity of big data in psychiatry requires careful evaluation of the most suitable computational strategies. A particularly intuitive and very timely problem is the optimal integration of multi-cohort data, where simple concatenation of datasets may give suboptimal results, and even more so when integration is performed across modalities.

The application of machine-learning techniques on biological problems in psychiatry has already yielded impressive results, including on the prediction of genetic risk, the identification of biomarker candidates or the exploration of etiological mechanisms (Iniesta, Stahl et al.). For example, the use of a Bayesian approach for the incorporation of LD information during polygenic risk score determination

led to a 5% improvement of accuracy in a large schizophrenia dataset (Vilhjalmsson, Yang et al.). In a study exploring the molecular basis of psychiatric comorbidity, an iterative LASSO approach was used for cross-tissue prediction and identified a schizophrenia expression signature that predicted a peripheral biomarker of T2D (Vos, Flaxman et al.). Beyond the analysis of individual data modalities, several machine-learning strategies have been developed for integrative multimodal analysis. For example, a study focusing on the IMAGEN cohort (Whelan, Watts et al.) applied an elastic net model to explore information patterns linked to binge drinking across multiple domains, including brain structure and function, personality traits, cognitive differences, candidate gene information, environmental factors, and life experiences. Similarly, another study (Xia, Ma et al.) explored the inherent data sparsity of neuroimaging and psychiatric symptom data, and successfully stratified subjects using sparse canonical correlation analysis. The study found four dimensions of psychopathology with different patterns of connectivity. In the present study we were particularly interested in the multi-task learning (MTL) which aims to improve generalizability by simultaneously learning multiple tasks (such as case-control associations in different datasets) and these learning processes exchange information to achieve a globally optimal solution (Caruana 1998). Historically, MTL was developed as an extension of neural networks (Caruana 1998), and has since been used across data-intensive research areas, including biomedical informatics (Widmer , Xu, Pan et al. 2011, Zhou, Liu et al. 2013, Feriante 2015, Li, Wang et al. 2016, Yuan, Paskov et al. 2016), speech and natural language processing(Collobert and Weston 2008, Wu, Valentini-Botinhao et al. 2015), image processing and computer vision(Xiaogang, Cha et al. 2009, Zhang, Luo et al. 2014), and web based applications(Chapelle, Shivaswamy et al. 2010, Ahmed, Aly et al. 2012). In psychiatric research, MTL has been applied for integrating measures of cognitive functioning and structural neuroimaging(Marquand, Brammer et al. 2014), as well as for improved fMRI pattern recognition(Jing, Zhilin et al. 2012). In other research fields, MTL approaches have been proposed to combine different sources of biological data, including the linking of MRI or expression with genetic data (Wang, Nie et al. 2012, Lin, Zhang et al. 2014), as well as the integrative analysis of multi-cohort expression data (Xu, Xue et al. 2011).

In the present study, we used MTL to differentiate schizophrenia patients from controls across multiple transcriptome-wide expression datasets. We hypothesized that MTL is particularly suited for this tasks, since it allows the consideration of different cohorts as separate classification tasks. As MTL aims to identify predictive patterns that are shared across tasks, it should uncover expression patterns that are biologically reproducible across cohorts. This may result in better and biologically more relevant classifiers compared to those derived from conventional single task learning (STL), which may be unduly influenced by strong signals present in individual cohorts. To test this, we performed a comparative analysis of different MTL and STL approaches in five transcriptome-wide datasets of schizophrenia brain expression. A 'Leave-dataset-out' procedure was applied to explore and compare the generalizability of the models, with specific focus on classification accuracy, and variability thereof, as well as model sensitivity to cross-dataset and sampling variability.

3.3 Materials and Methods

3.3.1 Datasets

In the present study, five transcriptome-wide expression datasets from schizophrenia post-mortem brains and controls were used for analysis. Details of the datasets are shown in **Table 2**. All datasets were downloaded from the GEO (Gene Expression Omnibus).

Table 2. Overview of demographic details. Values are shown as mean ± sd.

	GSE12679	**GSE35977**	**GSE17612**	**GSE21935**	**GSE21138**
Reference	(Harris, Wayland et al. 2008)	(Chen, Cheng et al. 2013)	(Maycox, Kelly et al. 2009)	(Barnes, Huxley-Jones et al. 2011)	(Narayan, Tang et al. 2008)
n SZ	11	50	22	19	29
n HC	11	50	22	19	29
age SZ	46.1 ± 5.9	42.4 ± 9.9	76 ± 12.9	77.6 ± 11.4	43.3 ± 17.3
age HC	41.7 ± 7.9	45.5 ± 9	68 ± 21.5	67.7 ± 22.2	44.7 ± 16.1
sex SZ (m/f)	7/4	37/13	16/6	11/8	23/6
sex HC (m/f)	8/3	35/15	11/11	10/9	24/5
PMI SZ	33 ± 6.7	31.8 ± 15.4	6.2 ± 4.1	5.5 ± 2.6	38.1 ± 10.8
PMI HC	24.2 ± 15.7	27.3 ± 11.8	10.1 ± 4.3	9.1 ± 4.3	40.5 ± 14
brain pH SZ	NA	6.4 ± 0.3	6.1 ± 0.2	6.1 ± 0.2	6.2 ± 0.2
brain pH HC	NA	6.5 ± 0.3	6.5 ± 0.3	6.5 ± 0.3	6.3 ± 0.2
Genechip	HGU	HuG	HGU	HGU	HGU
Brain Region	PFC	PC	APC	STC	PFC

HGU: HG-U133_Plus_2; HuG = HuGene-1_0-st; APC: anterior prefrontal cortex; PFC: Prefrontal cortex; PC: parietal cortex; STC: superior temporal cortex; HC: healthy control; SZ: schizophrenia.

3.3.2 Preprocessing

Preprocessing was performed using the statistical software R (https://cran.r-project.org/). First, raw expression data were read using the 'ReadAffy' function. Then RMA (Multi-Array Average(Irizarry, Hobbs et al. 2003)) was applied for background correction, quantile normalization and $\log_2$-transformation. Subsequently, multiple probes associated to one gene symbol were averaged. This was followed by selection of common genes across all datasets (17061 genes). For each dataset, propensity score matching was used to obtain a sample with approximate 1:1 matching for diagnosis, sex, ph, age and post-mortem interval (pmi). Next, all datasets were concatenated for quantile normalization and covariate correction. Specifically, the 'Combat' function from the R library *sva*(Leek, Johnson et al. 2012) was applied to correct for covariates (sex, ph, age, age^2, pmi and a dataset indicator). Finally, datasets were separated again for feature standardization (z-score) to remove bias from the expressed genes with large variance and for downstream machine learning analysis.

3.3.3 Machine learning approaches

For MTL, multiple across-task regularization strategies were tested, such as MTL with network structure (**MTL_NET**), sparse network structure (**MTL_SNET**), joint feature learning (**MTL_L21**), joint feature learning with elastic net (**MTL_EN**) and low-rank structure (**MTL_Trace**). As a comparison, we selected logistic regression with lasso (**LR**), linear support vector machines (**SVM**) and random forests (**RF**) as representatives of conventional STL methods. For all models (except for **RF**), stratified 5-fold cross validation was used to select hyper-parameters. Methodological details of the respective methods are described below. All machine-learning analyses were performed using Matlab (R2016b).

Multi-task learning

$$\mathcal{L}(W, C) = \frac{1}{n_i}\sum_{j=1}^{n_i} \log\left(1 + e^{\left(-Y_{i,j}\left(X_{i,j}W_i^T + C_i\right)\right)}\right) \quad (1)$$

For all MTL formulations, logistic loss was used as the common loss function $\mathcal{L}(\cdot)$, where X, Y, W and C referred to the gene expression matrixes, diagnosis status, weight vectors and constants of all tasks, respectively. In addition, i and j denoted the index of the dataset and subject respectively, i.e. n_i and W_i^T referred to the number of subject and weight vector of task i. This model aimed to estimate the effect size of each feature such that the likelihood (i.e. the rate of successful prediction in the training data) is maximized. During the prediction procedure, given the expression profile of a previously unseen individual, the model calculates the probability of belonging to the schizophrenia class (with subjects where the probability exceeded 0.5 being assigned to the patient group). Notably, while we focused on classification due to the categorical outcomes of the investigated datasets, the cross-task regularization strategies explored in the present study are not limited to classification but can also be applied for regression. All MTL formulations were used as implemented in the Matlab library Malsar(Zhou, Chen et al. 2012) or based on custom Matlab implementations.

$$\min_{W,C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda \sum_{i=1}^{t} ||W_i - \frac{1}{t} \sum_{j=1}^{t} W_j ||_2^2 \qquad (2)$$

We selected the mean-regularized multi-task learning method(Evgeniou and Pontil 2004) as an algorithm for the **MTL_NET** framework. This algorithm assumes that a latent model exists underlying all tasks, which can be estimated as the mean model across tasks. Based on this assumption, the formulation attempts to identify the most discriminative pattern in the high-dimensional feature space, while limiting the dissimilarity between pairwise models. Dissimilarity is quantified with respect to the effect size of a given predictor and the sign of its association with diagnosis. We expected this combined dissimilarity measure to lead to biologically plausible predictive patterns that are characterized by consistent differences across tasks, both in terms of magnitude as well as directionality. Here, λ had a range of $10^{(-6:1:2)}$.

$$\min_{W,C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda(\alpha \sum_{i=1}^{t} ||W_i - \frac{1}{t} \sum_{j=1}^{t} W_j ||_2^2 + (1-\alpha)\big||W|\big|_1) \qquad (3)$$

MTL_SNET was the sparse version of **MTL_NET**, and the sparsity was introduced by the l_1 norm (i.e. coefficients of predictors with low utility are set to 0). Here, λ controls the entire penalty and α distributes the penalty to full-sparse and non-sparse terms. λ had a range of $10^{(-6:1:2)}$and α was chosen from the range [0:0.1:1].

$$\min_{W,C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda ||W||_{2,1} \qquad (4)$$

The formulation of **MTL_L21** introduced the group sparse term $||W||_{2,1} = \sum_{i=1}^{p} ||W_i||_2$, which aimed to select or reject the same group of genes across datasets. λ controlled the level of sparsity with a range of $10^{(-6:0.1:0)}$.

$$\min_{W,C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda((1-\alpha)||W||_{2,1} + \alpha||W||_2^2) \qquad (5)$$

The **MTL_EN** was formulated by adding the composite penalties, where $||W||_2^2$ is the squared Frobenius norm. Similar to elastic net in conventional STL, such regularization helped to stabilize the solution when multiple highly correlated genes existed in the high-dimensional space(Tibshirani 2013). Here, λ had a range of $10^{(-6:0.1:0)}$and α was chosen from the range [0:0.1:1].

$$\min_{W,C} \sum_{i=1}^{t} \mathcal{L}(W, C) + \lambda ||W||_* \quad (6)$$

MTL_Trace encouraged a low-rank model W by penalizing the sum of its eigenvalues $||W||_*$. λ had a range of $10^{(-6:0.1:1)}$. By compressing the subspace spanned by weight vectors, models were structured (i.e. clustered structure). Thus, the models that were clustered together demonstrated high pairwise correlation.

Conventional, single-task machine learning

LR_L1: we trained logistic regression with lasso using the package "Glmnet". The lambda parameter was chosen among the set $10^{(-10:0.5:1)}$.

SVM: linear support vector machine was trained using the built-in Matlab function 'fitcsvm' with the box constraints in the range $10^{(-5:1:5)}$. We only used the linear kernel to facilitate determination of predictor importance.

RF: We used the Matlab built-in function 'TreeBagger' to train a random forest model with 5000 trees. The predictor importance was calculated according to the average error decrement for all splits on a given predictor.

Assessment of predictive performance

To quantify predictive performance and capture stability of decision rules against cross-dataset and sampling variability, we used a leave-dataset-out procedure. Specifically, the set of five expression datasets was denoted as $D = \{d_1, d_2, \dots, d_5\}$ and we calculated the power set $\mathbb{p}(D)$ of D. Then for each subset $d \in \mathbb{p}(D)$, we trained a given algorithm on d and tested the model on $D - d$. For example, for $d = \{d_1, d_2\}$, we trained using the combination of datasets $\{d_1, d_2\}$ and then tested on $\{d_3, d_4, d_5\}$. For convenience, we organized these training procedures according to the size of d, noted as $n_d \in \{2,3, \dots,5\}$. We thus obtained a series of models trained using all subsets of the five datasets (except for single dataset) and they are referred to using n_d.

The comparison of the predictive performance between methods was mainly based on $n_d = 4$, i.e. when all but one dataset were used for training. To understand how dataset-specific confounders affect the prediction, models were trained on a range of n_d from 2 to 4. Finally, to explore the convergence of genes' coefficients across different training datasets, we compared the models trained when $n_d = i, i \in \{2,3 \dots 5\}$.

During cross-validation (CV), as illustrated in **Figure A1,** subjects were randomly allocated to 5 folds, stratified for diagnosis and the dataset indicator. Subsequently, different strategies were specified for MTL and STL. For MTL, the training_{cv}datasets were trained in parallel, and the models were tested on each test_{cv} dataset by averaging the prediction scores. To determine the final accuracy of the current fold, the accuracies retrieved from all test_{cv} datasets were averaged. For STL the training_{cv}datasets were combined to train a single algorithm that was then predicted on the combined test_{cv}datasets. Similar to CV, in the training procedure, MTL trained on datasets in parallel, while combining the prediction scores for testing.

Consistency and stability analysis

To compare the consistency and stability of markers between algorithms, we use the correlation coefficient as the similarity measure of pairwise transcriptomic profiles (i.e. the coefficient vector for all genes) learnt by algorithms. A high similarity between profiles implied that models shared important

predictors with respect to their weights and signs. Using this similarity measure, 'consistency' and 'stability' were defined, respectively. These measures were derived from 100-fold stratified bootstrapping of subjects from a set of datasets. In each bootstrapping sample, we tested across the number of training sets ($n_d = i, i \in \{2,3,\ldots,5\}$). For MTL, since the training procedure would output multiple coefficient vectors (i.e. training on three datasets would output three coefficient vectors), to compare the similarity between algorithms, the coefficient vectors were averaged.

Consistency: With 'consistency' we quantified the pairwise similarity of models trained using overlapping or non-overlapping (i.e. 2 training datasets) datasets. For this, we differentiated two types of consistencies: 'horizontal' and 'vertical' consistency as illustrated in **Figures A2a and A2b,** respectively. Horizontal consistency quantified model robustness against cross-dataset variability. For this, we fixed the number of training datasets (n_d), and determined the pairwise similarity between models. This was performed for all possible choices of n_d (see supplementary methods for details). Vertical consistency measured the sensitivity of models to the number of training datasets. For this, we varied n_d and quantified similarity between the model determined on all training datasets ($n_d = 5$) and all models derived from lower training datasets numbers ($n_d = i, i \in \{2,3,4\}$) (see supplementary methods for details). Low vertical consistency would, for example, be observed when models trained on two training datasets led to vastly different transcriptomic profile compared to that using all five datasets for training.

Stability: To quantify the stability of an algorithm against the sampling variability, we observed the variation of transcriptomic profiles learnt from different bootstrapping samples as illustrated in **Figure A3**. Then the variation of all models given n_d was summarized as the stability (see supplementary methods for details).

Success rate: In addition to consistency and stability, to perform a side-by-side comparison of algorithms, we defined the success rate as the proportion of cases where one algorithm outperformed the other. For example, we quantified the success rate of consistency as the proportion of bootstrapping samples where the first algorithm demonstrated higher consistency than the second (see supplementary methods for details). The success rate of stability was quantified as the proportion of models, which were more stable for the first algorithm than that for the second (see supplementary methods for details).

3.4 Results

3.4.1 Accuracy comparison between MTL and STL

Figure 1 shows a comparison of average classification accuracies when 4 out of 5 datasets were used for training and the remaining dataset for testing. The distributions of accuracies are shown for 10 repetitions of the classification procedure, to assess the variability caused by parameter tuning via cross-validation. With an average accuracy of 0.73, **MTL_NET** outperformed all other methods, followed by **SVM** that had a marginally inferior accuracy of 0.72. Moderate accuracies were observed for **MTL_Trace** (0. 69), **MTL_L21** (0. 66) and **RF** (0.68). The sparse logistic regression performed worst (0. 64). As an extension of **MTL_NET** and **MTL_L21** respectively, **MTL_SNET** (0. 71) and **MTL_EN** (0. 66) achieved similar accuracies to their original algorithms. In the following analysis, we focused on the comparison of **MTL_NET** and **SVM** as representatives of MTL and STL, respectively.

In **Figure 1**, the standard error of accuracies for **SVM** (0.011) was slightly smaller than that for **MTL_NET** (0.012), indicating that **SVM** might be more robust regarding parameter selection. A possible reason was that **SVM** obtained higher statistical power by comparing cases and controls across datasets. In

contrast, **MTL_NET** derived transcriptomic signatures using cases and controls within datasets, limiting the statistical power.

3.4.2 Dependency of classification performance on the number of training datasets

We performed a side-by-side comparison of **MTL_NET** and **SVM** to explore the dependency of classification performance on the number of available training datasets. **Figure 2a** shows that increasing accuracy was observed for both **MTL_NET** and **SVM** with increasing numbers of training datasets. Notably, **MTL_NET** only outperformed SVM at $n_d = 4$ (4 datasets used for training), suggesting that MTL required a higher dataset number to identify a reproducible biological pattern. However, we observed that the variation of accuracies for **MTL_NET** substantially decreased with increasing numbers of training datasets (**Figure 2b**), which was not the case for **SVM**. This suggested that **MTL_NET** was more conservative in that accuracy was not driven by highly successful prediction on individual test set, but by improved predictability observed for all test sets.

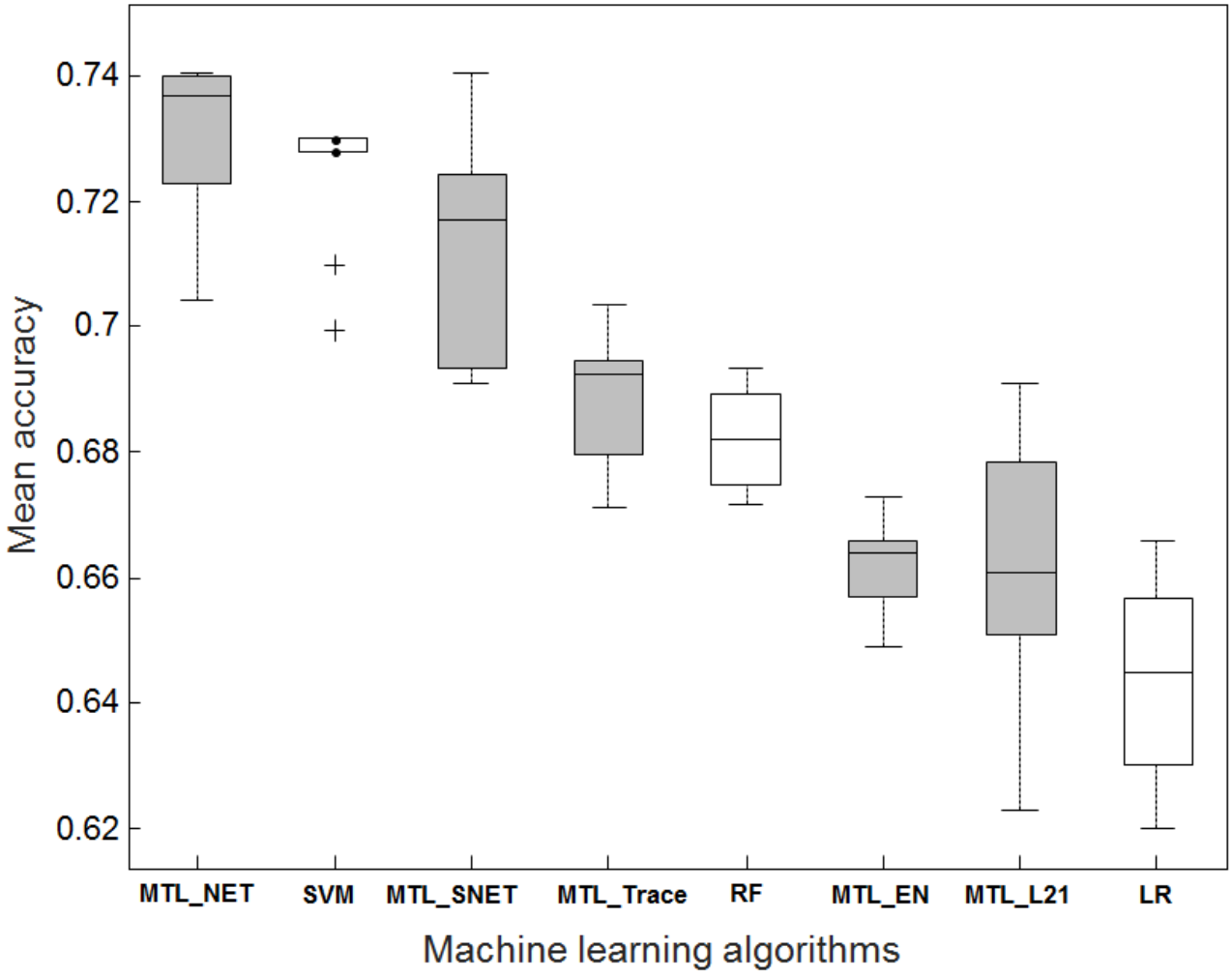


Figure 1. Predictive performance comparison between 8 algorithms. The 'leave- dataset-out' procedure was used for comparison. Four out of five datasets were combined for training, and then the model was tested on the remaining dataset. The distribution of accuracy estimates indicated the variation of parameter selection across 10 repetitions. The boxplots in gray denote the multi-task learning algorithms.

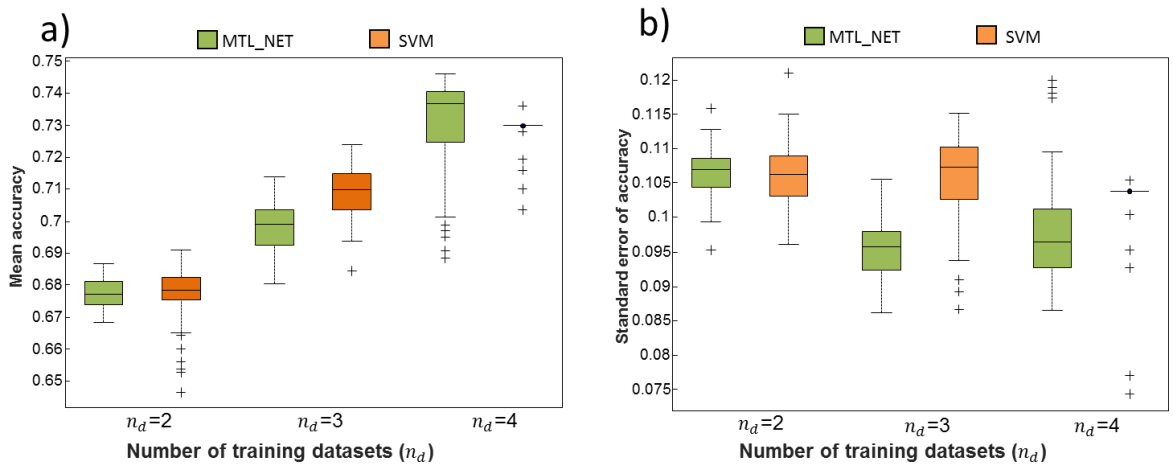


Figure 2. Distribution of classification accuracies and their standard errors across different numbers of training datasets. The Figure shows the mean (a) and standard error (b) of classification accuracies obtained for different numbers of training datasets (n_d). Performance was evaluated from the test datasets not used for training. The variation of the boxplot was due to the sampling variability during cross-validation.

3.4.3 Consistency and stability of trained models

Figures 3a and 3b show that, in terms of vertical and horizontal consistency, **MTL_NET** outperformed **SVM**, independently of the number of training datasets. This indicated that similar discriminative patterns of genes were identified by MTL across training datasets, and implied strong robustness against cross-dataset variability. In particular, the superior performance of vertical consistency for **MTL_NET** showed that this algorithm was less sensitive to the small numbers of training datasets compared to **SVM**. **Table 1** shows the mean consistency (both horizontal and vertical) across bootstrapping samples. Compared to **SVM**, **MTL_NET** achieved higher mean consistency by approximately 1.6% for horizontal and 2.2% for vertical consistency. Notably, the success rate of consistency was 100%, independent of the number of training sets, showing that **MTL_NET** models consistently identified higher transcriptomic profile robustness across bootstrapping samples than **SVM.**

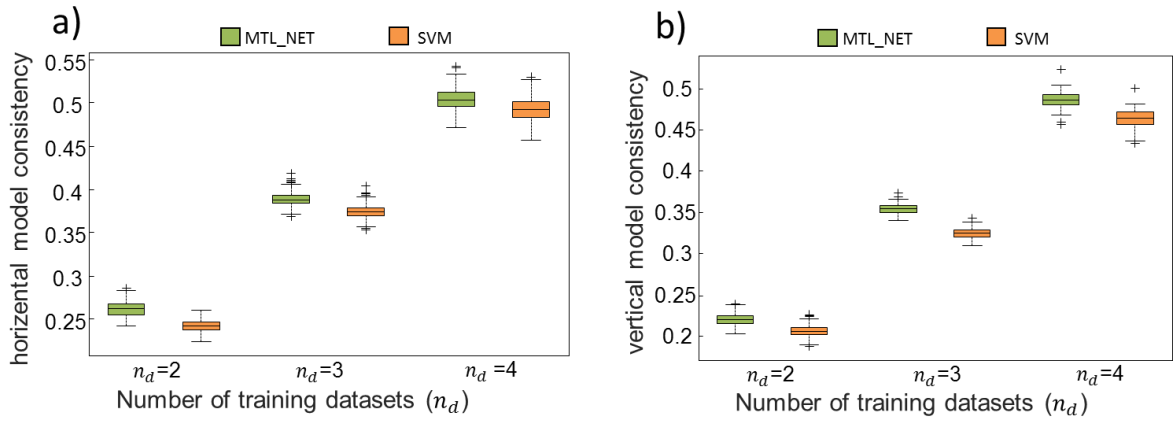


Figure 3. Horizontal and vertical model consistency. To analyze the consistency of a given machine-learning algorithm against the cross-dataset variability, we quantified the horizontal and vertical model consistency. Specifically, horizontal consistency quantified the similarity

between models trained using the same number of number of datasets, and vertical consistency quantified the pairwise similarity of models, where one was trained using all datasets and the other was trained using less datasets. Stratified 100-fold bootstrapping procedure was applied to quantify the variation of the consistency.

Table 1. Mean consistency, stability and success rate across the number of training sets n_d

MTL_NET/SVM	$n_d = 2$	$n_d = 3$	$n_d = 4$	$n_d = 5$
Horizontal consistency	**0.26**/ 0.24	**0.39**/ 0.37	**0.51**/ 0.49	-
Vertical consistency	**0.22**/ 0.21	**0.35**/ 0.33	**0.49**/ 0.46	-
Stability	**0.64**/ 0.63	**0.65**/ 0.64	**0.65**/ 0.64	**0.654**/ 0.645
Success rate (horizontal consistency)	1	1	1	-
Success rate (vertical consistency)	1	1	1	-
Success rate (stability)	1	1	1	1

To further identify the robustness of models against sampling variability, we quantified the algorithms' stability. In **Figure 4**, across the number of training datasets n_d, the increasing trend of stability demonstrated that both **MTL_NET** and **SVM** gained more robustness against sampling variability with an increasing number of subjects used for training. However, **MTL_NET** demonstrated higher stability than **SVM** independently of the number of training datasets (**Figure 4**). The mean stability across models also supported the result (**Table 1**). Moreover, the mean stability for **MTL_NET** was 1.2% higher than **SVM** (100% success rate of stability across all n_d, **Table 2**).

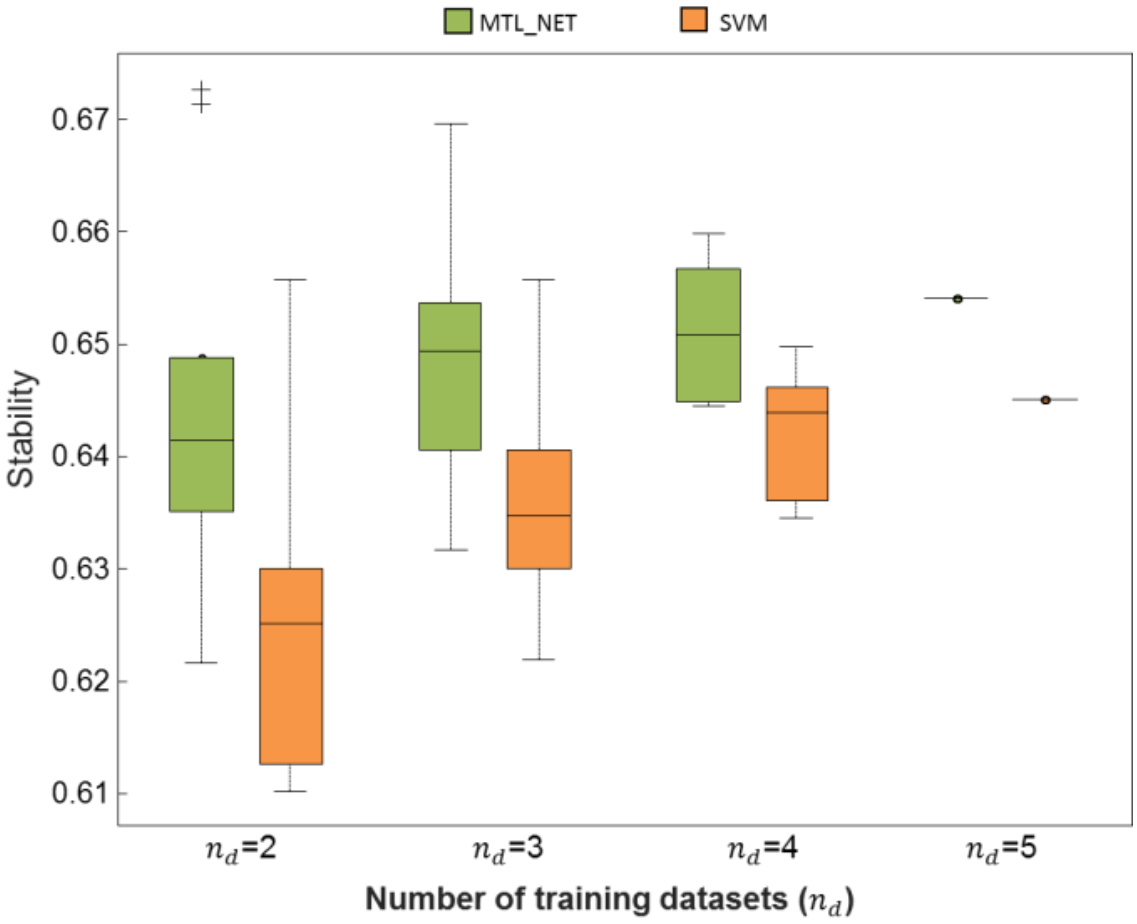


Figure 4. Stability comparison. The stability quantified the robustness of an algorithm against sampling variability. For each n_d, stability was computed as the pairwise similarity of models trained from two given bootstrap samples. The stability was then averaged across bootstrap

samples. In the figure, the distribution of the stability was due to the different combination of training datasets given n_d.

We did not perform comparative functional analysis of markers identified by the two algorithms, since marker sets were quite similar. For example, using all five datasets for training, the average similarity over all bootstrapping samples was 98.75%, suggesting that similar functional implications would be derived for these algorithms.

3.5 Discussion

The present study provides a comparative evaluation of using MTL for integrative machine learning, compared to classical, single task learning in five transcriptome-wide datasets of schizophrenia brain expression. Overall, MTL showed similar accuracy, albeit with lower variability, compared to STL. Accuracy estimates varied by up to approximately 10% between algorithms, suggesting different sensitivities of algorithms to cross-dataset heterogeneity as well as sampling variability. Among all MTL formulations, **MTL_NET** was most predictive. This was likely due to the fact that it harmonized algorithms across tasks with respect to both predictor weight and sign of diagnosis association, resulting in biologically plausible predictive patterns. In contrast, **MTL_L21** ignores the sign of association and **MTL_Trace** improves models' correlation in each subspace but failed to modulate the cross-subspace correlation. Contrary to the usual assumption that simpler models show improved generalizability(O'Brien 2016), a sparse version of **MTL_NET** (**MTL_SNET**) did not improve the prediction. This may be due to the fact that the sparse model was trained by constructing a solution tree among an unlimited number of optimal solution trees. Although these solution trees have similar performance on the training dataset, they may show differently predictive ability on cross-modality test dataset because the i.i.d assumption may not hold. **MTL_NET** (as well as **SVM**), solves a strictly convex optimization problem, resulting in a uniform solution in the entire feature space, which may be equally effective when tested on independent test data.

The higher consistency and stability of **MTL_NET** implied that a set of similar differentially expressed genes were identified for multiple training datasets. In addition, these genes demonstrated higher predictability and robustness against study-specific effects, which is particularly important for data integration in multi-modal analyses, such as the integrative analysis of genetic and expression data(Gandal, Haney et al. 2018) or the analysis of shared markers across multiple comorbid conditions(International Schizophrenia, Purcell et al. 2009, Cross-Disorder Group of the Psychiatric Genomics, Lee et al. 2013, Bulik-Sullivan, Finucane et al. 2015).

An interesting observation of the present study was that for **MTL_NET**, the variance of the classification accuracy substantially decreased with increasing the number of training datasets. This suggested that **MTL_NET** selected biological signatures with similar effect sizes across independent training datasets, further supporting the biological reproducibility of the identified patterns. In contrast, **SVM** did not show a decreasing accuracy variance with increasing numbers of training datasets. This indicates that despite the increasing classification accuracy, the identified signatures worked well only for some, but not other test datasets. These results for these particular datasets highlight differences between single and multi-task learning regarding the variance of the test-set accuracy, which is a fundamentally important consideration for study design and interpretation of classifier reproducibility.

3.6 Supplements

3.6.1 Supplementary Methods

Consistency, stability and success rate

Notations:

- The model pairs trained using different (overlapping, or non-overlapping) combinations of datasets were represented as $M$and $\widetilde{M}$ respectively (i.e. M represented the model trained using the training set $\mathrm{d} = \{1, 2\}$, $\widetilde{M}$ was trained using a different dataset combination, for example $\mathrm{d} = \{3,4\}$ or $\mathrm{d} = \{1, 2, \dots ,5\}$)
- The notation of an algorithm: α, β (i.e. α=MTL_NET, β=SVM)
- The index of the bootstrapping sample: $b \in \{1,2, \dots 100\}$ and $\tilde{b} \in \{1,2, \dots 100\}$. For computational efficiency, bootstrapping was performed across all datasets $\mathrm{d} = \{1, 2, \dots ,5\}$ and data subsets were selected from this sampling.

As an example, a model M_b^α could be trained based on bootstrap sample $b = 3$, from which training sets $\mathrm{d} = \{1, 2\}$ were extracted, using algorithm $\alpha = \mathrm{SVM}$. The model trained on the same bootstrap sample based on a different combination of training sets and using algorithm $\alpha = \mathrm{SVM}$ would be denoted as $\widetilde{M}_b^\alpha$.

Consistency

Given $n_d = i, i \in \{2,3,4\}$ and algorithm α, we calculated the expected similarity for each bootstrapping sample b as

$C_b^{\alpha,n_d} = \mathbb{E}_{M,\widetilde{M},M\neq\widetilde{M}}[\![Cor(M_b^\alpha, \widetilde{M}_b^\alpha)]\!]$

Then the expected similarity list $C^{\alpha,n_d} = [C_1^{\alpha,n_d}, C_2^{\alpha,n_d}, \dots, C_{100}^{\alpha,n_d}]$ over b was the consistency list of algorithm α for a given n_d. Here, the expectation was calculated empirically by enumerating all pairs of models $M$and $\widetilde{M}$. By assigning different values to $M$and $\widetilde{M}$, horizontal and vertical consistency were differentiated. For horizontal consistency, $M$and $\widetilde{M}$ represented the pairwise models trained using the same number (n_d) of datasets. For vertical consistency, $\widetilde{M}$ was trained using $n_d = 5$ datasets and $M$was trained using fewer datasets.

Stability

Given $n_d = i, i \in \{2,3,4\}$, and algorithm α, we quantified the expected similarity between pairwise models (M_b^α and $M_{\tilde{b}}^\alpha$) which were trained using the same datasets (M) but different bootstrapping samples (b and $\tilde{b}$) as

$$S_M^{\alpha,n_d} = \mathbb{E}_{b,\tilde{b},b\neq\tilde{b}}[\![Cor(M_b^\alpha, M_{\tilde{b}}^\alpha)]\!]$$

Over all models (M), $S^{\alpha,n_d} = [S_1^{\alpha,n_d}, S_2^{\alpha,n_d}, \dots, S_{\binom{5}{n_d}}^{\alpha,n_d}]$ was quantified as the stability list of algorithm α, given n_d. The expectation was estimated empirically by enumerating all pairs of bootstrapping samples b and $\tilde{b}$.

Success rate

The success rate compared algorithms α and β side-by-side, and was measured as the proportion of cases where algorithm α outperformed β.

For example, given the consistency list of algorithm α and β (C^{α,n_d} and C^{β,n_d}), we determined the proportion of bootstrapping samples where algorithm α demonstrated higher consistency than β, yielding the success rate of consistency:

$$SR_C^{n_d} = \mathbb{E}_b\left[\!\left[1_{C_b^{\alpha,n_d} - C_b^{\beta,n_d} > 0}\right]\!\right]$$

Given the stability list of algorithm α and β ($S^{\alpha,n_d}\ and\ S^{\beta,n_d}$), we determined the proportion of models, which demonstrated higher stability for algorithm α, yielding the success rate of stability:

$$SR_S^{n_d} = \mathbb{E}_M\left[\left[1_{S_M^{\alpha,n_d} - S_M^{\beta,n_d} > 0}\right]\right]$$

3.6.2 Supplementary Figures

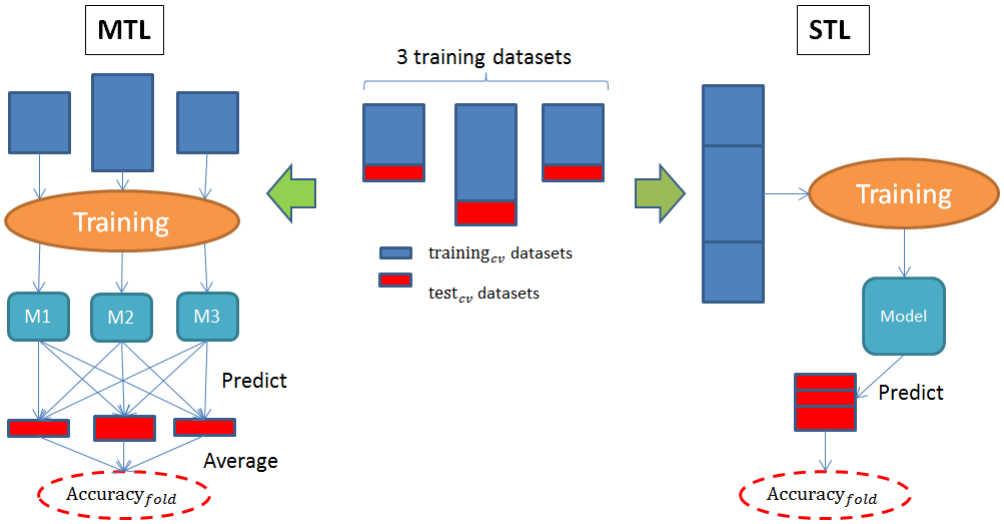


Figure A1. Procedure of 5-fold-stratified-cross-validation for STL and MTL (showing one fold as example). Using $\boldsymbol{n_d = 3}$ as an example, the specific procedure of the cross-validation procedure is shown. First, the subjects were randomly allocated to 5 folds, stratified for diagnosis per dataset. Subsequently, different strategies were specified for MTL and STL. For MTL, the training datasets were trained in parallel, and the three models (M1, M2 and M3) were tested on each test dataset by averaging the prediction score. The average across all accuracies was used as final accuracy for the current fold. In contrast, for STL the training datasets were combined to train a single algorithm that was then predicted on the combined test datasets.

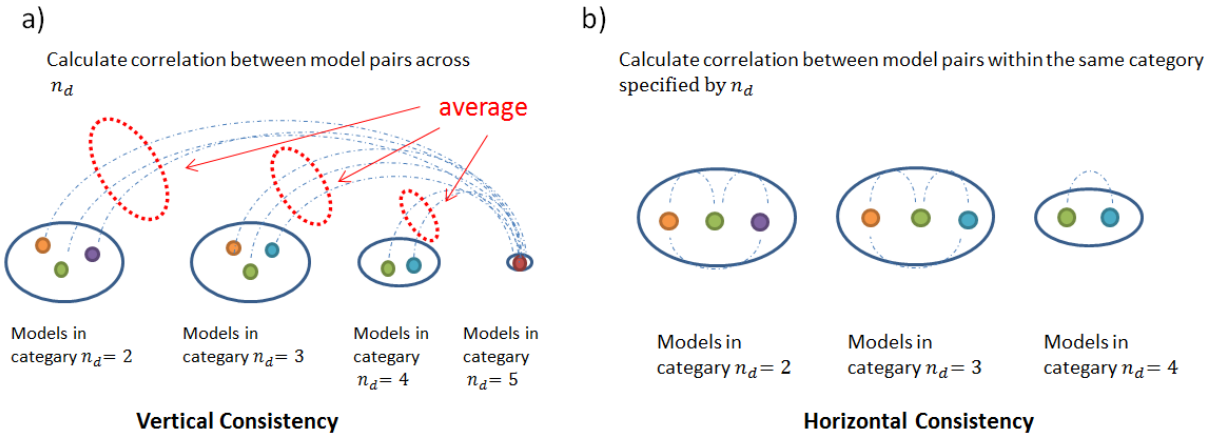


Figure A2. Illustration of model consistency calculation. Consistency quantified the robustness of an algorithm against the cross-dataset variability. To test this, we trained models using each subset of all 5 expression datasets and then categorized these models according to the number of training sets ($\boldsymbol{n_d}$). Different models were rendered as colored circles, categorized by $\boldsymbol{n_d}$. For vertical consistency (a) the similarity was determined between the models learned on $\boldsymbol{n_d = 2}$ to $\boldsymbol{n_d = 4}$ and the model trained on $\boldsymbol{n_d = 5}$.The resulting values were then averaged for a given category $\boldsymbol{n_d}$. For horizontal consistency (b) the model similarity was calculated in each category $\boldsymbol{n_d}$ and then averaged.

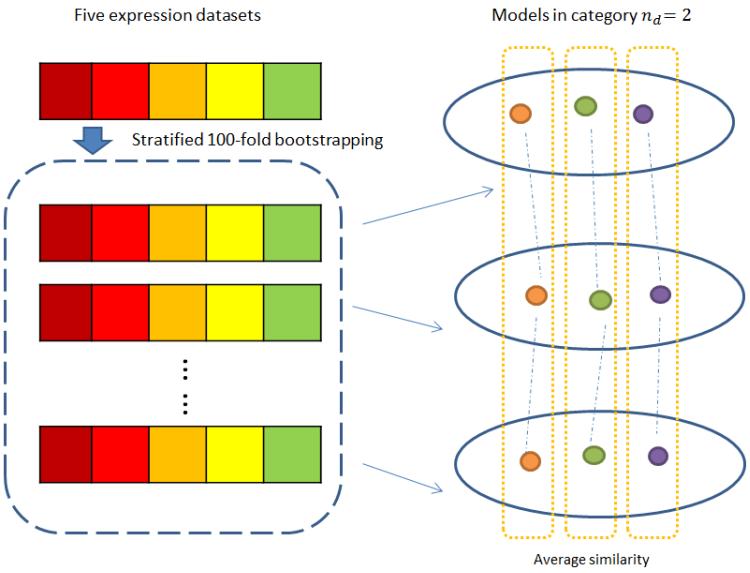


Figure A3. Illustration of model stability calculation. Stability quantified the robustness of an algorithm against sampling variability. This metric was computed by performing 100-fold-stratified-bootstrapping. In the left panel, 5 expression datasets are shown as colored boxes. Using $\boldsymbol{n_d = 2}$ as an example, two out of five datasets were combined for training in each bootstrapping sample. Thus, a series of models were obtained as illustrated as the colored circles in the right panel. The stability was determined as the average pairwise similarity for each model, calculated across all pairs of bootstrapping samples.

4 STUDY 3: DSMTL - A COMPUTATIONAL FRAMEWORK FOR PRIVACY-PRESERVING, DISTRIBUTED MULTI-TASK MACHINE LEARNING

4.1 Abstract

Multitask learning allows the simultaneous learning of multiple 'communicating' algorithms. It is increasingly adopted for biomedical applications, such as the modeling of disease progression. As data protection regulations limit data sharing for such analyses, an implementation of multitask learning on geographically distributed data sources would be highly desirable. Here, we describe the development of dsMTL, a computational framework for privacy-preserving, distributed multi-task machine learning that includes three supervised and one unsupervised algorithms. dsMTL is implemented as a library for the R programming language and builds on the DataSHIELD platform that supports the federated analysis of sensitive individual-level data. We provide a comparative evaluation of dsMTL for the identification of biological signatures in distributed datasets using two case studies, and evaluate the computational performance of the supervised and unsupervised algorithms. dsMTL provides an easy-to-use framework for privacy-preserving, federated analysis of geographically distributed datasets, and has several application areas, including comorbidity modeling and translational research focused on the simultaneous prediction of different outcomes across datasets. dsMTL is available at https://github.com/transbioZI/dsMTLBase (server-side package) and https://github.com/transbioZI/dsMTLClient (client-side package).

4.2 Introduction

The biology of many human illnesses is encoded in a vast number of genetic, epigenetic, molecular, and cellular parameters. The ability of Machine Learning (ML) to jointly analyze such parameters and derive algorithms with potential clinical utility has fueled a massive interest in biomedical ML applications. One of the fundamental requirements for such ML algorithms to perform well is the availability of data at a large scale, a challenge of steadily declining importance due to the ever-increasing availability of biological data(Jahanshad, Kochunov et al. 2013, Kochunov, Jahanshad et al. 2014, Schizophrenia Working Group of the Psychiatric Genomics 2014). As data can often not be freely exchanged across institutions due tothe need for protection of the individual privacy, the utility of 'bringing the algorithm to the data' is becoming apparent. Technological solutions for this task have thus risen in popularity and exist in various forms. One of the most straightforward approaches is the so-called federated ML, where algorithms are simultaneously learned at different institutions and optimized through a privacy-preserving exchange of parameters. Other approaches for this task include the training of ML algorithms on temporarily combined data stored in working memory(Carter, Francis et al. 2016) or the more recently introduced 'swarm-learning' approach(Warnat-Herresthal, Schultze et al. 2021). One commonality of most ML algorithms, federated or not, is the assumption that all investigated observations (e.g. illness-affected individuals) represent the same underlying population. However, in biomedicine, this is rarely the case, as biological and technological factors frequently induce cohort-specific effects that limit the ability to identify reproducible biological findings. Multitask Learning (MTL) can address this issue through the simultaneous learning of outcome (e.g. diagnosis) associated patterns across datasets with dataset-specific, as well as shared, effects. Multi-task learning has numerous exciting application areas, such as comorbidity modeling, and has already been applied successfully for e.g. disease progression analysis(Zhou, Liu et al. 2013).

Here, we describe the development of dsMTL ('Federated Multi-Task Learning for DataSHIELD'), a package of the statistical software R, for **Fe**derated **M**ulti-**T**ask **L**earning (FeMTL) analysis (**Figure 1**) . dsMTL was developed for DataSHIELD(Gaye, Marcon et al. 2014), a platform supporting the federated analysis of sensitive individual-level data that remains stored behind the data owner's firewall throughout analysis(Wilson, Butters et al. 2017). dsMTL includes three supervised and one unsupervised federated multi-task learning algorithms that extend algorithms previously developed for non-federated analysis (for R implementations, see (Yang and Michailidis 2016, Cao, Zhou et al. 2018)). Specifically, the **dsMTL_L21** approach allows for cross-task regularization, building on the popular LASSO method, in order to identify outcome-associated signatures with a reduced number of features shared across tasks. The non-federated version of this approach has previously been applied to simultaneously predict multiple oncological outcomes using gene expression data(Xu, Xue et al. 2011). The **dsMTL_trace** approach constrains the coefficient vectors in a low-dimensional space during the training procedure to penalize the complexity of task relationships, resulting in an improved generalizability of the models. In a non-federated implementation, this method has previously been used to predict the response to different drugs, and the identified models showed a high degree of interpretability in the context of the represented drug mechanism(Yuan, Paskov et al. 2016). **dsMTL_net** incorporates the task relationships that can be described as a graph, in order to improve biological interpretability. In a non-federated version, this technique has previously been used for the integrative analysis of heterogeneous cohorts(Cao, Meyer-Lindenberg et al. 2018) and for the prediction of disease progression(Zhou, Yuan et al. 2011). The **dsMTL_iNMF** approach is an unsupervised, integrative non-negative matrix factorization method that aims at factorizing the cohorts' data matrices into shared and dataset-specific components. Such modeling has been applied to explore dependencies in multi-omics data for biomarker identification(Yang and Michailidis 2016, Fujita, Mizuarai et al. 2018). In addition to the FeMTL methods, we also implemented a federated version of conventional Lasso (dsLasso) (Tibshirani 1996) in dsMTL package due to its wide usage in biomedicine and as a benchmark for testing the performance of the federated MTL algorithms.
To explore the utility of the dsMTL algorithms, we used a network comprising three servers. These servers hosted simulated data with variable degrees of cross-dataset heterogeneity, in order to test the ability of the MTL algorithms to suitably characterize shared and specific biological signatures. In addition, we analyzed actual RNA sequencing and microarray data across the three-server network, to show that the accurate analysis can be performed in acceptable runtime using dsMTL in real network latency.

4.3 Results

Here we show the results for two case studies. The first case study aims at demonstrating the utility of the supervised **dsMTL_L21** algorithm to identify 'heterogeneous' target signatures across the data network. With 'heterogeneous' we describe signatures that involve the same features (e.g. genes) but with potentially differing signs (indicating differential directions of influences) across datasets. In contrast, 'homogeneous' signatures relate to the same features and signs across datasets. The second case study focuses on the unsupervised **dsMTL_iNMF** method and explores the utility of the federated implementation, compared to the aggregation of local NMF models, to disentangle shared and cohort-specific components across datasets. For all case studies, we evaluated the signature identification accuracy as the major metric. For predictions of clinical outcomes, the prediction accuracy was also demonstrated.

Case study 1 – distributed MTL for identification of heterogeneous target signatures

With the aim to identify 'heterogeneous' signatures, we compared the performance of dsMTL_L21, dsLasso and the bagging of glmnet models. As part of this, we explored the sensitivity of these methods to different sample sizes (*n*) relative to the gene number (*p*). **Figure 2** shows the resulting prediction performance and gene selection accuracy, each averaged over 100 repetitions. dsLasso showed the worst prediction performance in this heterogeneous setting, and dsMTL_L21 slightly outperformed the aggregation of local models (glmnet). Similarly, the gene selection accuracy of dsLasso was inferior to that of dsMTL_L21 and glmnet-bagging, which showed similar performance when the sample size is sufficiently large, e.g. the number of subjects approximately equal to the number of genes (n/p ~1). However, with a decreasing n/p ratio, dsMTL_L21 showed an increasing superiority over the other methods, especially for n/p=0.15, where the gene selection accuracy of dsMTL_L21 was over 2.8 times higher than that of the bagging technique.

Case study 2 – distributed iNMF for disentangling shared and cohort-specific signatures

Figure 3 shows the performance of distributed and aggregated local NMF methods for disentangling shared and cohort-specific signatures from multi-cohort data, given different 'severities' of the signature heterogeneity. For both types of signatures, dsMTL_iNMF outperformed the ensemble of local NMF models for any heterogeneity severity setting. Notably, even with increasing heterogeneity, the accuracy of dsMTL_iNMF to capture shared genes remained stable at approximately 100%, illustrating the robustness of dsMTL_iNMF against the heterogeneity's severity shown in **Figure 3c**. In contrast, for the ensemble of local NMF, the gene selection accuracy of the shared signature continuously decreased to approximately 50% (20% of outcome-associated genes were shared among cohorts), while the gene selection accuracy of cohort-specific signatures continuously increased to 75% (20% of outcome-associated genes were shared among cohorts) as shown in **Figures 3a** and **3b**.

Efficiency of supervised dsMTL

We aimed at determining the efficiency of supervised dsMTL using the real molecular data and the actual latency of a distributed network. Using a three-server scenario (see **Table 2 Supplementary Results**; two servers at the Central Institute of Mental Health, Mannheim; one server at BioQuant, Heidelberg University) we analyzed four case-control gene expression datasets of patients with schizophrenia and controls (median n=80; 8013 genes). **Supplementary Table 3** shows the comparison between dsLasso and mean-regularized dsMTL_net, which were trained (cross-validation + training) and tested in approximately 8min and 10min, respectively, with the time-difference being due to the increased network access of dsMTL. The prediction accuracy of dsMTL was slightly higher than that of dsLasso, consistent with our previous study(Cao, Meyer-Lindenberg et al. 2018). Regarding model interpretability, dsLasso captured a signature comprising 38 genes but could not distinguish shared and cohort-specific effects. Mean regularized dsMTL identified a signature with 10 genes shared among all cohorts, with 163 genes shared by two cohorts, as well as three cohort-specific signatures comprising 1532 genes.

Efficiency of unsupervised dsMTL

The cohorts and server information is shown in **Supplementary Table 4**. It took 34.9 minutes (1,003 times network accesses) to train a dsMTL_iNMF model with 5 random initializations (~7 min for each initialization). The factorization rank k=4 was selected as the optimal parameter. In **Supplementary Figure 1**, the objective curve illustrates that the training time was sufficient for model convergence. In

this analysis, a shared signature comprising 473 genes between SCZ and BIP was identified, while two disease-specific signatures containing 37 genes for SCZ and 152 genes for BIP, respectively, were found.

4.4 Discussion

We here present dsMTL – a secure, federated multi-task learning package for the programming language R, building on DataSHIELD as an ecosystem for privacy-preserving and distributed analysis. Multi-task learning allows the investigation of research questions that are difficult to address using conventional ML, such as the identification of heterogeneous, albeit related, signatures across datasets. The implementation of a privacy-preserving framework for the distributed application of MTL is an essential requirement for the large-scale adoption of MTL. Using such a distributed server setup, we demonstrate the applicability and utility of dsMTL to identify biomarker signatures in different settings. For applications where the target biomarker signatures are different, but relate to an overlapping set of features (explored here as the 'heterogeneous' case), conventional machine learning would not be a meaningful algorithm choice. We show that MTL is able to identify the target signatures with high confidence and may thus be a reasonable choice for a diverse set of interesting analyses. As mentioned above, a particularly noteworthy application is comorbidity modeling, where the target signatures index the shared (although potentially heterogeneously manifested) biology of multiple, clinically comorbid conditions. Such analyses could potentially be a powerful, machine learning-based extension of comorbidity modeling approaches based on univariate statistics that have already been very useful for characterizing the shared biology of comorbid illness(Lichtenstein, Yip et al. 2009). We show that unsupervised MTL can disentangle the shared from cohort-specific effects, demonstrating its potential utility for comorbidity analysis. Other applications for this method include the analysis of biological patterns shared across clinical symptom domains, between clinical and demographic characteristics, or with digital measures, such as ecological momentary assessments.

The use of dsMTL follows the concept of the so-called "freely composing script" in the DataSHIELD ecosystem. It organizes a given dsMTL workflow as a free composition of dsMTL, DataSHIELD, and local R commands (e.g. R base functions, customer-defined functions and CRAN packages) into a script, such that the geo-distribution of datasets and the federated computation are transparent to users. This concept is similar to that of the "freely composing apps" used in a recently presented federated ML application(Matschinske, Späth et al. 2021), which allows flexible scheduling of functions in the form of apps and improves the federated data analysis flexibility for users. In addition to dsMTL, other packages in the DataSHIELD ecosystem exist for e.g. "big data" storage and management(Marcon, Bishop et al. 2021), various statistical tests(Gaye, Marcon et al. 2014, Marcon, Bishop et al. 2021) and deep learning(Lenz, Hess et al. 2021, Marcon, Bishop et al. 2021).

Interesting future developments of the dsMTL approach could include the implementation of asynchronous communication, which provides a probabilistically approximate solution but faster convergence(Xie, Baytas et al. 2017, Zhang and Liu 2020). Furthermore, integration of other popular systems for ML, such as tensorflow(Dahl, Mancuso et al. 2018), for which interfaces with the R language already exist, would provide valuable additions to the DataSHIELD system. Finally, a noteworthy consideration is an architecture underlying the distributed data infrastructure. DataSHIELD builds on a centralized ("client-server") architecture and each data provider needs to install a well-configured data warehouse. Such infrastructure is suitable for long-term collaboration scenarios and large consortia projects that conduct a broad spectrum of complex analyses requiring

high flexibility. However, in other scenarios that require more temporary and easy-compute collaboration setups, a server-free or decentralized architecture(Warnat-Herresthal, Schultze et al. 2020) might be more suitable, because the cost of data provider for participating is low.
In conclusion, the dsMTL library for the programming language R provides an easy-to-use framework for privacy-preserving, federated analysis of geographically distributed datasets. Due to its ability to disentangle shared and cohort-specific effects across these datasets, dsMTL has numerous interesting application areas, including comorbidity modeling and translational research focused on the simultaneous prediction of different outcomes across datasets.

4.5 Methods

Modeling
All methods part of dsMTL share the identical form,

$$\min_{\boldsymbol{\theta}} \boldsymbol{\mathcal{L}}(\boldsymbol{\theta}) + \boldsymbol{\lambda} \boldsymbol{S}(\boldsymbol{\theta}) + \boldsymbol{C} \boldsymbol{\aleph}(\boldsymbol{\theta})$$

where $\mathcal{L}(\theta)$ is the data fitting term (or loss function), the major determinant of the solutions obtained from model training. $\aleph(\theta)$ and $S(\theta)$ are the penalties of θ with the aim to incorporate the prior information. $\aleph(\theta)$ is a non-smooth function and able to create sparsity, while $S(\theta)$ is smooth. λ and C are the hyper-parameters to control the strength of the penalties. More technical details can be found in the supplementary methods.
In dsMTL, two approaches for sharing information across cohorts are included, 1) shared parameters and 2) cross-task regularization, leading to a slightly different distributed computation. The shared parameters are estimated using all cohorts. For cross-task regularization, the cohort-specific parameters are estimated using only the local data, and then tuned by considering parameters from other cohorts.
Efficiency
Most dsMTL methods aim at training an entire regularization tree. The determination of the λ sequence controls the tree's growth and is essential for computational speed. The λ sequence should be accurately scaled to both capture the highest posterior and avoid overwhelming computations. Inspired by a previous study(Friedman, Hastie et al. 2010), we estimate the largest and smallest λ from the data by characterizing the optima of the objective using the first-order optimal condition and then interpolate the entire λ sequence on a log scale (see supplementary methods for more details). In addition, several options are provided to improve the speed of the algorithms by decreasing the precision of the results, i.e., 1) the number of digits of parameters for transformation can be specified to reduce the network latency; 2) several termination rules are provided, some of which are relaxed; 3) the depth of the regularization tree can be shortened. More details can be found in supplementary methods.
Besides the efficiency of the federated ML/MTL methodology, the import/export of “big data” cohorts is also crucial for computational efficiency, where e.g. uncompressed GWAS data requires tens of gigabytes, leading to time-consuming data import. dsMTL was designed to support a wide variety of data types. For this, an architecture package resource (Marcon, Bishop et al. 2021) developed by the DataSHIELD community was incorporated to facilitate the efficient import and export of large-scale datasets in compressed formats. For example, in DataSHIELD, GWAS data of the PLINK file formats can be read and processed using the software PLINK(Purcell, Neale et al. 2007) as the backend(Marcon, Bishop et al. 2021).
Security

dsMTL was developed based on DataSHIELD(Wilson, Butters et al. 2017), which provides comprehensive security mechanisms not specific to machine learning applications. For example, 1) DataSHIELD requires the data analysis to only occur behind the firewall; 2) each server is only allowed to communicate with a set of clients with fixed IP addresses; 3) the network communication is protected by an SSL protocol; 4) an R parser(Wilson, Butters et al. 2017) implemented on the server rejects the calling of unwanted functions; and 5) the so-called 'disclosure control'(Wilson, Butters et al. 2017) on the server ensures that the returned response does not contain any disclosive information. In addition, several permissions can be set by the data providers to fully control the usage of their data. These permissions describe the degree of accessibility of data and functions on the server i.e. "*which users* can perform *what actions* on *what data"*. In an extremely secure example, a user could be granted to check the summary of a given dataset but cannot perform any actions because no functions were granted. With these settings, DataSHIELD allows customizing the security protection strategies according to the specific requirements of the applications. For statistical and machine learning analyses, DataSHIELD assumes that summary statistics are safe to share.

dsMTL inherits all these security mechanisms. In addition, we considered potential ML-specific privacy leaks, such as membership inference attacks(Hu, Salcic et al. 2021) and model inverse attacks(Fredrikson, Lantz et al. 2014). Inverse attacks aim at extracting the individual observation-level information from the models. Membership inference attempts to decide if an individual was included in a given training set using the model. All these techniques require a complete model for inference. Since multi-task learning returns multiple matrices, returning an incomplete model could be one strategy against these attacks. For example, dsMTL_iNMF in dsMTL only returns the homogenous matrix (H), whereas the cohort-specific components (V_k, W_k) never leave the server. For example, in a two-server scenario, one (H) out of five output matrices is transmitted between the client and the servers. With such an incomplete model, inverse construction of the raw data matrix becomes difficult, and the risk of an inverse attack and membership inference is reduced. For most biomedical analyses, the H matrix is sufficient for subsequent studies. In addition, if the analyst was authorized to access the raw data of the server, the so-called "data key mechanism" (see supplement) would allow the analyst to retrieve all component matrices. For supervised multi-task learning methods in dsMTL, all models have to be aggregated within the clients, and thus we suggest the data providers enable the option on the server that rejects a returned coefficient vector containing parameter numbers exceeding the number of subjects. In this way, the model is not saturated and more robust to an inverse attack.

Proof of concept with simulation and actual data

Two case studies and speed-tests were conducted to demonstrate the suitability of dsMTL methods to analyze heterogeneous cohorts, compared to federated ML methods and ensemble of local models regarding the prediction performance, interpretability and computational speed. An overview of methodological aspects related to the case studies is detailed below. For an extensive methodological description, please see the supplementary Methods.

Case study 1. In this case study, the heterogeneous cohorts were generated with the same set of outcome-associated genes. These however showed different directionality of their respective associations with the outcome. A three-server scenario was simulated. 150 out of 500 features with random signs across cohorts were simulated. Seven tests were created for simulating different n/p ($\frac{\text{sample size}}{\text{gene number}}$) ratios. The n/p ratio was $\{1.2, 1, 0.9, 0.6, 0.5, 0.3, 0.15\}$ with the number of subjects $\{600, 500, 450, 300, 250, 150, 75\}$ for each test. 500 genes were created for each server. The test sample consisted of 200 subjects for each server. Data were generated as follows:

Given gene number $\mathrm{p} = 500$, the models of three cohorts were $\{w^{(1)}, w^{(3)}, w^{(3)}\}$ where $w^{(.)} = \mathrm{p} \times 1$. A shared signature comprising 150 genes was generated for each $w^{(.)}$ but with random signs, ${w^{(.)}}_i = \begin{cases} 2 \times (\rho - 0.5) \times N(1, 0.1) & 1 < i < 150 \\ 0 & \text{others} \end{cases}$, $\rho \sim \mathrm{Bernoulli}(\frac{1}{2})$. The expression values of each subject across cohorts were generated as $\mathrm{x} = 1 \times \mathrm{p}$ where $x_j \sim N(0,1)$. The numeric outcome (e.g. symptom severity) $\mathrm{y} = \mathrm{x}\mathrm{w}^{(i)}$ in cohort i was standardized in a normal distribution $N(0,1)$, then model-irrelevant noise with 50% of the variance of the true signal was added $\mathrm{y} = \mathrm{y} + N(0, 0.5)$.
dsMTL_L21 and dsLasso were trained as the federated learning system, and the hyper-parameter was selected using 10 fold in-cohort cross-validation. For glmnet, the ensemble technique was only applied on the gene selection due to the consistent gene set of their signatures. The mean squared error (mse) was used as the measure of prediction performance. To account for the sampling variance, we repeated each analysis 100 times.

Case study 2. In this case study, two heterogeneous RNA-seq cohorts were created to simulate a comorbidity analysis, where the genes were separated to be part of either a shared signature among cohorts, cohort-specific signatures or diagnosis-unassociated genes. The dsMTL_iNMF was compared to the ensemble of local NMF regarding the selection accuracy of shared/cohot-specific genes, in particular impacted by the severity of heterogeneity. Here the severity of heterogeneity refers to the proportion of the genes harbored by the shared signature over all diagnosis-associated genes. The data simulation protocol for RNA-seq data can be found in the **Supplementary Methods**.
A two-server scenario was simulated. As shown in **Supplementary Table 1**, for the data of each server, 1000 genes and 200 subjects were simulated, 50% of the genes were diagnosis-unassociated and the remaining genes were part of the disease signature. The genes comprised by shared signatures were identical for data of two servers, and the genes comprised by cohort-specific signatures did not overlap. The case-control ratio was balanced for each server. Four tests were performed by varying the proportion of genes in the shared signature over all diagnosis-associated genes from 20% to 80%.
The training of dsMTL_iNMF results in three outputs related to the original input data: the shared gene 'exposure' (H), cohort-specific gene 'exposure' (V) and sample 'exposure' (W). We measured the association between the sample exposure and the diagnosis as the weight of each latent factor. The shared (or specific) gene signature was identified as the weighted summation of the shared (or specific) gene exposures over latent factors. To quantify the important genes related to a given signature, we binarized the gene signature according to the mean (0-1 vector, values larger than the mean were assigned). To assess the performance of the gene identification, we associated the selected genes set with the ground truth (0-1 vector, signature genes were 1). The assessment was applied to shared and cohort-specific genes in parallel. Based on this metric, three gene sets were derived as output from dsMTL_iNMF, called dsMTL_iNMF-H, dsMTL_iNMF-V1 and dsMTL_iNMF-V2, and these related to the shared, cohort 1 specific and cohort 2 specific gene signature, respectively. The same strategy was applied to analyze the ensemble of local NMF models. For each cohort, the specific gene signature was the weighted summation of gene exposure over latent factors, and then binarized as the specific gene set (called local-NMF1 and local-NMF2). The shared gene signature was identified as the sum of the specific gene signature over cohorts, and then binarized as the shared gene set (NMF-bagging). We then compared 1) NMF-bagging and dsMTL_iNMF-H for the accuracy related to the isolation of shared genes; 2) dsMTL_iNMF-V1 and local-NMF1 as well as dsMTL_iNMF-V2 and local-NMF2 for the accuracy of isolating cohort-specific genes.

Computational speed of supervised dsMTL. We aimed at identifying the efficiency of supervised dsMTL using real molecular data and given the real network latency. Four independent schizophrenia case-control cohorts were used for this analysis. The training cohorts consisted of three datasets

comprising prefrontal cortex gene expression data (available from the GEO repository under accession numbers GSE53987, GSE21138 and GSE35977). A detailed description of these datasets can be found in their respective original publications (Tang, Capitao et al. 2012, Chen, Cheng et al. 2013, Lanz, Reinhart et al. 2019). The dataset used for algorithm testing was from the HBCC (n=422) cohort comprising genome-wide gene expression data quantified by microarray (dbGAP ID: phs000979.v3.p2). A detailed description of this dataset can be found in the original publication (Fromer, Roussos et al. 2016). As shown in **Supplementary Table 2**, three servers were used for training algorithms. Two servers were held at the Central Institute of Mental Health, Mannheim while the third was positioned at the BioQuant institute, Heidelberg.

Using this data, we repeated a previously described analysis(Cao, Meyer-Lindenberg et al. 2018), in order to evaluate computational speed in a federated analysis setting. Here we show the formulation of the mean regularized MTL using dsMTL_net:

The cohort-level batch effect was assumed to be Gaussian noise affecting the true coefficient of gene i and cohort j $w_{ij} = w_i + \epsilon_j$, $\epsilon_j \in N(\mu, \sigma)$. Hence, the average model $\overline{w_i}$ across cohorts was an unbiased estimator for the true coefficient, and therefore the squared penalty $\left|w_{ij} - \overline{w_i}\right|^2$ was incorporated to penalize the departure of each model j to the mean. The complete formulation was

$$\min_W \sum_{k=1}^{3} \sum_{i=1}^{n_k} \frac{1}{n_k} \log\left(1 + e^{-Y_i^{(k)}\left(X_i^{(k)} W_{,k}\right)}\right) + \lambda ||W||_1 + C||WG||_2^2,$$

$$\text{where } G = \begin{bmatrix} \frac{2}{3} & 0 & \frac{-1}{3} & \frac{2}{3} & \frac{-1}{3} & 0 \\ \frac{-1}{3} & \frac{2}{3} & 0 & 0 & \frac{2}{3} & \frac{-1}{3} \\ 0 & \frac{-1}{3} & \frac{2}{3} & \frac{-1}{3} & 0 & \frac{2}{3} \end{bmatrix}$$

Computational speed of unsupervised dsMTL. Here, we analyzed the time efficiency in applying dsMTL_iNMF on two real datasets based on the real network latency. Two processed RNA-seq case-control cohorts comprising patients with schizophrenia (GSE164376(A; and R; 2021)) and bipolar disorder (GSE134497(Kathuria, Lopez-Lengowski et al. 2020)) were retrieved from the GEO database and converted into a matrix format for the analysis. As shown in **Supplementary Table 4**, the data were stored on servers in Mannheim and Heidelberg.

Figures

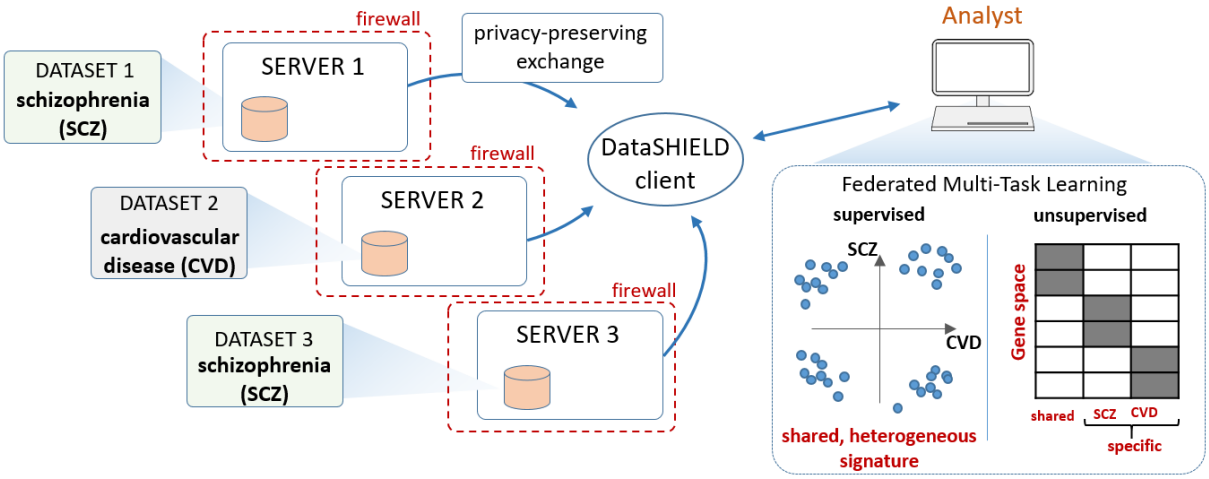


Figure 1. Schematic illustration of dsMTL using comorbidity modeling of schizophrenia and cardiovascular disease as an example. Multiple datasets stored at different institutions are used as a basis for federated MTL. dsMTL was developed in the DataSHIELD ecosystem, which provides

functionality regarding data management, transmission and security. Data are analyzed behind a given institution's firewall and only algorithm parameters that do not disclose personally identifiable information are exchanged across the network. dsMTL contains algorithms for supervised and unsupervised multi-task machine learning. The former aims at identifying shared, but potentially heterogeneous signatures across tasks (here, diagnostic classification for schizophrenia and cardiovascular disease). Unsupervised learning separates the original data into shared and cohort-specific components, and aims at revealing the corresponding outcome-associated biological profiles.

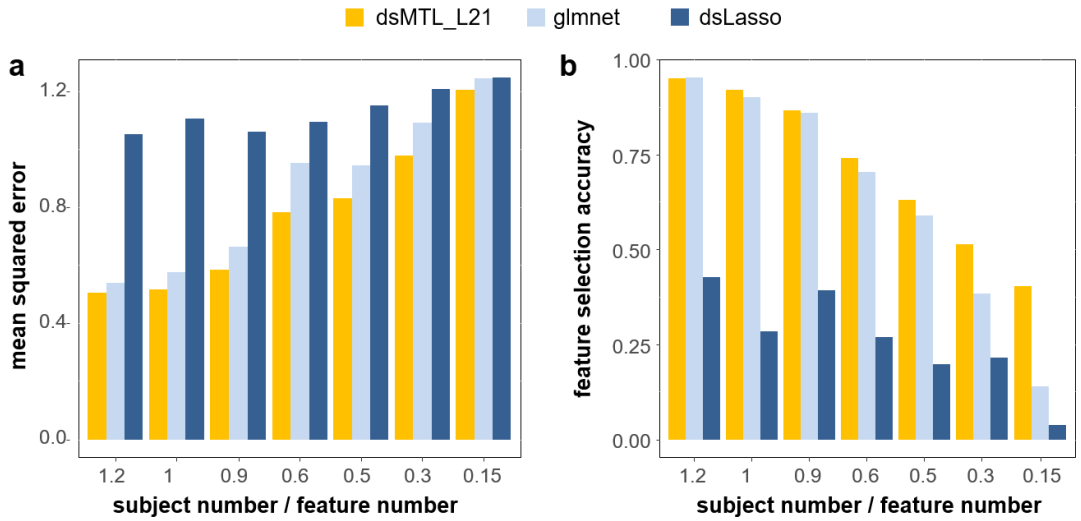


Figure 2. Analysis of 'heterogeneous' signatures of continuous outcomes in simulated data stored on three servers. The figure shows the **a)** prediction accuracy expressed as the mean squared error and **b)** the feature selection accuracy for different subject/feature number ratios. The respective values were averaged across the three servers, and across 100 repetitions, in order to account for the effect of sampling variability.

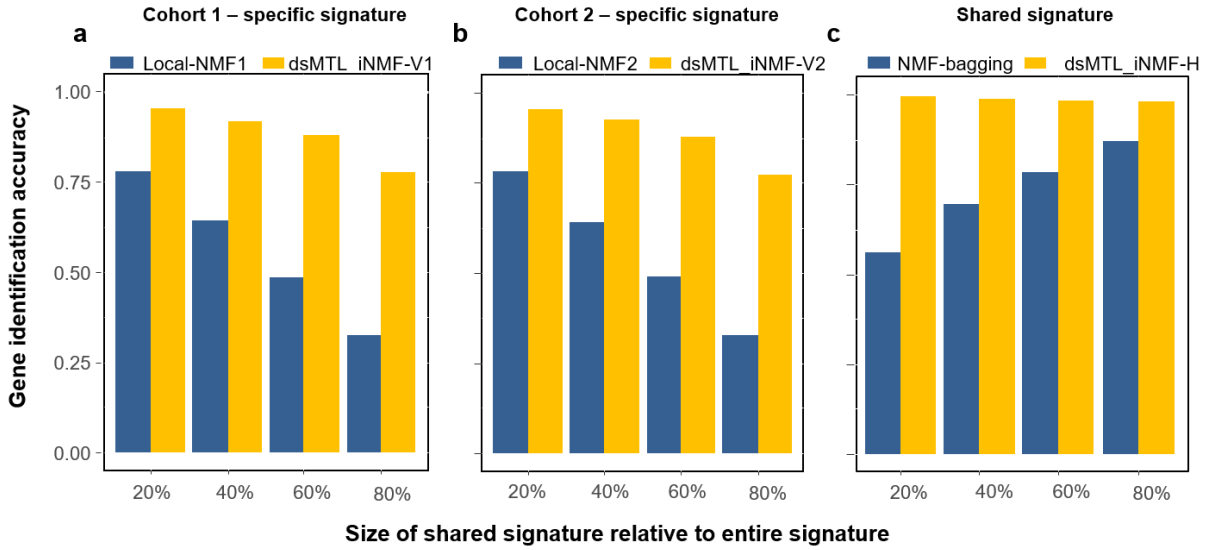


Figure 3. The gene identification accuracy for shared and specific signatures using simulated data. a) the identification accuracy of important genes for cohort 1. **b)** the identification accuracy of important genes for cohort 2. **c)** the identification accuracy of genes comprised in the shared signature**.** Local-NMF1 and Local-NMF2 were the cohort-specific gene sets identified by local NMF, which were combined into "NMF-bagging" for the shared gene set. dsMTL_iNMF-H was the predicted shared gene

set using dsMTL_iNMF. dsMTL_iNMF-V1 and dsMTL_iNMF-V2 were the predicted cohort-specific gene sets identified using dsMTL_iNMF (see Supplementary Figure 1). The proportion of genes harbored by the shared signature was varied from 20% to 80% illustrating the impact of the heterogeneity severity. The model was trained using rank=4 as model parameter. The results for a broader spectrum of rank choices can be found in **Supplementary Figure 2** illustrating that the superior performance of dsMTL_iNMF was not due to the choice of ranks.

4.6 Supplements

4.6.1 Supplementary methods

dsMTL framework

In dsMTL, we included four federated multi-task (FeMTL) and one machine learning (FeML) methods covering supervised and unsupervised learning procedures. All models followed the consistent formulation,

$$\min_{\theta} \mathcal{L}(\theta) + \lambda S(\theta) + C\aleph(\theta) \quad (1)$$

$\mathcal{L}(\theta)$ was the data fitting term (or loss function), the major determinant of the solutions of the model training. $\aleph(\theta)$ and $S(\theta)$ were the regularization/penalty terms with the aim to incorporate the prior information and prevent overfitting. $\aleph(\theta)$ was a non-smooth function for creating the sparsity, while $S(\theta)$ was smooth with the ability to stabilize the solution. λ and C were the hyper-parameters to control the strength of the penalty, λ was learnt from cross-validation (CV) and C was the constant.

There are three loss functions in dsMTL, achieving the tasks of regression, classification and matrix factorization. They are summarized in **Supplementary Table 1**.

	Unsupervised Learning	Supervised Learning	
	Matrix factorization	Regression	Classification
Model	$[X_1, \dots, X_k, \dots X_t] = [(H + Hv_1) \times W_1, \dots, (H + Hv_t) \times W_t])$	$f(x) = xw$	$P(x) = \frac{1}{1 + e^{-(xw)}}$
Loss function	$\min_{\substack{H, \\ W_1 \dots, W_K, \\ V_1 \dots, V_K}} \sum_{k=1}^{t} \|\|X_k - (H + V_k)W_k\|\|_F^2$	$\min_w \frac{1}{2n} \sum_{i=1}^{n} \|\|y_i - x_i w\|\|^2$	$\min_w \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i(x_i w)})$
Gradient	$\nabla_{H_{i,j}} = 2\sum_{k=1}^{t}(H_{i,j}W_{k_i}W_{k_i}^T - X_{k_i}W_{k_i}^T)$ $\nabla_{W_{k_{i,j}}} = 2\sum_{m=1}^{n_t}(HV_k)_{m,i}\left[(HV_k)_{m,i}W_{k_{ij}} - X_{k_{m,j}}\right]$ $\nabla_{V_{k_{i,j}}} = 2(V_{i,j}W_{k_i}W_{k_i}^T - X_{k_i}W_{k_i}^T)$	$\nabla_w = \frac{1}{n}(x^T xw - x^T Y)$	$\nabla_w = -\frac{1}{n}X^T \times \begin{bmatrix} \frac{y_1}{1 + e^{y_1(x_1 w)}} \\ \dots \\ \frac{y_n}{1 + e^{y_n(x_n w)}} \end{bmatrix}$

Supplementary Table 1. Summaries of loss functions used in dsMTL

Termination rules Four termination rules were included in dsMTL to determine whether the optimization converges. The first three rules were applied to all methods in dsMTL, while the last was new designed for matrix factorization. The first rule checked whether the current objective value was close enough to 0. The second rule investigates the last two objective values and checks whether the decrement was close enough to 0. The third rule allowed the optimization to be performed for a certain

maximum number of iterations. The last rule specific to matrix factorization was described in the next section.

Federated unsupervised method in dsMTL

To discover the hidden structure in heterogeneous, high-dimensional biological data, we integrated the integrative matrix factorization method(Yang and Michailidis 2016) (iNMF) in our distributed learning framework, called dsMTL_iNMF. The major concept of dsMTL_iNMF is shown in **Supplementary Figure 1**, where the cohort matrices on three servers can be factorized simultaneously with the shared component matrix (H) and cohort-specific component matrices(V, W). The objective function was

$$\min_{\substack{H,\\ W_1 \dots, W_K,\\ V_1 \dots, V_K}} \sum_{k=1}^{t} ||X_k - (H + V_k)W_k||_F^2 + \lambda \sum_{k=1}^{t} ||V_k W_k||_F^2 + \lambda_s \sum_{k=1}^{t} |W_k|_1$$

The robustness of the model was due to the decoupled setting of H and V_k, where H was to capture the shared information across cohorts and V_k was to capture the cohort-specific information. To integrate the more information into shared component H, the magnitude of cohort-specific component was penalized $\aleph(.) = \sum_{k=1}^{t} ||V_k W_k||_F^2$. The sparse term $S(.) = \sum_{k=1}^{t} |W_k|_1$ was used to remove the redundant coefficients from the component matrices.

Distributed variables update

$$W_{k_{ij}} \leftarrow W_{ij} \frac{\left((H+V_k)^T X_k\right)_{i,j}}{\left((H^T H + H V_k{}^T + H^T V_k + (1+\lambda) V_k{}^T V_k) W_k\right)_{i,j} + \lambda_s} \quad (2)$$

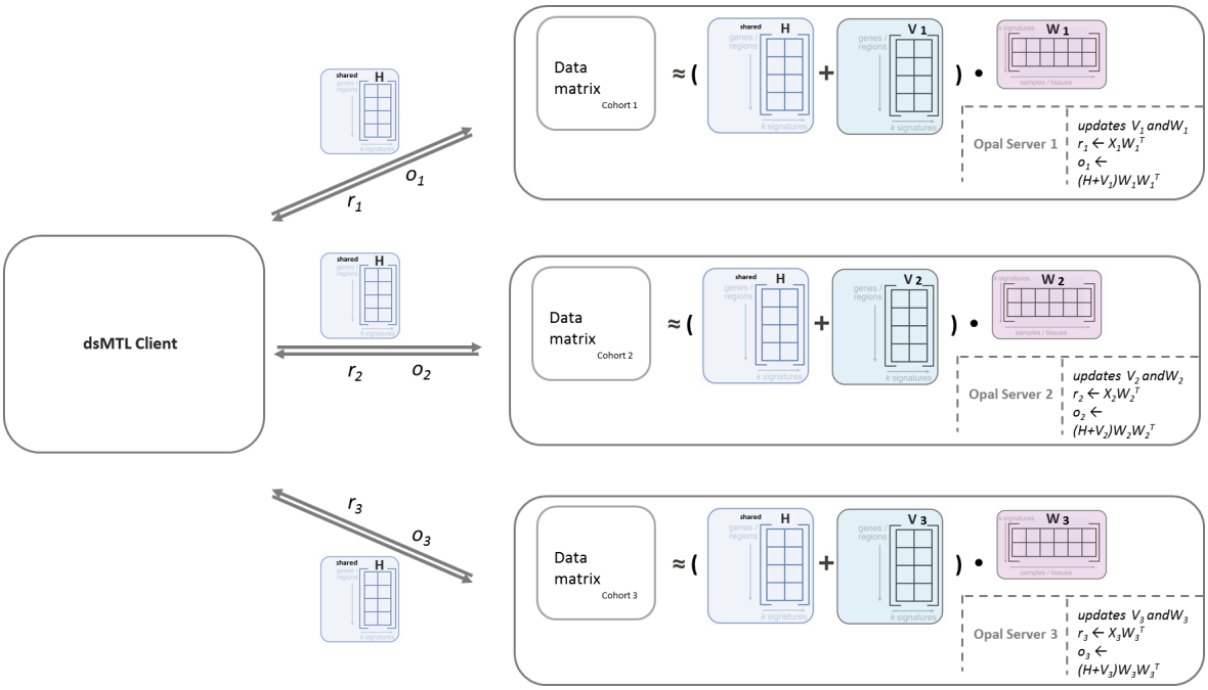
$$V_{k_{ij}} \leftarrow V_{ij} \frac{\left(X_k W_k^T\right)_{i,j}}{\left(H W_k W_k^T + (1+\lambda) V_k W_k W_k^T\right)_{i,j} + \lambda_s} \quad (3)$$

$$H_{ij} \leftarrow H_{ij} \left(\frac{X_1 W_1^T + \cdots + X_t W_t^T}{(H+V_1) W_1 W_1^T + \cdots + (H+V_t) W_t W_t^T} \right)_{i,j} \quad (4)$$

The variables were updated for non-federated applications as demonstrated in formulas (2) to (4). In the federated scenario, the cohort-specific variables W_k and V_k were updated on server k using the local data as formulas (2) and (3). The shared matrix H was updated on the client after receiving summary data (see **Supplementary Figure 1**) from all servers, where these were not-disclosed and calculated behind a given institution's firewall. The distributed update of H was

$$H_{ij} \leftarrow H_{ij} \left(\frac{server_1(X_1 W_1^T) + \cdots + server_t(X_t W_t^T)}{server_1((H+V_1) W_1 W_1^T) + \cdots + server_t((H+V_t) W_t W_t^T)} \right)_{i,j} \quad (5)$$

After the aggregation, the client updates H and a new iteration begin. The communications between the client and the servers are illustrated in **Supplementary Figure 1**.



Supplementary Figure 1. Communications between the client and opal servers for dsMTL_iNMF

Algorithms

Distributed solver

For the privacy-preserving purpose, only the shared matrix H was returned. The distributed solver of dsMTL_iNMF is shown in Algorithm 1.

Algorithm 1, Solver of distributed iNMF in dsMTL

Input: $\lambda > 0, \lambda_s > 0, \text{maxIter} > 0, H, W_1 \dots, W_K, V \dots, V_K$

Output: H

1: **for** $i = 1$ to maxIter **do**

2: Update H according to (5) on client

3: Send H to all servers

3: Update $W_1 \dots, W_K$ according to (2) on server $1, \dots, k$

4: Update $V_1 \dots, V_K$ according to (3) on server $1, \dots, k$

5: Send summary statistics back client according to (5)

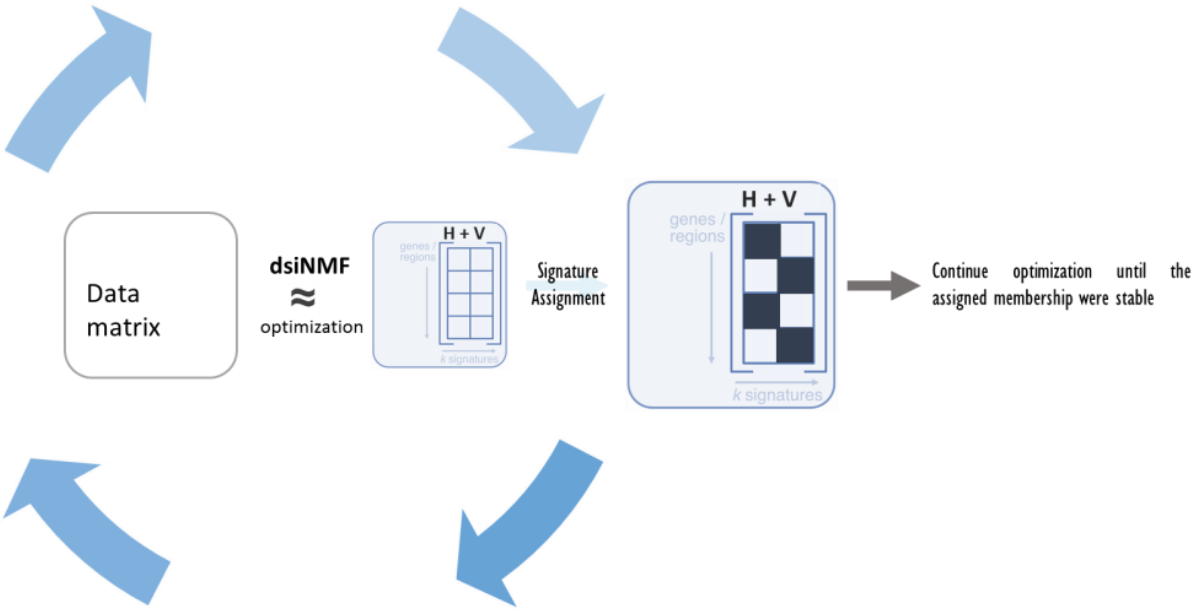
6: If termination rule satisfied, **return**

6: **end for**

Termination rules

For dsMTL_iNMF, we provided an additional termination rule developed in the ButchR package(Quintero, Hubschmann et al. 2020) to determine the convergence of the algorithm. In this method, each of the samples was assigned to a hidden factor (clustering membership) by $j = \arg\max_j |H_{i,j}|$ at every iteration. The convergence was determined when the assignments of samples

remained unchanged. By default, if the samples were assigned to the same hidden factors consistently for over 10 iterations, the memberships were seen as stable, and the algorithm stopped. The rationale behind this procedure is that in order to maximize the power of clustering, the variance of the determined memberships must be small. Therefore, the proposed rule terminates the algorithm when the samples found stable memberships, such that the clustering can make a stable decision.



Supplementary Figure 2. Schematic illustration of cluster membership optimization.

Distributed model training

In the default setting, Algorithm 1 was performed with 10 random initial points to approximately sample the distribution of the local optima considering the non-convex nature of the problem. The initialization of these component matrices were uniformly sampled from $[0, 2]$. For each initialization, Algorithm 1 was performed. A set of shared matrices were returned as the final results for subsequent analysis

Algorithm 2 Training procedure of iNMF in dsMTL
Input: $\lambda > 0$, $\lambda_s > 0$, maxIter > 0, rank, nInitialization, $\{X_1, \dots X_k, \dots X_t\}$
Output: H_1, H_2,…
1: **for** $i = 1$ to nInitialization **do**
2: Initialize $H_i \sim U_{n \times rank}(0,1)$, for each k, $V_k \sim U_{n \times rank}(0,1)$, $W_k \sim U_{rank \times p_k}(0,1)$
3: H_i =**Algorithm 1** $(\lambda = \lambda, \lambda_s = \lambda_s, H = H_i, \{W_1, \dots, W_t\}, \{V_1, \dots, V_t\})$
4: **end for**

Federated supervised methods in dsMTL

We included one machine-learning (ML) and three multi-task learning (MTL) algorithms into supervised methods of dsMTL. In the federated scenario, the ML model was trained by averaging the summary statistics from geo-distributed cohorts with the synchronous communication, which leads to a model equivalent to the standalone ML model training on the concatenated cohorts. MTL, in the federated scenario, exchanges a small amount of information by regularization, such that the commonality of multi-cohort models was reinforced, but the cohort-specific element remained unchanged. In dsMTL, dsLasso was included as the federated ML variant of Lasso(Tibshirani 2011). The federated multi-task learning methods were adapted from our previously published package RMTL(Cao, Zhou et al. 2018). These methods adopted various strategies of cross-cohort regularization to select joint features, explore the low-rank structure and incorporate the network structure in multiple tasks.

In the next sections, we introduce the theoretical derivations of each method to form the federated algorithm. Then a federated optimization framework is derived that was applied to all supervised methods. At the end, we show an executable algorithms for solving the objective and training the model.

Models

For each method, the objective function is first introduced as the major problem to solve. Second, we derive the subgradient of the objective to characterize the properties of the optima. Third, since all models are sparse, we aimed to solve the entire regularization tree(Zou and Hastie 2005) with a given positive λ sequence in decreasing order. The λ_{max} , the estimate of largest λ in the λ sequence, was derived from the subgradient, such that λ_{max} was the smallest λ guaranteeing the existence of 0 optima. Last, to solve the non-smooth objective efficiently, the proximal mapping was applied, and the proximal point estimator was derived as the solution of the iteration-level sub-problem (see next section).

Lasso (dsLasso)

Objective function:

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(w; X, Y) + \lambda |w| + C||w||_2^2$$

Subgradient:

$$\partial_w = \nabla_w \mathcal{L} + 2Cw + \lambda \frac{w}{|w|}$$

Estimated λ_{max}:

$$\lambda_{max} = \begin{cases} \frac{1}{n} \max_j |X_{,j}^T Y| & for\ least\ square\ loss \\ \frac{1}{2n} \max_j |X_{,j}^T Y| & for\ logit\ loss \end{cases}$$

Proximal point estimation:

$$\underset{f=\frac{\lambda|x|}{L}}{\text{prox}}(w) = \text{sign}(w)\max\{w - \lambda, 0\}$$

The Lasso model(Tibshirani 2011) aimed to learn a sparse parameter vector w. λ was identified by cross-validation. $||w||_2^2$ was used to stabilize the solution and incorporate the correlated features. C was selected by the user.

MTL with Feature selection (dsMTL_L21)

Objective function

$$\min_W \sum_{k=1}^{t} \sum_{i=1}^{n_k} \frac{1}{n_k} \mathcal{L}\left(W_{,k}; X_i^{(k)}, Y_i^{(k)}\right) + \lambda \sum_{j=1}^{p} \sqrt{||W_{j,}||_2^2} + C||W||_2^2$$

Subgradient:

$$\partial_{W_{j,}} = \nabla_{W_{j,}} \mathcal{L} + 2CW_{j,} + \lambda v,\ v = \{x \in R^t : ||x||_2 \leq 1\}$$

Estimated λ_{max}:

$$\lambda_{max} = \begin{cases} \max_j \sqrt{\sum_{k=1}^{t} \left(\frac{< X_{,j}^{(k)}, Y^{(k)} >}{n_k} \right)^2} & for\ least\ square\ loss \\ \max_j \sqrt{\sum_{k=1}^{t} \left(\frac{< X_{,j}^{(k)}, Y^{(k)} >}{2n_k} \right)^2} & for\ logit\ loss \end{cases}$$

Proximal point estimation:

$$\underset{f=\frac{\lambda||x||_2}{L}}{\text{prox}}(w) = \left(1 - \frac{\lambda}{\max\left\{||w||_2, \lambda\right\}}\right) w$$

The method(Liu, Ji et al. 2009) aimed to find a model with the same set of features. For this, $\sum_{j=1}^{p} \sqrt{||W_{j,}||_2^2}$ was used to penalize the magnitudes of the coefficients of a given feature across the datasets. The $W = p \times t$ was the solution matrix of t tasks and p features.

MTL with low-rank structure(dsMTL_Trace)

Objective function:

$$\min_W \sum_{k=1}^{t} \sum_{i=1}^{n_k} \frac{1}{n_k} \mathcal{L}\left(W_{,k}; X_i^{(k)}, Y_i^{(k)}\right) + +\lambda||W||_* + C||W||_2^2$$

Subgradient:

$$\partial_W = \nabla_W \mathcal{L} + 2CW + \lambda \partial ||W||_*$$

Estimated λ_{max}:

$$\lambda_{max} = \begin{cases} \max_j \sigma_1\left(\left[\frac{X^{(1)^T}Y^{(1)}}{n_1}, \dots, \frac{X^{(t)^T}Y^{(t)}}{n_t}\right]\right) & for\ least\ square\ loss \\ \max_j \sigma_1\left(\left[\frac{X^{(1)^T}Y^{(1)}}{2n_1}, \dots, \frac{X^{(t)^T}Y^{(t)}}{2n_t}\right]\right) & for\ logit\ loss \end{cases}$$

Where $\sigma_1(A)$is the largest singular value of matrix A

Proximal point estimation:

$$\underset{f=\frac{\lambda||x||_*}{L}}{\text{prox}}(W) = U \times I_{\max\{\sigma-\lambda,0\}} \times V,$$

where $W = U\Sigma V$, σ is the diagonal vector of Σ

The method(Pong, Tseng et al. 2010) aimed to identify the coefficient vectors of multiple cohorts existing in the compressed low-dimensional space. For this, the trace norm of the coefficient matrix was used to compress the models' space.

MTL with network structure(dsMTL_Net)

Objective function

$$\min_W \sum_{k=1}^{t} \sum_{i=1}^{n_k} \frac{1}{n_k} \mathcal{L}\left(W_{,k}; X_i^{(k)}, Y_i^{(k)}\right) + \lambda||W||_1 + C||GW||_2^2$$

Subgradient:

$$\partial_W = \nabla_W \mathcal{L} + 2CGG^T + \lambda \frac{W}{|W|}$$

Estimated λ_{max}:

$$\lambda_{max} = \begin{cases} \max_{j,k} \frac{X_{,j}^{(k)^T} Y^{(k)}}{n_k} & for\ least\ square\ loss \\ \max_{j,k} \frac{X_{,j}^{(k)^T} Y^{(k)}}{2n_k} & for\ logit\ loss \end{cases}$$

Where $\sigma_1(A)$is the largest singular value of matrix A

Proximal point estimation:

$$\underset{f=\frac{\lambda|x|}{L}}{\text{prox}}(W) = \text{sign(W)}\max\{|W| - \lambda, 0\}$$

The method aimed to incorporate the relationships between cohorts as a graph into the model training procedure. $||GW||_2^2$ was used for this aim, where G was an pre-defined matrix describing the network structure. More details of G for variant applications can be found in (Cao and Schwarz). $||W||_1$ was used to remove redundant coefficients. λ was identified by cross-validation.

Distributed Optimization procedure

To solve these composite objective functions efficiently in the same framework, we rewrite the objective (1) as

$$\min_x F(x) + \lambda\Omega(x) \tag{9}$$

where $F(x) = \mathcal{L}(\theta) + CS(\theta)$ was smooth component function and $\Omega(x) = \aleph(\theta)$ was non-smooth

Solving sub-problem in each iteration

Given the Lipschitz constant L of the objective function above, the sequence of estimation points $\{x_0, x_1, x_2, \dots\}$ were found by solving the below iteration-wise sub-problem ((Beck and Teboulle 2009))

$$x_{i+1} = \arg\min_y \mathcal{M}_{L,x_i}(y) \tag{10}$$

$$\mathcal{M}_{L,x_i}(y) = F(x_i) + \langle \nabla F(x_i), \mathrm{y} - x_i \rangle + \frac{L}{2}\left|\left|y - x_i\right|\right|_2^2 + \lambda\Omega(x) \tag{11}$$

The first three terms on the right side were the second order approximation of $F(.)$ using Tylor expansion on point x_i. After re-organization, we have (12) equal to (10).

$$x_{i+1} = \arg\min_y \frac{L}{2}\left(y - \left(x_i - \frac{\nabla F(x_i)}{L}\right)\right)^2 + \lambda\Omega(x) \tag{12}$$

Since $x_i - \frac{\nabla F(x_i)}{L}$ was known after the ith iterations, the above problem was applicable to the proximal algorithm framework(Parikh and Boyd 2014). For all sparse regularizations ($\Omega(x)$) used in dsMTL, they can be simplified and solved analytically in (13), and the results were derived and summarized above (see the "Proximal point estimation") for each dsMTL method.

$$x_{i+1} = \underset{f=\frac{\lambda\Omega(x)}{L}}{\mathrm{prox}}\left(x_i - \frac{\nabla F(x_i)}{L}\right) \tag{13}$$

Line search

Since L was unknown in our framework, we estimated it using a backtracking line search approach. Set an increasing sequence of $L \in \{L_0, 2L_0, 4L_0, 16L_0, \dots\}$given an constant L_0, the smallest L satisfying the condition $\mathcal{M}_{L,x_i}(x_{i+1}) \geq F(x_{i+1})$ was selected. Here, x_{i+1} was determined based on (12).

Federated computation

For supervised MTL, the variable matrix $\boldsymbol{W} = \boldsymbol{p} \times \boldsymbol{t} = \left[\boldsymbol{w_{,1}}, \boldsymbol{w_{,2}}, \dots, \boldsymbol{w_{,t}}\right]$, where each column represents one task. So distributed proximal operator was:

$$W_{i+1} = \underset{f=\frac{\lambda\Omega(x)}{L}}{\mathrm{prox}}\left(W_i - \frac{\nabla F(W_i)}{L}\right) = \underset{f=\frac{\lambda\Omega(x)}{L}}{\mathrm{dist\,prox}}\left(\left[w_{,1_i} - \frac{\nabla F\left(w_{,1_i}; D^{(1)}\right)}{L}, \quad \dots, \quad w_{,t_i} - \frac{\nabla F\left(w_{,t_i}; D^{(t)}\right)}{L}\right]\right) \tag{14}$$

where $w_{,k_i} - \frac{\nabla F\left(w_{,k_i}\right)}{L}$ was calculated on server k and sent back. $\left\{D_{1,\dots}, D_t\right\}$ represented the data on t servers. Similarly, objective function $O(W)$ has to be evaluated in a distributed fashion,

$$O(W) = \text{dist } O(W) = \sum_{k=1}^{t} F\left(W_{,k}; D^{(k)}\right) + \lambda\Omega(W)$$

For supervised ML, the information aggregation was different. The variable vector $\boldsymbol{w} = \boldsymbol{p} \times \mathbf{1}$. The distributed proximal operator is

$$w_{i+1} = \underset{f=\frac{\lambda\Omega(x)}{L}}{\text{prox}}\left(w_i - \frac{\nabla F(w_i)}{L}\right) = \underset{f=\frac{\lambda\Omega(x)}{L}}{\text{dist prox}}\left(w_i - \frac{1}{L}\left(\sum_{j=1}^{t} \nabla\mathcal{L}\left(w_i; D^{(j)}\right)\frac{n_t}{n} + C\nabla\aleph(w_i)\right)\right)$$

where $\nabla\mathcal{L}\left(w_i; D^{(j)}\right)$ was calculated on server j and sent back. Similarly, objective function $O(w)$ has to be evaluated in distributed federated fashion,

$$O(w) = \text{dist } O(w) = \sum_{j=1}^{t} \mathcal{L}\left(w_i; D^{(j)}\right)\frac{n_j}{n} + C\aleph(w_i) + \lambda\Omega(w).$$

Accelerated algorithms

Distributed solver

To accelerate the optimization procedure, we applied Nesterov's acceleration approach(Liu and Jieping , Beck and Teboulle 2009, Nesterov 2012). In the beginning of iteration i, the search point was first defined as the weighted combination of the results from the previous two steps: $\mathrm{S_i} = \frac{\alpha_{i-1}}{\alpha_i}\mathrm{x_i} + \frac{1-\alpha_{i-1}}{\alpha_i}\mathrm{x_{i-1}}$. Then the formulas (13) was applied on the $\mathrm{S_i}$.

Algorithm 3 Distributed solver of supervised learning methods in dsMTL

Input: $\lambda > 0$, $L_0 > 0$, W_0, $maxIter > 0$

Output: W_{i+1}

1: Initialize $W_1 = W_0$, $\alpha_{-1} = \alpha_0 = 0$, and $L = L_0$

2: **for** $i = 1$ to $maxIter$ **do**

3: Set $\mathrm{S_i} = \mathrm{W_i} + \frac{\alpha_{i-1}-1}{\alpha_i}(\mathrm{W_i} - \mathrm{W_{i-1}})$

4: Find smallest $L \in \{L_{i-1}, 2L_{i-1}, 4L_{i-1}, 16L_{i-1}, \dots\}$ such that

$$\mathcal{M}_{L,x_i}(W_{i+1}) \geq \text{dist } O(W_{i+1}),$$

where $W_{i+1} = \underset{f=\frac{\lambda\Omega(x)}{L}}{\text{dist prox}}\left(W_i - \frac{\nabla F(W_i)}{L}\right)$

5: Set $L_i = L$, and $\alpha_{i+1} = \frac{1+\sqrt{1+4\alpha_i^2}}{2}$

6: If termination rule satisfied, **return**

7: **end for**

Training for sparse model

In the high-dimensional data analysis, the performance of sparse ML models was highly related to the accuracy of sparse structure identification, thus λ selection was crucial. In dsMTL, we trained the entire regularization tree for a given hyper-paratemer C. Similar to the study(Friedman, Hastie et al. 2010), we estimated the λ_{max} as the largest λ of the sequence from the data. λ_{max} was selected by looking for the smallest λ such that the equation $\partial_W\big(F(x) + \lambda\Omega(x)\big) \ni \mathbf{0}$ hold. Due to the differential objective functions, λ_{max} of calssification model was different from that of regression model. The λ_{min} was determined based on λ_{max}, i.e. $0.1\lambda_{max}$. Then the entire sequence was interpolated based on the log scale of λ_{max} and λ_{min}. For each method, λ_{max} was therotically different, and summarized above.

Algorithm 4 Training procedure of sparse models in dsMTL

Input: $\lambda_1 > \lambda_2 > \cdots > 0$

Output: W_1, W_2,…

1: Initialize $W_0 = p \times t = 0$

2: **for** $i = \{1, 2, \dots\}$ **do**

3: $W_i =$**Algorithm 3** $(\lambda = \lambda_i, L_0 = 1, W_0 = W_{i-1}, maxIter = 100)$

4: **end for**

Cross-validation

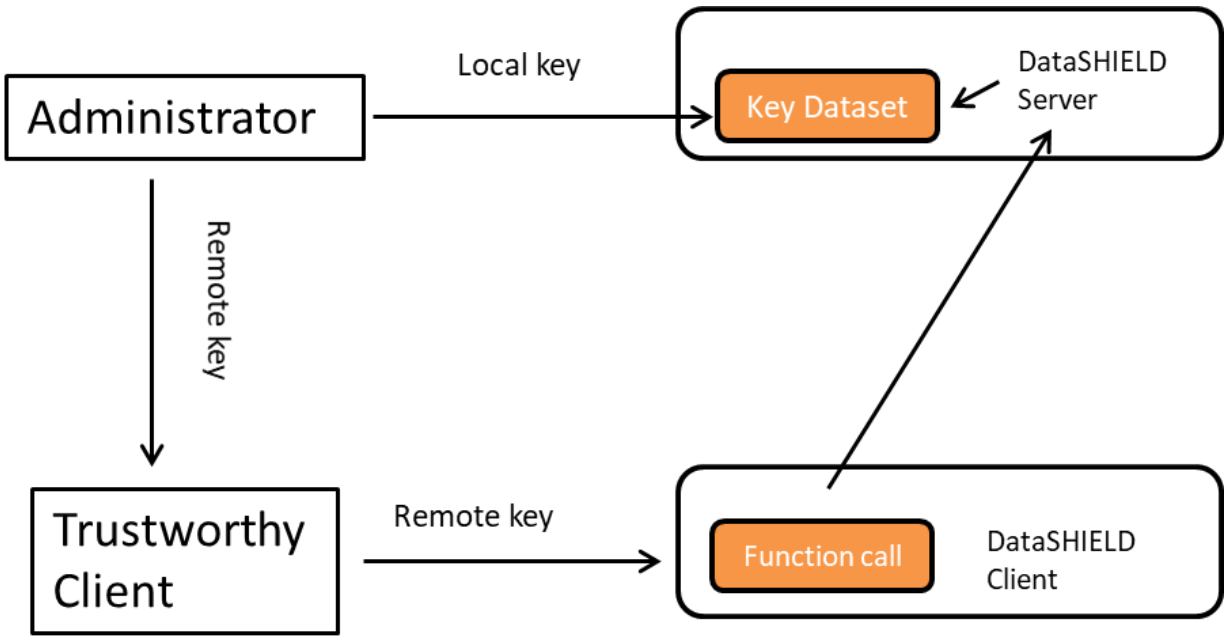
We set up cross-cohort and in-cohort CV in dsMTL for all ML/MTL methods. For cross-cohort CV, t folds CV were established for *t* cohorts. In fold *i*, cohort *i* was as the test cohort and the model was trained on rest cohorts. The prediction performances were averaged and used to select λ. Such CV aimed to identify a λ with an optimized generalization performance. For k-folds in-cohort CV, the samples of each cohort were randomly separated into k folds, such that the test folds across cohorts were combined for testing, and the training folds were combined for training. Such CV aimed to identify a λ with the most representative sparse model across all cohorts.

Introduction of DataSHIELD

DataSHIELD(Wilson, Butters et al. 2017) is a platform software supporting federated data analysis without disclosing personally identifiable information. Two modules were included, the Ranalytic environment and the data warehouse opal. To mitigate the risk of sensitive data disclosure, the design of DataSHIELD considers to broader aspects: software architecture and statistics. The architecture of DataSHIELD provides several non-disclosure mechanisms to improve the system security, such as 1) data analysis only occurs behind the firewall; 2) each server is only allowed to communicate with a single client of a fixed IP; 3) the network communication was protected by the SSL protocol; 4) an R parser was implemented on the DataSHIELD server to reject unpermitted behaviors. A comprehensive set of permission settings were provided for data providers to fully control access to their data. These permissions were about users, data and functions for characterizing i.e. "*which users* could perform *what behaviors* to *what data*". In an extremely secure example, a user could be granted to access a dataset but cannot perform any actions because no functions were granted. With these settings, DataSHIELD allows to customize the security strategies according to the specific requirement of the applications. From a statistical perspective, DataSHIELD assumes sharing summary statistics are safe

for privacy-preserving applications. Such assumption is quite common in the biomedical field, and there is a large number of websites providing summary data for free download, such as the GWAS summary data of certain traits(Zheng, Erzurumluoglu et al. 2017) and eQTLs of tissues(Consortium 2015). Another study(Jones, Sheehan et al. 2012) confirmed the non-disclosure property of DataSHIELD for regression analysis from a biostatistical perspective.

Data key mechanism



Supplementary Figure 3. Schematic illustration of the data key mechanism

This mechanism allows the authorized users to obtain the complete model identified by multi-task learning from the server. The administrator generates two keys, stores the local key in the key database, and gives the remote key to a trustworthy user. Then by sending the remote key, the client is seen as the data provider of the server, and can retrieve the complete model from the multi-task learning method.

This mechanism was built for two reasons. 1) The custom-defined functions in DataSHIELD cannot obtain identity information from the users, and 2) specifying the identity of users via the IP address is not sufficiently safe.

Generate RNA-seq count data for case study 2

The RNA-seq count data was generated using the Negative Binomial distribution (NB distribution), which was the most common distribution used to model RNA-seq data. In case study 2, a two-cohort scenario was simulated. Four tests were conducted for different severity of heterogeneity. Here the severity of heterogeneity was characterized by the proportion (20%, 40%, 60% and 80%) of genes in the shared signature over all diagnosis-associated genes. The simulation procedure contained the following steps. First, the background data of two cohorts X_1 and X_2 were generated as sampled from the NB distribution $X_1 \sim \mathrm{NB}_{p\times n_1\times}(\mathrm{r}=2,\mathrm{p}=0.3)$ and $X_2 \sim \mathrm{NB}_{p\times n_2}(\mathrm{r}=2,\mathrm{p}=0.3)$, where p was shared gene dimension, n_1 and n_2 referred to the respective sample size. Then in cohort i, the first 50% samples, $X_i[,1{:}\frac{n_i}{2}]$, were selected as patients while the first 50% genes, $X_i[1{:}\frac{p}{2},]$, were selected as the

diagnosis-related genes. According to the specific proportion $\varphi\epsilon\{20\%, 40\%, 60\%, 80\%\}$ of shared genes over all signature genes, the shared and cohort-specific disease effect was added to the background data of patients. Specifically, the diagnosis-related effect shared by both cohorts was added as $X_i\left[1{:}\frac{p\varphi}{2}, 1{:}\frac{n_i}{2}\right] = X_i\left[1{:}\frac{p\varphi}{2}, 1{:}\frac{n_i}{2}\right] + \mathrm{NB}_{\frac{p\varphi}{2}\times\frac{n_i}{2}}(\mathrm{r}=2, \mathrm{p}=0.002)$ for cohort$i$. The diagnosis-related effect specific to cohort 1 was added as $X_1\left[\frac{2+p\varphi}{2}{:}\frac{p(1+\varphi)}{4}, 1{:}\frac{n_1}{2}\right] = X_1\left[\frac{2+p\varphi}{2}{:}\frac{p(1+\varphi)}{4}, 1{:}\frac{n_1}{2}\right] + \mathrm{NB}_{\frac{p(1-\varphi)}{4}\times\frac{n_1}{2}}(\mathrm{r}=2, \mathrm{p}=0.002)$. The diagnosis-related effect specific to cohort 2 was added as $X_2\left[\frac{4+p(1+\varphi)}{4}{:}\frac{p}{2}, 1{:}\frac{n_2}{2}\right] = X_1\left[\frac{4+p(1+\varphi)}{4}{:}\frac{p}{2}, 1{:}\frac{n_2}{2}\right] + \mathrm{NB}_{\frac{p(1-\varphi)}{4}\times\frac{n_2}{2}}(\mathrm{r}=2, \mathrm{p}=0.002)$. Here the specific effects were not overlapped between cohorts 1 and 2.

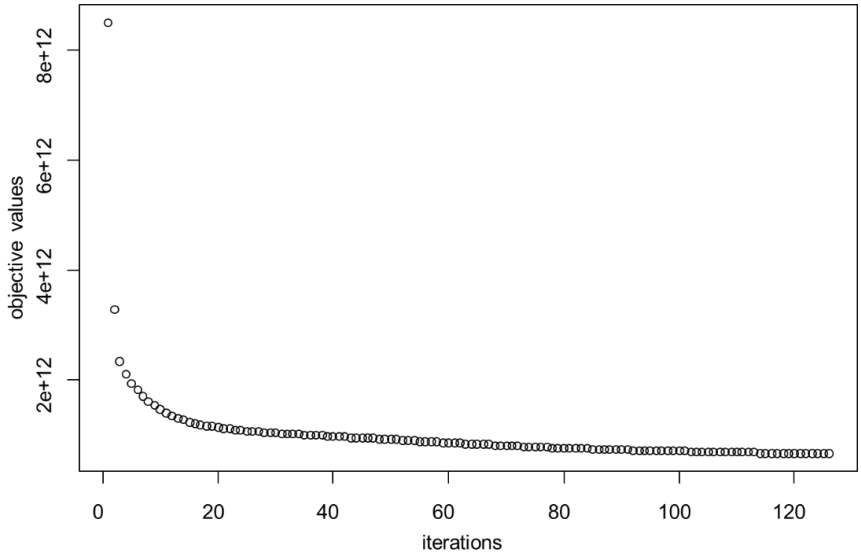
Microarray expression data pre-processing

Four independent cortical microarray gene expression datasets from schizophrenia case-control cohorts were used in this study. Three datasets were downloaded from the GEO repository with id: GSE53987, GSE21138 and GSE35977. A detailed data description can be found on GEO and the respective original studies(Tang, Capitao et al. 2012, Chen, Cheng et al. 2013, Lanz, Reinhart et al. 2019). The fourth dataset was the HBCC microarray dataset (dbGAP ID: phs000979.v3.p2). Th data description and sample acquisition methods can be found on dbGAP and the original publication(Fromer, Roussos et al. 2016).

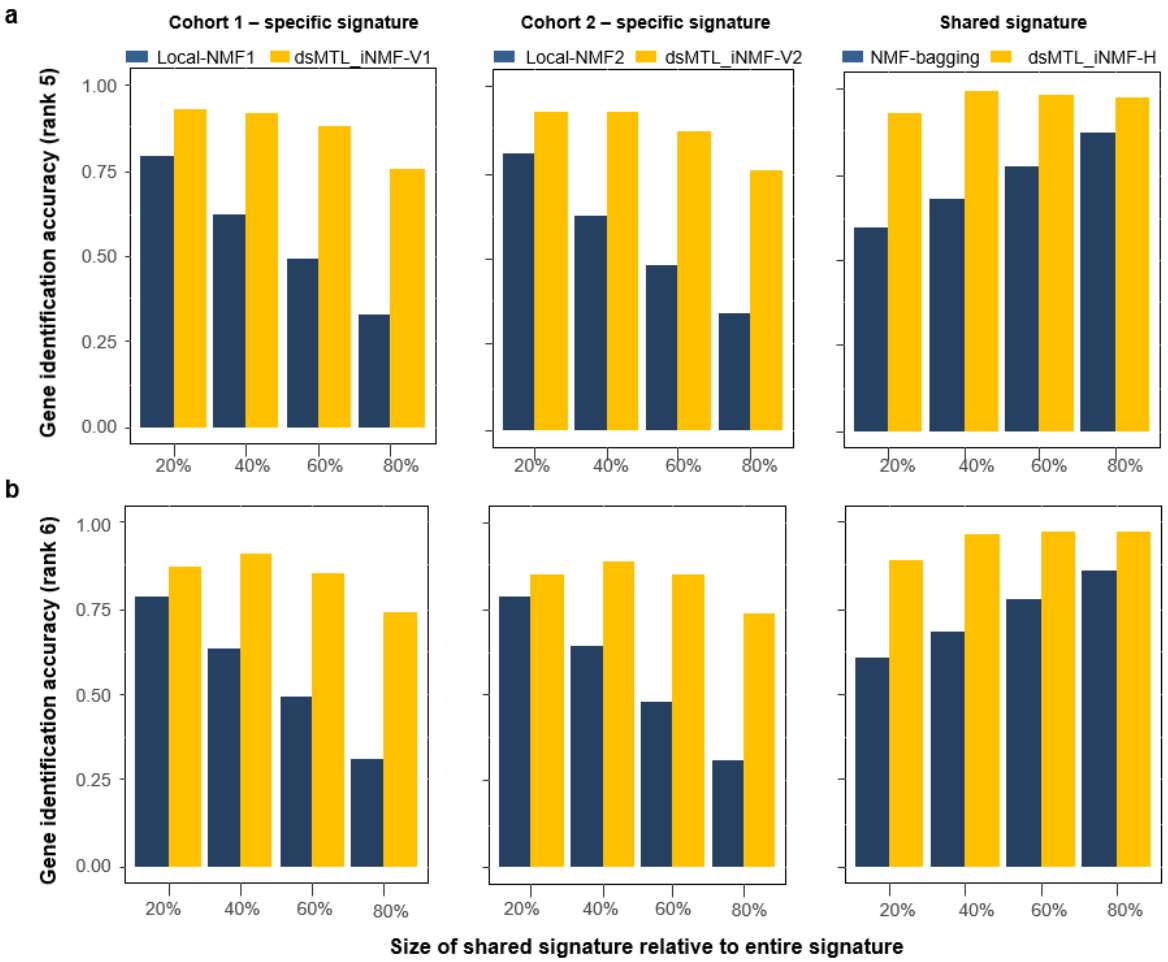
For GSE53987 and GSE21138, the expression levels were measured using the Affymetrix GeneChip Human Genome U133 Plus 2.0 Array, while the data of GSE35977 was measured using Affymetrix Human Gene 1.0 ST Array. A consistent pre-processing procedure was applied to all datasets. First, the raw data was extracted by the function *ReadAffy*() of the R package affy 1.64.0(Gautier, Cope et al. 2004), followed by the rma(Bolstad, Irizarry et al. 2003) (Robust Multi-array Average) procedure for normalization. Values from multiple probes related to the same gene were averaged. Second, subjects with ages < 18 were excluded. Third, outliers were deleted as those outside of four standard deviations from the mean of the first two principal components. Fourth, 10 surrogate variables were determined using SVA(Leek, Johnson et al. 2012) from the R package sva 3.34.0. Fifth, multiple linear regression analysis was used to correct for the effect of potential confounders with the covariates age, age^2, sex, PMI, pH, RIN, batch ID and 10 surrogate variables. Sixth, the resulting expression genes were z-standardized.

HBCC data was normalized and quality controlled as previously described(Fromer, Roussos et al. 2016). First, we extracted the raw dlpfc expression data using the function *read.idat*() from the R package limma 3.42.2 (Ritchie, Phipson et al. 2015). Second, we corrected for background noise using the negative probes followed by quantile normalization and log-transformation. Third, we retained the robustly expressed probes as those with a detection p-value<0.01 in at least half of individuals. Prior to sva analysis, the missing “pH” and “PMI” values were imputed using the average of available data. The covariates contained age, age^2, sex, PMI, pH, RIN, ethnicity and 10 surrogate variables. The cohort contained 321 healthy controls and 191 patients with schizophrenia. All four datasets shared 8013 overlapping genes.

4.6.2 Supplementary Results



Supplementary Figure 1: the curve of objectives training dsMTL_iNMF with an initialization in case study 5. 100 iterations were sufficient to converge to a solution with an acceptable precision.



Supplementary Figure 2. The gene identification accuracy for shared and specific signatures using simulated data using rank=5 (**a**) and rank=6 (**b**) as model parameters.

Index of test	Proportion of homogenous signatures	Number of samples	Number of features	Number of signatures	Proportion of patients
1	20%	200	1000	500	50%
2	40%				
3	60%				
4	80%				

Supplementary Table 1. The simulation data of each server. These parameters were same to each of two servers. The only difference is the set of heterogeneous signatures.

		Server 1	Server 2	Server 3	Client
Type		Training	Training	Training	Testing
Location		Mannheim	Mannheim	Heidelberg	Mannheim
Hardware	CPU	I7-4790 3.6GHz	I7-4790 3.6GHz	Intel Xeon 2.4 GHz	I7-4790 3.6GHz
	Memory	4G	4G	4G	16G
ID		GSE35977	GSE21138	GSE53987	HBCC
Number of Subjects		101	59	34	422
Number of Genes		8013	8013	8013	8013

Supplementary Table 2. Client-server architecture for the real data analysis.

		dsLasso	Mean regularized dsMTL
Misclassification rate		0.34	0.29
Time consumed	5-fold CV	7.5 mins	7.3 mins
	Training	1.7 mins	2.9 mins
Number of network accesses for training		70	137
Non-zero coefficients		38	173

Supplementary Table 3. Performance of dsML/MTL models on real data in real network.

		Server 1	Server 2
Type		Training	Training
Location		Heidelberg	Mannheim
Hardware	CPU	Intel Xeon 2.4 GHz	I7-4790 3.6GHz
	Memory	4G	4G
ID		GSE164376	GSE134497
Number of Subjects		17	16
Number of Genes		15215	15215

Supplementary Table 4. Details of server configurations used for real data analysis.

5 DISCUSSION

The biological understanding of complex disorders such as schizophrenia relies on the large-scale analysis of molecular data. The advances in sequencing technologies reduced the cost of data acquisition, leading to a fast increase in the availability of such data. In addition, the rapid progress in ML technology development has led to widespread availability of sophisticated computational tools to utilize these rich data resources to address complex biological questions. However, this progress has thus far not impacted significantly on the clinical management of severe mental illness, such as schizophrenia. An important factor contributing to this yet lacking clinical translation is our still incomplete mechanistic understanding of illness biology. ML approaches applied in biological psychiatry have thus far focused primarily on the investigation of individual data modalities, which likely captures only part of the complex disease process. The integrative analysis of multiple modalities, however, is a technological challenge, as cross-modality dependencies need to be appropriately captured and potential confounding effects accounted for. The overarching goal of this thesis was thus to develop a ML framework that could capture these effects using cross-task regularization, which led to a spectrum of multi-task learning algorithms that allow a dimensional deconstruction of schizophrenia and that identify an interpretable cross-modality landscape of biological signatures.

In a preliminary work (Cao, Chen et al. 2017), we found a brain expression signature of schizophrenia that could predict a glycemic marker of type 2 diabetes (T2D) in blood. This work followed the conventional “sequential approach” in molecular studies that identifies an illness-related signature in a given data modality and then tests the association of this signature with clinical or biological outcomes in another. This thesis is based on the hypothesis that this sequential approach can be improved, as in high-dimensional data there may be a dimension that is optimally associated with multiple outcomes (e.g. as in the example above with schizophrenia and type 2 diabetes) and biologically interpretable but that cannot be identified from the analysis of an individual data modality. We hypothesize that this methodological gap can be closed via a “simultaneous approach” that jointly analyzed multi-modal data, and identifies a signature predictive of multiple outcomes. In Study 1, this was addressed with an algorithmic framework of MTL and the implementation of the first MTL software for the programming language R, called RMTL. The framework comprises ten well-known MTL algorithms, and simulation analysis supported their improved prediction performance and ability to identify reproducible biological signatures when applied to heterogeneous cohorts. This study further indicated that the prediction models built on multi-modal data could potentially be improved by incorporating an appropriate measure of task-relatedness. Using RMTL, Study 2 investigated the ability to identify gene expression signatures in multi-site gene expression data from patients with schizophrenia and controls, and explored the associated cohort-level heterogeneity. The comparative analysis against several conventional ML algorithms indicated that MTL-derived algorithms showed superior prediction accuracy and signature interpretability. One significant obstacle in applying such algorithms to multimodal data at large scale is the fact that no IT solutions exist where such data can be effectively brought together in a single data warehouse. Frequently, logistic and legal concerns prevent such data integration and have led to a geographically distributed landscape of individually large-scale data resources. To make maximal use of such distributed data, Study 3 developed dsMTL, a computational framework for privacy-preserving, federated multi-task machine learning. dsMTL models demonstrated strong robustness against the institution-level heterogeneity, supporting the suitability of dsMTL for the integrative analysis of heterogeneous, high-dimensional and geographically distributed data cohorts.

5.1 Shared molecular alterations between schizophrenia and T2D

Our preliminary work (Cao, Chen et al. 2017) identified a schizophrenia signature in brain tissue (25% explained variance), that was reproducibly associated (23%~26% explained variance) with a glycemic marker of T2D in blood. The negative association between schizophrenia and HbA_{1c} level indicated that T2D patients had lower predicted schizophrenia scores than healthy controls. This may be due to 1) compensatory mechanism present in the brain or 2) the shared molecular risk causing opposite expression in different tissues. This finding highlighted the molecular complexity of schizophrenia and supported the relevance of antipsychotic treatment for T2D risk.
Four genes in the two pathways "Kidney development" and "respiratory electron transport chain" contributed to the cross-tissue prediction. These findings supported mitochondrial dysfunction and oxidative stress as a unifying theme underlying the comorbidity between schizophrenia and T2D. Notably, the identified signature demonstrated specificity for T2D compared to HIV encephalitis and AD. This is interesting, as the pathology of both AD and HIV is linked to the common pathway "oxidative stress", which may imply that the molecular comorbidity between schizophrenia and T2D is linked to the same pathway (as HIV or AD) but might be regulated differently.

5.2 Standalone and distributed MTL

Study 1 and 3 implemented an algorithmic framework for a standalone (RMTL) and a distributed (dsMTL) MTL package, respectively. Study 1 indicated that MTL was able to capture specific task-relatedness by incorporating cross-task regularization. This implied that MTL could suitably address the heterogeneity across cohorts, e.g., cohort-specific noise, during integrative data analysis. Besides accounting for unwanted heterogeneity, a core objective for the MTL method was to differentiate outcome-specific effects from those shared across outcomes. One important example for this type of investigation is comorbidity analysis, which is aimed at differentiating illness specific biological hallmarks from those indexing the shared biology of the comorbid conditions. For such applications, the simultaneous learning process employed by MTL is a promising technology to accelerate future translational research. In order to make this technology applicable for geographically distributed data, we developed dsMTL as a secure federated MTL system that supports the joint analysis of geographically distributed data warehouses. The analysis of simulated data showed that dsMTL could better differentiate the shared and cohort-specific effects in heterogeneous data cohorts for both supervised and unsupervised analysis compared to the meta-analysis of local ML models. The raw expression (RNA-seq and microarray) data analysis in the actual network illustrated dsMTL was efficient for most gene expression studies (i.e., hundreds of subjects with ~10000 genes getting involved). Most security mechanisms of dsMTL were provided by the DataSHIELD ecosystem, which was not specific to ML applications. An interesting focus of future work is to implement cryptographic algorithms in dsMTL, in order to prevent ML-oriented attacks.

5.3 Heterogeneity in brain expression data of individuals with schizophrenia

Study 2 comprised a comparative analysis of 7 ML/MTL methods applied to five cortex expression datasets. After standard preprocessing, these datasets were considered as "near-homogenous" cohorts and were ready for data concatenation. The comparative analysis indicated the mean regularized MTL outperformed other methods (accuracy: 0.73). Comparing between mean regularized MTL and SVM (best ML model) models trained on varying numbers of training datasets, we found higher consistency and stability of the MTL model. These findings implied that the "near-homogenous"

expression cohorts were still sufficiently heterogeneous for the MTL application to yield a significant benefit in building predictive models through consideration of cohort-level noise. The superiority of the mean regularized MTL in the comparative analysis suggested that the cohort-level noise could be seen as Gaussian noise added to the true underlying signal. Therefore, the network constraint regularization was successful in disentangling the structure of this source of noise. In addition, the superior consistency and stability of MTL suggested that MTL can significantly capture shared effects in high-dimensional, heterogeneous cohorts against the variability of subjects and cohorts, as well as the data scarcity. These advantages are important for the identification of molecular signatures because 1) molecular data is commonly high-dimensional with far fewer observations than features, and 2) the cross-cohort reproducibility is an essential factor for assessing the quality of a given signature.

5.4 Limitations

Most methodologies included in RMTL and dsMTL assume that all tasks share the same task-relatedness. For example, the MTL model with joint feature selection assumes that the identical set of features is relevant for all tasks. However, this is unlikely to be the case in real biological analysis. Due to the frequently occurring and often unmeasurable confounding effects, the signature set of otherwise “homogenous cohorts” could be slightly (or very) different. An extreme case is an outlier task with low data quality that does not share any significant features with other tasks. Including this outlier task in MTL analysis would lead to a twisted model with misleading signatures. A possible countermeasure against this is to wrap the MTL analysis in a cross-validation and statistical testing framework, in order to locate and remove the outlier tasks. However, this would lead to a data split required for the hold-out validation and reduce the statistical power of the MTL analysis. Alternatively, one could solve this issue inside the framework of MTL, with the so-called "dirty statistical models”(Yang and Ravikumar 2013). The approach superposed base models to represent a complex pattern, where one base model is used to capture the shared effect and the other is used to capture the unwanted effect. Based on this approach, one work (Gong, Ye et al. 2012) designed an MTL algorithm to simultaneously detect and remove the outlier tasks from the training procedure. The regularization takes the form ($\Omega(W) = ||P||_{2,1} + ||Q^T||_{2,1}$ S.t. $W = P + Q$). Here, the coefficient matrix is decomposed into a base model P capturing the shared signature over all tasks and a model Q capturing the signature of the outlier tasks. A similar idea was applied in a hybrid model (Jalali, Sanghavi et al. 2010). There, the regularization effect was stratified into a base model P capturing the shared signature ($||P||_{\infty,1}$) across tasks and one Q model capturing a task-specific effect ($||Q||_1$). This algorithm allowed learning any extent of features shared across tasks.

5.5 Outlook

There have been many exciting developments in the MTL community in recent years. Here I enumerate two fields with the highest potential for advancing neuroscience and psychiatry.

High-order MTL. Figure 3 illustrates the evolutional path of machine learning technologies along an increasing availability of different data modalities. Conventional ML can be seen as a “one-to-one” functional mapping for predicting a single outcome on a single modality. Here, the model is a one-dimensional vector. MTL coordinates the joint training procedure on multi-modal data and is still based on a “one-to-one” functional mapping. The MTL model can be represented as a two-dimensional matrix. In the future, as multimodal data becomes increasingly available, MTL would generalize to a “multiple-to-multiple” functional mapping, in order to capture more complex relationships between

modalities and outcomes. For example, multiple data modalities (X1~X3 in Figure 3) and multiple outcomes (Y1~Y3 in Figure 3) could be available, where any data modality can predict any outcome. The model of this "high-order" MTL can be represented as the three-dimensional tensor. This approach has enormous potential in neuroscience and psychiatry due to the representation of rich structures (e.g., biological annotation data) among modalities. For example, one could select genetic, expression and methylation data as X1~X3, where each is used to predict imaging-derived measures of three brain regions (Y1~Y3). Here, the landscape of omics' markers related to the neuroimaging-derived measures can be fully captured by the "high-order" model.

From the experience of conventional MTL, the "high-order" model could overfit without regularization. One study (Romera-Paredes, Aung et al. 2013) built on the concept of traditional low-rank MTL model and proposed a tensor-based trace-norm model ($\Omega(\mathcal{W}) = ||\mathcal{W}||_*, \mathcal{W} = \mathrm{p} \times \mathrm{t}_1 \times \mathrm{t}_2$) to identify the shared low-dimensional space across tasks. Similarly, another study (Lin, Xu et al. 2016) used a similar technique to explore the pairwise feature interactions explicitly in MTL. The functional mapping for the task was $\mathrm{y} = \sum_{\mathrm{i}=1}^{\mathrm{p}} \mathrm{x_i w_i} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{x_i x_j} \mathcal{Q}_{\mathrm{ij}}$, where the coefficients form a $\mathcal{Q} = \mathrm{p} \times \mathrm{p}$ interaction matrix. The cross-task penalty was performed on the stacked Q over tasks. We summarize the formulations in Table 1. In mental illness, a recent work (Brand, Wang et al. 2018) successfully utilized this technique to integrate longitudinal brain scans for modeling the progression of AD.

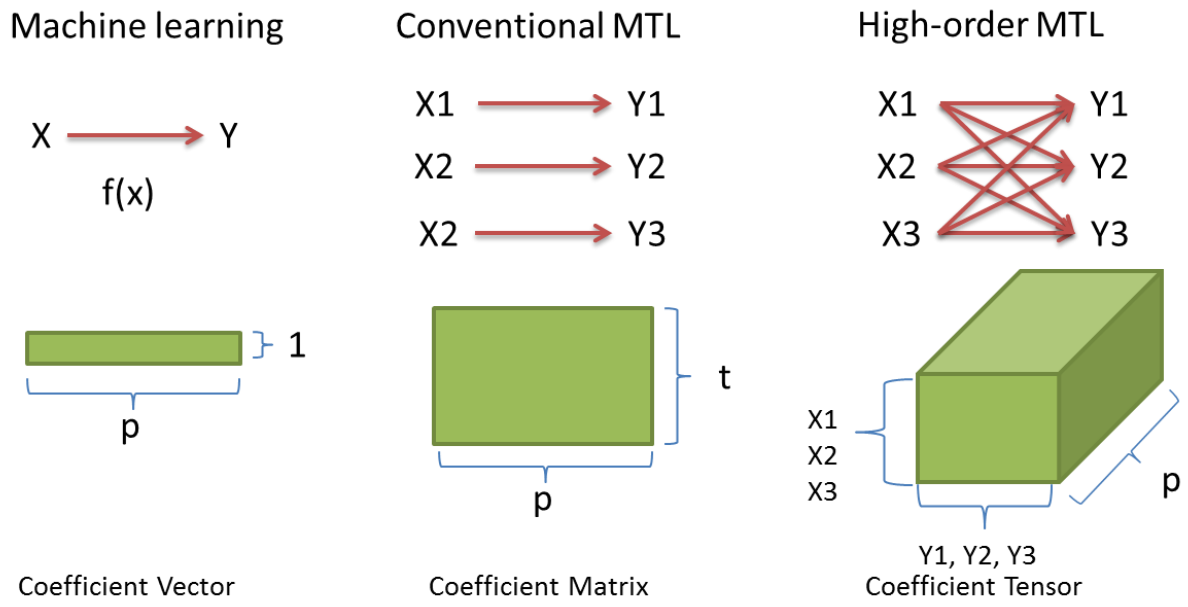


Figure 3. The evolutionary path of ML/MTL. With an increasing availability of data modalities comes an increasing spectrum of interesting functional mappings. In ML, one data modality is mapped to one outcome. Conventional MTL adopts multiple "one-to-one" mappings to link different modalities to outcomes. In high-order MTL, the "multiple-to-multiple" mappings have the potential to connect every modality to every outcome.

Deep MTL. Another fascinating progress regarding MTL development is the so-called "deep MTL", which integrates deep learning techniques into the MTL framework. The core benefit of deep learning is to replace the so-called 'feature engineering' with "representation learning" based on the layered structures of neural networks. Here, the appropriate features can be extracted and learnt from the raw data automatically. This approach has been highly successful for the analysis of several data modalities, including images and text. In psychiatry, using neural imaging data, deep learning achieved

significant success (accuracies > 90%) for predicting the diagnosis of AD (Li, Habes et al. 2017), ADHD (Deshpande, Wang et al. 2015) and schizophrenia (Plis, Hjelm et al. 2014). A detailed review of deep learning in psychiatry can be found elsewhere (Durstewitz, Koppe et al. 2019).

The deep MTL adopts a slightly different strategy for integrating information across tasks compared to regularization-based MTL. Instead of the cross-task regularization, deep MTL commonly shares hidden layers across tasks directly. The underlying principle is the abstraction of natural features into "semantic" features. For example, for imaging-based diagnosis prediction, the low-level layers (closer to the raw data) tend to capture pixel-related patterns, e.g., points or lines. In contrast, the high-level layers (closer to the diagnosis) tend to capture clinical phenotype-related patterns, e.g., behaviors. Such abstraction of natural features is useful for the training in large-scale datasets, in order to avoid overfitting. With the increasing data availability in psychiatry, it is likely that deep MTL will be a highly useful approach to support translational research.

6 SUMMARY

Schizophrenia is a severe and heritable disorder affecting approximately 1% of the population. It has become clear that an improved mechanistic understanding of its underlying biology is a critical factor for improving the clinical management of schizophrenia, for refining the current diagnostic system, and for advancing psychiatry closer to precision medicine. Advanced sequencing technology has led to a fast accumulation of molecular data. Combined with the availability of extensive computing resources and sophisticated ML methods, data science is playing an increasingly important role in schizophrenia research. However, dimensionality of data, as well as the availability of different modalities, may increase faster than the number of individuals for whom such data is available, which may lead to a loss of predictive value and interpretability of algorithms derived from molecular studies (Xu, Xue et al. 2011, Cao, Meyer-Lindenberg et al. 2018). To address this, the present thesis conducted methodological developments in two areas.

First, a significant effort at the algorithmic and computational level has been made to provide the MTL algorithms as a useful tool for both individual researchers and large-scale collaboration projects. RMTL (standalone MTL package) supports a "simultaneous approach" for signature identification in heterogeneous, multi-modal datasets, e.g., for comorbidity analysis and the prediction of multiple clinical outcomes. We showed that such heterogeneity could be captured by cross-task regularization. dsMTL (federated MTL package) supports a secure MTL analysis for geographically distributed datasets. Due to the requirement of privacy protection, the institution-level heterogeneity is challenging to remove when each dataset is analyzed individually. To address this, dsMTL provides a distributed learning system resilient to such heterogeneity. We also showed that dsMTL was computationally efficient for the typical scale of molecular studies.
Second, focusing on gene expression studies of schizophrenia, this thesis explored computational approaches to extract meaningful and biologically reproducible signatures. We found an expression signature associated with schizophrenia as well as T2D, which implied mitochondrial dysfunction and oxidative stress as a unifying theme underlying the comorbidity of these conditions. We also identified a highly accurate, consistent and robust signature in heterogeneous expression cohorts of schizophrenia and controls using MTL.

In summary, the work presented in this thesis introduced the multi-task learning paradigm for modeling the task-level heterogeneity during multi-modal data analysis, and explored its impact on signature identification in standalone and geo-distributed scenarios of molecular studies. This work expands our knowledge of complex learning systems with rich biological structures among data modalities and advances our insight into the molecular biology of schizophrenia. Hopefully, this work will also provide a foundation for the future development of more advanced and efficient MTL methods to further advance translational research in psychiatry.

7 REFERENCES

A;, K. and K. R;. (2021). "GSE164376 dataset." from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164376.
Abi-Dargham, A. and G. Horga (2016). "The search for imaging biomarkers in psychiatric disorders." Nat Med **22**(11): 1248-1255.

Ahmed, A., M. Aly, A. Das, A. J. Smola and T. Anastasakos (2012). "Web-scale multi-task feature selection for behavioral targeting." 1737.
Argyriou, A., T. Evgeniou and M. Pontil (2007). Multi-task feature learning. Advances in neural information processing systems.

Avey, S., S. Mohanty, J. Wilson, H. Zapata, S. R. Joshi, B. Siconolfi, S. Tsang, A. C. Shaw and S. H. Kleinstein (2017). "Multiple network-constrained regressions expand insights into influenza vaccination responses." Bioinformatics **33**(14): i208-i216.

Barnes, M. R., J. Huxley-Jones, P. R. Maycox, M. Lennon, A. Thornber, F. Kelly, S. Bates, A. Taylor, J. Reid, N. Jones, J. Schroeder, C. A. Scorer, C. Davies, J. J. Hagan, J. N. Kew, C. Angelinetta, T. Akbar, S. Hirsch, A. M. Mortimer, T. R. Barnes and J. de Belleroche (2011). "Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia." J Neurosci Res **89**(8): 1218-1227.

Bauerle, A., M. Teufel, V. Musche, B. Weismuller, H. Kohler, M. Hetkamp, N. Dorrie, A. Schweda and E. M. Skoda (2020). "Increased generalized anxiety, depression and distress during the COVID-19 pandemic: a cross-sectional study in Germany." J Public Health (Oxf) **42**(4): 672-678.

Beck, A. and M. Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." SIAM journal on imaging sciences **2**(1): 183-202.

Bhalala, O. G., A. P. Nath, U. K. B. E. Consortium, M. Inouye and C. R. Sibley (2018). "Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue." PLoS Genet **14**(8): e1007607.

Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Bondell, H. D. and B. J. Reich (2008). "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR." Biometrics **64**(1): 115-123.
Brand, L., H. Wang, H. Huang, S. Risacher, A. Saykin, L. Shen and Adni (2018). "Joint High-Order Multi-Task Feature Learning to Predict the Progression of Alzheimer's Disease." Med Image Comput Comput Assist Interv **11070**: 555-562.

Bulik-Sullivan, B., H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P. R. Loh, C. ReproGen, C. Psychiatric Genomics, C. Genetic Consortium for Anorexia Nervosa of the Wellcome Trust

Case Control, L. Duncan, J. R. Perry, N. Patterson, E. B. Robinson, M. J. Daly, A. L. Price and B. M. Neale (2015). "An atlas of genetic correlations across human diseases and traits." Nat Genet **47**(11): 1236-1241.

Bulik-Sullivan, B. K., P. R. Loh, H. K. Finucane, S. Ripke, J. Yang, C. Schizophrenia Working Group of the Psychiatric Genomics, N. Patterson, M. J. Daly, A. L. Price and B. M. Neale (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." Nat Genet **47**(3): 291-295.

Campbell, M., N. Jahanshad, M. Mufford, K. W. Choi, P. Lee, R. Ramesar, J. W. Smoller, P. Thompson, D. J. Stein and S. Dalvie (2021). "Overlap in genetic risk for cross-disorder vulnerability to mental disorders and genetic risk for altered subcortical brain volumes." J Affect Disord **282**: 740-756.

Cao, H., J. Chen, A. Meyer-Lindenberg and E. Schwarz (2017). "A polygenic score for schizophrenia predicts glycemic control." Transl Psychiatry **7**(12): 1295.

Cao, H., A. Meyer-Lindenberg and E. Schwarz (2018). "Comparative Evaluation of Machine Learning Strategies for Analyzing Big Data in Psychiatry." Int J Mol Sci **19**(11).

Cao, H. and E. Schwarz. "An Tutorial for Regularized Multi-task Learning using the package RMTL."

Cao, H. and E. Schwarz (2019). "Computational Approaches for Identification of Pleiotropic Biomarker Profiles in Psychiatry." Reviews on Biomarker Studies of Metabolic and Metabolism-Related Disorders: 111-128.

Cao, H. and E. Schwarz (2020). Opportunities and challenges of machine learning approaches for biomarker signature identification in psychiatry. Personalized Psychiatry, Elsevier**:** 117-126.

Cao, H., J. Zhou and E. Schwarz (2018). "RMTL: An R Library for Multi-Task Learning." Bioinformatics.

Carter, K. W., R. W. Francis, K. Carter, R. Francis, M. Bresnahan, M. Gissler, T. Grønborg, R. Gross, N. Gunnes and G. Hammond (2016). "ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data." International journal of epidemiology **45**(2): 408-416.

Caruana, R. (1998). Multitask Learning, Springer US.

Chapelle, O., P. Shivaswamy, S. Vadrevu, K. Weinberger, Y. Zhang and B. Tseng (2010). "Multi-task learning for boosting with application to web search ranking." 1189.
Chapelle, O., P. Shivaswamy, S. Vadrevu, K. Q. Weinberger, Y. Zhang and B. Tseng (2010). Multi-task learning for boosting with application to web search ranking. . Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10.

Chen, C., L. Cheng, K. Grennan, F. Pibiri, C. Zhang, J. A. Badner, C. Members of the Bipolar Disorder Genome Study, E. S. Gershon and C. Liu (2013). "Two gene co-expression modules differentiate psychotics and controls." Mol Psychiatry **18**(12): 1308-1314.

Chen, J., H. Cao, T. Kaufmann, L. T. Westlye, H. Tost, A. Meyer-Lindenberg and E. Schwarz (2020). "Identification of reproducible BCL11A alterations in schizophrenia through individual-level prediction of coexpression." Schizophrenia bulletin **46**(5): 1165-1171.

Chen, J., Z. Zang, U. Braun, K. Schwarz, A. Harneit, T. Kremer, R. Ma, J. Schweiger, C. Moessnang, L. Geiger, H. Cao, F. Degenhardt, M. M. Nothen, H. Tost, A. Meyer-Lindenberg and E. Schwarz (2020). "Association of a Reproducible Epigenetic Risk Profile for Schizophrenia With Brain Methylation and Function." JAMA Psychiatry **77**(6): 628-636.

Choi, S. W., T. S. Mak and P. F. O'Reilly (2020). "Tutorial: a guide to performing polygenic risk score analyses." Nat Protoc **15**(9): 2759-2772.

Collobert, R. and J. Weston (2008). "A unified architecture for natural language processing: deep neural networks with multitask learning." 160-167.

Consortium, G. T. (2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." Science **348**(6235): 648-660.

Cross-Disorder Group of the Psychiatric Genomics, C., S. H. Lee, S. Ripke, B. M. Neale, S. V. Faraone, S. M. Purcell, R. H. Perlis, B. J. Mowry, A. Thapar, M. E. Goddard, J. S. Witte, D. Absher, I. Agartz, H. Akil, F. Amin, O. A. Andreassen, A. Anjorin, R. Anney, V. Anttila, D. E. Arking, P. Asherson, M. H. Azevedo, L. Backlund, J. A. Badner, A. J. Bailey, T. Banaschewski, J. D. Barchas, M. R. Barnes, T. B. Barrett, N. Bass, A. Battaglia, M. Bauer, M. Bayes, F. Bellivier, S. E. Bergen, W. Berrettini, C. Betancur, T. Bettecken, J. Biederman, E. B. Binder, D. W. Black, D. H. Blackwood, C. S. Bloss, M. Boehnke, D. I. Boomsma, G. Breen, R. Breuer, R. Bruggeman, P. Cormican, N. G. Buccola, J. K. Buitelaar, W. E. Bunney, J. D. Buxbaum, W. F. Byerley, E. M. Byrne, S. Caesar, W. Cahn, R. M. Cantor, M. Casas, A. Chakravarti, K. Chambert, K. Choudhury, S. Cichon, C. R. Cloninger, D. A. Collier, E. H. Cook, H. Coon, B. Cormand, A. Corvin, W. H. Coryell, D. W. Craig, I. W. Craig, J. Crosbie, M. L. Cuccaro, D. Curtis, D. Czamara, S. Datta, G. Dawson, R. Day, E. J. De Geus, F. Degenhardt, S. Djurovic, G. J. Donohoe, A. E. Doyle, J. Duan, F. Dudbridge, E. Duketis, R. P. Ebstein, H. J. Edenberg, J. Elia, S. Ennis, B. Etain, A. Fanous, A. E. Farmer, I. N. Ferrier, M. Flickinger, E. Fombonne, T. Foroud, J. Frank, B. Franke, C. Fraser, R. Freedman, N. B. Freimer, C. M. Freitag, M. Friedl, L. Frisen, L. Gallagher, P. V. Gejman, L. Georgieva, E. S. Gershon, D. H. Geschwind, I. Giegling, M. Gill, S. D. Gordon, K. Gordon-Smith, E. K. Green, T. A. Greenwood, D. E. Grice, M. Gross, D. Grozeva, W. Guan, H. Gurling, L. De Haan, J. L. Haines, H. Hakonarson, J. Hallmayer, S. P. Hamilton, M. L. Hamshere, T. F. Hansen, A. M. Hartmann, M. Hautzinger, A. C. Heath, A. K. Henders, S. Herms, I. B. Hickie, M. Hipolito, S. Hoefels, P. A. Holmans, F. Holsboer, W. J. Hoogendijk, J. J. Hottenga, C. M. Hultman, V. Hus, A. Ingason, M. Ising, S. Jamain, E. G. Jones, I. Jones, L. Jones, J. Y. Tzeng, A. K. Kahler, R. S. Kahn, R. Kandaswamy, M. C. Keller, J. L. Kennedy, E. Kenny, L. Kent, Y. Kim, G. K. Kirov, S. M. Klauck, L. Klei, J. A. Knowles, M. A. Kohli, D. L. Koller, B. Konte, A. Korszun, L. Krabbendam, R. Krasucki, J. Kuntsi, P. Kwan, M. Landen, N. Langstrom, M. Lathrop, J. Lawrence, W. B. Lawson, M. Leboyer, D. H. Ledbetter, P. H. Lee, T. Lencz, K. P. Lesch, D. F. Levinson, C. M. Lewis, J. Li, P. Lichtenstein, J. A. Lieberman, D. Y. Lin,

D. H. Linszen, C. Liu, F. W. Lohoff, S. K. Loo, C. Lord, J. K. Lowe, S. Lucae, D. J. MacIntyre, P. A. Madden, E. Maestrini, P. K. Magnusson, P. B. Mahon, W. Maier, A. K. Malhotra, S. M. Mane, C. L. Martin, N. G. Martin, M. Mattheisen, K. Matthews, M. Mattingsdal, S. A. McCarroll, K. A. McGhee, J. J. McGough, P. J. McGrath, P. McGuffin, M. G. McInnis, A. McIntosh, R. McKinney, A. W. McLean, F. J. McMahon, W. M. McMahon, A. McQuillin, H. Medeiros, S. E. Medland, S. Meier, I. Melle, F. Meng, J. Meyer, C. M. Middeldorp, L. Middleton, V. Milanova, A. Miranda, A. P. Monaco, G. W. Montgomery, J. L. Moran, D. Moreno-De-Luca, G. Morken, D. W. Morris, E. M. Morrow, V. Moskvina, P. Muglia, T. W. Muhleisen, W. J. Muir, B. Muller-Myhsok, M. Murtha, R. M. Myers, I. Myin-Germeys, M. C. Neale, S. F. Nelson, C. M. Nievergelt, I. Nikolov, V. Nimgaonkar, W. A. Nolen, M. M. Nothen, J. I. Nurnberger, E. A. Nwulia, D. R. Nyholt, C. O'Dushlaine, R. D. Oades, A. Olincy, G. Oliveira, L. Olsen, R. A. Ophoff, U. Osby, M. J. Owen, A. Palotie, J. R. Parr, A. D. Paterson, C. N. Pato, M. T. Pato, B. W. Penninx, M. L. Pergadia, M. A. Pericak-Vance, B. S. Pickard, J. Pimm, J. Piven, D. Posthuma, J. B. Potash, F. Poustka, P. Propping, V. Puri, D. J. Quested, E. M. Quinn, J. A. Ramos-Quiroga, H. B. Rasmussen, S. Raychaudhuri, K. Rehnstrom, A. Reif, M. Ribases, J. P. Rice, M. Rietschel, K. Roeder, H. Roeyers, L. Rossin, A. Rothenberger, G. Rouleau, D. Ruderfer, D. Rujescu, A. R. Sanders, S. J. Sanders, S. L. Santangelo, J. A. Sergeant, R. Schachar, M. Schalling, A. F. Schatzberg, W. A. Scheftner, G. D. Schellenberg, S. W. Scherer, N. J. Schork, T. G. Schulze, J. Schumacher, M. Schwarz, E. Scolnick, L. J. Scott, J. Shi, P. D. Shilling, S. I. Shyn, J. M. Silverman, S. L. Slager, S. L. Smalley, J. H. Smit, E. N. Smith, E. J. Sonuga-Barke, D. St Clair, M. State, M. Steffens, H. C. Steinhausen, J. S. Strauss, J. Strohmaier, T. S. Stroup, J. S. Sutcliffe, P. Szatmari, S. Szelinger, S. Thirumalai, R. C. Thompson, A. A. Todorov, F. Tozzi, J. Treutlein, M. Uhr, E. J. van den Oord, G. Van Grootheest, J. Van Os, A. M. Vicente, V. J. Vieland, J. B. Vincent, P. M. Visscher, C. A. Walsh, T. H. Wassink, S. J. Watson, M. M. Weissman, T. Werge, T. F. Wienker, E. M. Wijsman, G. Willemsen, N. Williams, A. J. Willsey, S. H. Witt, W. Xu, A. H. Young, T. W. Yu, S. Zammit, P. P. Zandi, P. Zhang, F. G. Zitman, S. Zollner, B. Devlin, J. R. Kelsoe, P. Sklar, M. J. Daly, M. C. O'Donovan, N. Craddock, P. F. Sullivan, J. W. Smoller, K. S. Kendler, N. R. Wray and C. International Inflammatory Bowel Disease Genetics (2013). "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs." Nat Genet **45**(9): 984-994.

Dahl, M., J. Mancuso, Y. Dupis, B. Decoste, M. Giraud, I. Livingstone, J. Patriquin and G. Uhma (2018). "Private machine learning in tensorflow using secure computation." arXiv preprint arXiv:1810.08130.

Dalmau, J., A. J. Gleichman, E. G. Hughes, J. E. Rossi, X. Peng, M. Lai, S. K. Dessain, M. R. Rosenfeld, R. Balice-Gordon and D. R. Lynch (2008). "Anti-NMDA-receptor encephalitis: case series and analysis of the effects of antibodies." The Lancet Neurology **7**(12): 1091-1098.
De Hert, M., C. U. Correll, J. Bobes, M. Cetkovich-Bakmas, D. Cohen, I. Asai, J. Detraux, S. Gautam, H.-J. Möller and D. M. Ndetei (2011). "Physical illness in patients with severe mental disorders. I. Prevalence, impact of medications and disparities in health care." World psychiatry **10**(1): 52.

de Wit, S. J., P. Alonso, L. Schweren, D. Mataix-Cols, C. Lochner, J. M. Menchon, D. J. Stein, J. P. Fouche, C. Soriano-Mas, J. R. Sato, M. Q. Hoexter, D. Denys, T. Nakamae, S. Nishida, J. S. Kwon, J. H. Jang, G. F. Busatto, N. Cardoner, D. C. Cath, K. Fukui, W. H. Jung, S. N. Kim, E. C. Miguel, J. Narumoto, M. L. Phillips, J. Pujol, P. L. Remijnse, Y. Sakai, N. Y. Shin, K. Yamada, D. J. Veltman and O. A. van den Heuvel (2014). "Multicenter voxel-based morphometry mega-

analysis of structural brain scans in obsessive-compulsive disorder." Am J Psychiatry **171**(3): 340-349.

DeLisi, L. E. (1992). "The significance of age of onset for schizophrenia." Schizophrenia bulletin **18**(2): 209-215.

Deshpande, G., P. Wang, D. Rangaprakash and B. Wilamowski (2015). "Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data." IEEE Trans Cybern **45**(12): 2668-2679.

Durstewitz, D., G. Koppe and A. Meyer-Lindenberg (2019). "Deep neural networks in psychiatry." Mol Psychiatry.

Dwork, C. (2006). Differential privacy. International Colloquium on Automata, Languages, and Programming, Springer.

Egerton, A., O. D. Howes, S. Houle, K. McKenzie, L. R. Valmaggia, M. R. Bagby, H.-H. Tseng, M. A. Bloomfield, M. Kenk and S. Bhattacharyya (2017). "Elevated striatal dopamine function in immigrants and their children: a risk mechanism for psychosis." Schizophrenia bulletin **43**(2): 293-301.

Evgeniou, T. and M. Pontil (2004). Regularized multi-task learning. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Feriante, J. (2015). Massively Multitask Deep Learning for Drug Discovery, University of Wisconsin-Madison.

Franke, B., J. L. Stein, S. Ripke, V. Anttila, D. P. Hibar, K. J. van Hulzen, A. Arias-Vasquez, J. W. Smoller, T. E. Nichols, M. C. Neale, A. M. McIntosh, P. Lee, F. J. McMahon, A. Meyer-Lindenberg, M. Mattheisen, O. A. Andreassen, O. Gruber, P. S. Sachdev, R. Roiz-Santianez, A. J. Saykin, S. Ehrlich, K. A. Mather, J. A. Turner, E. Schwarz, A. Thalamuthu, Y. Yao, Y. Y. Ho, N. G. Martin, M. J. Wright, C. Schizophrenia Working Group of the Psychiatric Genomics, C. Psychosis Endophenotypes International, C. Wellcome Trust Case Control, C. Enigma, M. C. O'Donovan, P. M. Thompson, B. M. Neale, S. E. Medland and P. F. Sullivan (2016). "Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept." Nat Neurosci **19**(3): 420-431.

Fredrikson, M., E. Lantz, S. Jha, S. Lin, D. Page and T. Ristenpart (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. 23rd {USENIX} Security Symposium ({USENIX} Security 14).

Friedman, J., T. Hastie and R. Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software **33**(1).

Fromer, M., P. Roussos, S. K. Sieberts, J. S. Johnson, D. H. Kavanagh, T. M. Perumal, D. M. Ruderfer, E. C. Oh, A. Topol, H. R. Shah, L. L. Klei, R. Kramer, D. Pinto, Z. H. Gumus, A. E. Cicek, K. K. Dang, A. Browne, C. Lu, L. Xie, B. Readhead, E. A. Stahl, J. Xiao, M. Parvizi, T.

Hamamsy, J. F. Fullard, Y.-C. Wang, M. C. Mahajan, J. M. J. Derry, J. T. Dudley, S. E. Hemby, B. A. Logsdon, K. Talbot, T. Raj, D. A. Bennett, P. L. De Jager, J. Zhu, B. Zhang, P. F. Sullivan, A. Chess, S. M. Purcell, L. A. Shinobu, L. M. Mangravite, H. Toyoshiba, R. E. Gur, C.-G. Hahn, D. A. Lewis, V. Haroutunian, M. A. Peters, B. K. Lipska, J. D. Buxbaum, E. E. Schadt, K. Hirai, K. Roeder, K. J. Brennand, N. Katsanis, E. Domenici, B. Devlin and P. Sklar (2016). "Gene expression elucidates functional impact of polygenic risk for schizophrenia." Nat Neurosci **19**(11): 1442-1453.

Fujita, N., S. Mizuarai, K. Murakami and K. Nakai (2018). "Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses." Sci Rep **8**(1): 9743.
Gaebel, W. and J. Zielasek (2015). "Schizophrenia in 2020: Trends in diagnosis and therapy." Psychiatry Clin Neurosci **69**(11): 661-673.

Gandal, M. J., J. R. Haney, N. N. Parikshak, V. Leppa, G. Ramaswami, C. Hartl, A. J. Schork, V. Appadurai, A. Buil, T. M. Werge, C. Liu, K. P. White, C. CommonMind, E. C. Psych, P.-B. W. G. i, S. Horvath and D. H. Geschwind (2018). "Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap." Science **359**(6376): 693-697.

Gautier, L., L. Cope, B. M. Bolstad and R. A. Irizarry (2004). "affy--analysis of Affymetrix GeneChip data at the probe level." Bioinformatics **20**(3): 307-315.

Gaye, A., Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, E. M. Jones, J. Minion, A. W. Boyd, C. J. Newby and M.-L. Nuotio (2014). "DataSHIELD: taking the analysis to the data, not the data to the analysis." International journal of epidemiology **43**(6): 1929-1944.

Gifford, G., R. McCutcheon and P. McGuire (2020). Neuroimaging studies in people at clinical high risk for psychosis. Risk Factors for Psychosis, Elsevier**:** 167-182.

Gong, P., J. Ye and C. Zhang (2012). Robust multi-task feature learning. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
Greenlaw, K., E. Szefer, J. Graham, M. Lesperance, F. S. Nathoo and I. Alzheimer's Disease Neuroimaging (2017). "A Bayesian group sparse multi-task regression model for imaging genetics." Bioinformatics **33**(16): 2513-2522.

Han Cao and E. Schwarz "dsMTL - a computational framework for privacy-preserving, distributed multi-task machine learning (in preparation)."

Harris, L. W., M. Wayland, M. Lan, M. Ryan, T. Giger, H. Lockstone, I. Wuethrich, M. Mimmack, L. Wang, M. Kotter, R. Craddock and S. Bahn (2008). "The cerebral microvasculature in schizophrenia: a laser capture microdissection study." PLoS One **3**(12): e3964.

Hastie, T., R. Tibshirani and J. Friedman (2009). "The Elements of Statistical Learning."
Hietala, J. and E. Syvälahti (1996). "Dopamine in schizophrenia." Annals of medicine **28**(6): 557-561.

Hoerl, A. E. and R. W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics **12**(1): 55.

Hood, L. E., M. W. Hunkapiller and L. M. Smith (1987). "Automated DNA sequencing and analysis of the human genome." Genomics **1**(3): 201-212.

Hu, H., Z. Salcic, G. Dobbie and X. Zhang (2021). "Membership Inference Attacks on Machine Learning: A Survey." arXiv preprint arXiv:2103.07853.

Huo, Y., L. Xin, C. Kang, M. Wang, Q. Ma and B. Yu (2020). "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso." J Theor Biol **486**: 110098.

Iniesta, R., D. Stahl and P. McGuffin (2016). "Machine learning, statistical learning and the future of biological research in psychiatry." Psychol Med **46**(12): 2455-2465.

International Schizophrenia, C., S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan and P. Sklar (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." Nature **460**(7256): 748-752.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Jablensky, A. (2010). "The diagnostic concept of schizophrenia: its history, evolution, and future prospects." Dialogues in clinical neuroscience **12**(3): 271.

Jacob, L., J.-p. Vert and F. R. Bach (2008). Clustered Multi-Task Learning: A Convex Formulation. Advances in Neural Information Processing Systems 21 (NIPS 2008).

Jahanshad, N., P. V. Kochunov, E. Sprooten, R. C. Mandl, T. E. Nichols, L. Almasy, J. Blangero, R. M. Brouwer, J. E. Curran and G. I. de Zubicaray (2013). "Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group." Neuroimage **81**: 455-469.

Jalali, A., S. Sanghavi, C. Ruan and P. K. Ravikumar (2010). A dirty model for multi-task learning. Advances in neural information processing systems.

James, S. L., D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela and A. Abdelalim (2018). "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017." The Lancet **392**(10159): 1789-1858.

Ji, S. and J. Ye (2009). An accelerated gradient method for trace norm minimization. Proceedings of the 26th annual international conference on machine learning, ACM.
Jing, W., Z. Zhilin, Y. Jingwen, L. Taiyong, B. D. Rao, F. Shiaofen, K. Sungeun, S. L. Risacher, A. J. Saykin and S. Li (2012). "Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease." 940-947.

Johnstone, E., D. Owens, G. Bydder, N. Colter, T. Crow and C. Frith (1989). "The spectrum of structural brain changes in schizophrenia: age of onset as a predictor of cognitive and clinical impairments and their cerebral correlates." Psychological medicine **19**(1): 91-103.

Jones, E. M., N. A. Sheehan, N. Masca, S. E. Wallace, M. J. Murtagh and P. R. Burton (2012). "DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective." Norsk Epidemiologi **21**(2).

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects." Science **349**(6245): 255-260.

Kathuria, A., K. Lopez-Lengowski, M. Vater, D. McPhie, B. M. Cohen and R. Karmacharya (2020). "Transcriptome analysis and functional characterization of cerebral organoids in bipolar disorder." Genome Med **12**(1): 34.

Kochunov, P., N. Jahanshad, E. Sprooten, T. E. Nichols, R. C. Mandl, L. Almasy, T. Booth, R. M. Brouwer, J. E. Curran and G. I. de Zubicaray (2014). "Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling." Neuroimage **95**: 136-150.

Kohannim, O., D. P. Hibar, J. L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, A. W. Toga, C. R. Jack, Jr., M. W. Weiner, G. I. de Zubicaray, K. L. McMahon, N. K. Hansell, N. G. Martin, M. J. Wright, P. M. Thompson and I. Alzheimer's Disease Neuroimaging (2012). "Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression." Front Neurosci **6**: 115.

Konečný, J., B. McMahan and D. Ramage (2015). "Federated optimization: Distributed optimization beyond the datacenter." arXiv preprint arXiv:1511.03575.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer,

G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki and C. International Human Genome Sequencing (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Lanz, T. A., V. Reinhart, M. J. Sheehan, S. J. S. Rizzo, S. E. Bove, L. C. James, D. Volfson, D. A. Lewis and R. J. Kleiman (2019). "Postmortem transcriptional profiling reveals widespread increase in inflammation in schizophrenia: a comparison of prefrontal cortex, striatum, and hippocampus among matched tetrads of controls with subjects diagnosed with schizophrenia, bipolar or major depressive disorder." Transl Psychiatry **9**(1): 151.

Laursen, T. M., M. Nordentoft and P. B. Mortensen (2014). "Excess early mortality in schizophrenia." Annual review of clinical psychology **10**: 425-448.

Lee, S. H., S. Ripke, B. M. Neale, S. V. Faraone, S. M. Purcell, R. H. Perlis, B. J. Mowry, A. Thapar, M. E. Goddard and J. S. Witte (2013). "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs." Nature genetics **45**(9): 984-995.

Lee, S. H., J. Yang, M. E. Goddard, P. M. Visscher and N. R. Wray (2012). "Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood." Bioinformatics **28**(19): 2540-2542.

Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe and J. D. Storey (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." Bioinformatics **28**(6): 882-883.

Lehmann, H. E. and T. A. Ban (1997). "The history of the psychopharmacology of schizophrenia." The Canadian Journal of Psychiatry **42**(2): 152-162.

Lenz, S., M. Hess and H. Binder (2021). "Deep generative models in DataSHIELD." BMC Med Res Methodol **21**(1): 64.

Li, C. and H. Li (2008). "Network-constrained regularization and variable selection for analysis of genomic data." Bioinformatics **24**(9): 1175-1182.

Li, H., M. Habes and Y. Fan (2017). "Deep ordinal ranking for multi-category diagnosis of alzheimer's disease using hippocampal MRI data." arXiv preprint arXiv:1709.01599.

Li, J., X. Liu, W. Yin, M. Yang, L. Ma and Y. Jin (2020). "Empirical evaluation of multi-task learning in deep neural networks for natural language processing." Neural Computing and Applications **33**(9): 4417-4428.

Li, Y., J. Wang, J. Ye and C. K. Reddy (2016). A Multi-Task Learning Formulation for Survival Analysis. Proceedings of the 22nd ACM SIGKDD International Conference on Knowedge Discovery and Data Mining (KDD'16).

Li, Y., J. Wang, J. Ye and C. K. Reddy (2016). "A Multi-Task Learning Formulation for Survival Analysis." 1715-1724.

Li, Y., C. Willer, S. Sanna and G. Abecasis (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.

Lichtenstein, P., B. H. Yip, C. Björk, Y. Pawitan, T. D. Cannon, P. F. Sullivan and C. M. Hultman (2009). "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study." The Lancet **373**(9659): 234-239.

Lieberman, J., R. Girgis, G. Brucato, H. Moore, F. Provenzano, L. Kegeles, D. Javitt, J. Kantrowitz, M. Wall and C. Corcoran (2018). "Hippocampal dysfunction in the pathophysiology of schizophrenia: a selective review and hypothesis for early detection and intervention." Molecular psychiatry **23**(8): 1764-1772.

Lin, D., J. Zhang, J. Li, V. D. Calhoun, H. W. Deng and Y. P. Wang (2013). "Group sparse canonical correlation analysis for genomic data integration." BMC Bioinformatics **14**: 245.
Lin, D., J. Zhang, J. Li, H. He, H. W. Deng and Y. P. Wang (2014). "Integrative analysis of multiple diverse omics datasets by sparse group multitask regression." Front Cell Dev Biol **2**: 62.

Lin, K., J. Xu, I. M. Baytas, S. Ji and J. Zhou (2016). "Multi-Task Feature Interaction Learning." 1735-1744.

Liu, J., S. Ji and J. Ye (2009). Multi-task feature learning via efficient l2, 1-norm minimization. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence.
Liu, J. and Y. Jieping "Efficient L1/Lq Norm Regularization."

Liu, J., K. Wang, S. Ma and J. Huang (2013). "Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method." Stat Interface **6**(1): 99-115.

Liu, J. and J. Ye (2009). "Efficient l_1/lq-norm regularization." Technical report, Arizona State University.

Liu, S., S. J. Pan and Q. Ho (2017). "Distributed Multi-Task Relationship Learning." 937-946.

Major Depressive Disorder Working Group of the Psychiatric, G. C., S. Ripke, N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman, G. Breen, E. M. Byrne, D. H. Blackwood, D. I. Boomsma, S. Cichon, A. C. Heath, F. Holsboer, S. Lucae, P. A. Madden, N. G. Martin, P. McGuffin, P. Muglia, M. M. Noethen, B. P. Penninx, M. L. Pergadia, J. B. Potash, M. Rietschel, D. Lin, B. Muller-Myhsok, J. Shi, S. Steinberg, H. J. Grabe, P. Lichtenstein, P. Magnusson, R. H. Perlis, M. Preisig, J. W. Smoller, K. Stefansson, R. Uher, Z. Kutalik, K. E. Tansey, A. Teumer, A. Viktorin, M. R. Barnes, T. Bettecken, E. B. Binder, R. Breuer, V. M. Castro, S. E. Churchill, W. H. Coryell, N. Craddock, I. W. Craig, D. Czamara, E. J. De Geus, F. Degenhardt, A. E. Farmer, M. Fava, J. Frank, V. S. Gainer, P. J. Gallagher, S. D. Gordon, S. Goryachev, M. Gross, M. Guipponi, A. K. Henders, S. Herms, I. B. Hickie, S. Hoefels, W. Hoogendijk, J. J. Hottenga, D. V. Iosifescu, M. Ising, I. Jones, L. Jones, T. Jung-Ying, J. A. Knowles, I. S. Kohane, M. A. Kohli, A. Korszun, M. Landen, W. B. Lawson, G. Lewis, D. Macintyre, W. Maier, M. Mattheisen, P. J. McGrath, A. McIntosh, A. McLean, C. M. Middeldorp, L. Middleton, G. M. Montgomery, S. N. Murphy, M. Nauck, W. A. Nolen, D. R. Nyholt, M. O'Donovan, H. Oskarsson, N. Pedersen, W. A. Scheftner, A. Schulz, T. G. Schulze, S. I. Shyn, E. Sigurdsson, S. L. Slager, J. H. Smit, H. Stefansson, M. Steffens, T. Thorgeirsson, F. Tozzi, J. Treutlein, M. Uhr, E. J. van den Oord, G. Van Grootheest, H. Volzke, J. B. Weilburg, G. Willemsen, F. G. Zitman, B. Neale, M. Daly, D. F. Levinson and P. F. Sullivan (2013). "A mega-analysis of genome-wide association studies for major depressive disorder." Mol Psychiatry **18**(4): 497-511.

Marcon, Y., T. Bishop, D. Avraam, X. Escriba-Montagut, P. Ryser-Welch, S. Wheater, P. Burton and J. R. Gonzalez (2021). "Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD." PLoS Comput Biol **17**(3): e1008880.

Marquand, A. F., M. Brammer, S. C. Williams and O. M. Doyle (2014). "Bayesian multi-task learning for decoding multi-subject neuroimaging data." Neuroimage **92**: 298-311.

Matschinske, J., J. Späth, R. Nasirigerdeh, R. Torkzadehmahani, A. Hartebrodt, B. Orbán, S. Fejér, O. Zolotareva, M. Bakhtiari and B. Bihari (2021). "The FeatureCloud AI Store for Federated Learning in Biomedicine and Beyond." arXiv preprint arXiv:2105.05734.

Maycox, P. R., F. Kelly, A. Taylor, S. Bates, J. Reid, R. Logendra, M. R. Barnes, C. Larminie, N. Jones, M. Lennon, C. Davies, J. J. Hagan, C. A. Scorer, C. Angelinetta, M. T. Akbar, S. Hirsch, A. M. Mortimer, T. R. Barnes and J. de Belleroche (2009). "Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function." Mol Psychiatry **14**(12): 1083-1094.

McLaughlin, R. L., D. Schijven, W. van Rheenen, K. R. van Eijk, M. O'Brien, R. S. Kahn, R. A. Ophoff, A. Goris, D. G. Bradley, A. Al-Chalabi, L. H. van den Berg, J. J. Luykx, O. Hardiman, J. H. Veldink, E. G. C. Project Min and C. Schizophrenia Working Group of the Psychiatric Genomics (2017). "Genetic correlation between amyotrophic lateral sclerosis and schizophrenia." Nat Commun **8**: 14774.

Medland, S. E., K. L. Grasby, N. Jahanshad, J. N. Painter, L. Colodro-Conde, J. Bralten, D. P. Hibar, P. A. Lind, F. Pizzagalli, S. I. Thomopoulos, J. L. Stein, B. Franke, N. G. Martin, P. M. Thompson and E. G. W. Group (2020). "Ten years of enhancing neuro-imaging genetics through meta-analysis: An overview from the ENIGMA Genetics Working Group." Hum Brain Mapp.

Meier, L., S. Van De Geer and P. Bühlmann (2008). "The group lasso for logistic regression." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(1): 53-71.

Meyer-Lindenberg, A. and D. R. Weinberger (2006). "Intermediate phenotypes and genetic mechanisms of psychiatric disorders." Nat Rev Neurosci **7**(10): 818-827.

Narayan, S., B. Tang, S. R. Head, T. J. Gilmartin, J. G. Sutcliffe, B. Dean and E. A. Thomas (2008). "Molecular profiles of schizophrenia in the CNS at different stages of illness." Brain Res **1239**: 235-248.

Nesterov, Y. (2012). "Gradient methods for minimizing composite functions." Mathematical Programming **140**(1): 125-161.

O'Brien, C. M. (2016). "Statistical Learning with Sparsity: The Lasso and Generalizations." International Statistical Review **84**(1): 156-157.

O'Dushlaine, C., L. Rossin, H. P. Lee, L. Duncan, N. N. Parikshak, S. Newhouse and S. Ripke (2015). "Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways." Nat Neurosci **18**(2): 199-209.

Olney, J. W. and N. B. Farber (1995). "Glutamate receptor dysfunction and schizophrenia." Archives of general psychiatry **52**(12): 998-1007.

Oswald, L. M., G. S. Wand, H. Kuwabara, D. F. Wong, S. Zhu and J. R. Brasic (2014). "History of childhood adversity is positively associated with ventral striatal dopamine responses to amphetamine." Psychopharmacology **231**(12): 2417-2433.

Parikh, N. and S. Boyd (2014). "Proximal algorithms." Foundations and Trends® in Optimization **1**(3): 127-239.

Passos, I. C., B. Mwangi and F. Kapczinski (2016). "Big data analytics and machine learning: 2015 and beyond." The Lancet Psychiatry **3**(1): 13-15.

Plis, S. M., D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner and V. D. Calhoun (2014). "Deep learning for neuroimaging: a validation study." Front Neurosci **8**: 229.

Pong, T. K., P. Tseng, S. Ji and J. Ye (2010). "Trace Norm Regularization: Reformulations, Algorithms, and Multi-Task Learning." SIAM Journal on Optimization **20**(6): 3465-3489.

Prive, F., B. J. Vilhjalmsson, H. Aschard and M. G. B. Blum (2019). "Making the Most of Clumping and Thresholding for Polygenic Scores." Am J Hum Genet **105**(6): 1213-1221.

Pruessner, J. C., F. Champagne, M. J. Meaney and A. Dagher (2004). "Dopamine release in response to a psychological stress in humans and its relationship to early life maternal care: a positron emission tomography study using [11C] raclopride." Journal of Neuroscience **24**(11): 2825-2831.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-575.

Quintero, A., D. Hubschmann, N. Kurzawa, S. Steinhauser, P. Rentzsch, S. Kramer, C. Andresen, J. Park, R. Eils, M. Schlesner and C. Herrmann (2020). "ShinyButchR: Interactive NMF-based decomposition workflow of genome-scale datasets." Biol Methods Protoc **5**(1): bpaa022.

Rao, M. S., T. R. Van Vleet, R. Ciurlionis, W. R. Buck, S. W. Mittelstadt, E. A. G. Blomme and M. J. Liguori (2018). "Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies." Front Genet **9**: 636.

Ripke, S., J. T. Walters, M. C. O'Donovan and S. W. G. o. t. P. G. Consortium (2020). "Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia." MedRxiv.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." Nucleic Acids Research **43**(7): e47-e47.

Romera-Paredes, B., H. Aung, N. Bianchi-Berthouze and M. Pontil (2013). Multilinear multitask learning. International Conference on Machine Learning.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the national academy of sciences **74**(12): 5463-5467.

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). "Biological insights from 108 schizophrenia-associated genetic loci." Nature **511**(7510): 421-427.

Schwarz, E., R. Izmailov, P. Lio and A. Meyer-Lindenberg (2016). "Protein Interaction Networks Link Schizophrenia Risk Loci to Synaptic Function." Schizophr Bull.

Seeman, P. (2004). "Atypical antipsychotics: mechanism of action." Focus **47**(1): 27-58.

Simon, N., J. Friedman, T. Hastie and R. Tibshirani (2013). "A Sparse-Group Lasso." Journal of Computational and Graphical Statistics **22**(2): 231-245.

Sivakumaran, S., F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson and H. Campbell (2011). "Abundant pleiotropy in human complex diseases and traits." The American Journal of Human Genetics **89**(5): 607-618.

Smith, V., C.-K. Chiang, M. Sanjabi and A. S. Talwalkar (2017). Federated multi-task learning. Advances in Neural Information Processing Systems.

Smith, V., S. Forte, M. Chenxin, M. Takáč, M. I. Jordan and M. Jaggi (2018). "CoCoA: A general framework for communication-efficient distributed optimization." Journal of Machine Learning Research **18**: 230.

Sonderby, I. E., D. van der Meer, C. Moreau, T. Kaufmann, G. B. Walters, M. Ellegaard, A. Abdellaoui, D. Ames, K. Amunts, M. Andersson, N. J. Armstrong, M. Bernard, N. B. Blackburn, J. Blangero, D. I. Boomsma, H. Brodaty, R. M. Brouwer, R. Bulow, R. Boen, W. Cahn, V. D. Calhoun, S. Caspers, C. R. K. Ching, S. Cichon, S. Ciufolini, B. Crespo-Facorro, J. E. Curran, A. M. Dale, S. Dalvie, P. Dazzan, E. J. C. de Geus, G. I. de Zubicaray, S. M. C. de Zwarte, S. Desrivieres, J. L. Doherty, G. Donohoe, B. Draganski, S. Ehrlich, E. Eising, T. Espeseth, K. Fejgin, S. E. Fisher, T. Fladby, O. Frei, V. Frouin, M. Fukunaga, T. Gareau, T. Ge, D. C. Glahn, H. J. Grabe, N. A. Groenewold, O. Gustafsson, J. Haavik, A. K. Haberg, J. Hall, R. Hashimoto, J. Y. Hehir-Kwa, D. P. Hibar, M. H. J. Hillegers, P. Hoffmann, L. Holleran, A. J. Holmes, G. Homuth, J. J. Hottenga, H. E. Hulshoff Pol, M. Ikeda, N. Jahanshad, C. Jockwitz, S. Johansson, E. G. Jonsson, N. R. Jorgensen, M. Kikuchi, E. E. M. Knowles, K. Kumar, S. Le Hellard, C. Leu, D. E. J. Linden, J. Liu, A. Lundervold, A. J. Lundervold, A. M. Maillard, N. G. Martin, S. Martin-Brevet, K. A. Mather, S. R. Mathias, K. L. McMahon, A. F. McRae, S. E. Medland, A. Meyer-Lindenberg, T. Moberget, C. Modenato, J. M. Sanchez, D. W. Morris, T. W. Muhleisen, R. M. Murray, J. Nielsen, J. E. Nordvik, L. Nyberg, L. M. O. Loohuis, R. A. Ophoff, M. J. Owen, T. Paus, Z. Pausova, J. M. Peralta, G. B. Pike, C. Prieto, E. B. Quinlan, C. S. Reinbold, T. R. Marques, J. J. H. Rucker, P. S. Sachdev, S. B. Sando, P. R. Schofield, A. J. Schork, G. Schumann, J. Shin, E. Shumskaya, A. I. Silva, S. M. Sisodiya, V. M. Steen, D. J. Stein, L. T. Strike, I. K. Suzuki, C. K. Tamnes, A. Teumer, A. Thalamuthu, D. Tordesillas-Gutierrez, A. Uhlmann, M. O. Ulfarsson, D. van 't Ent, M. B. M. van den Bree, P. Vanderhaeghen, E. Vassos, W. Wen, K. Wittfeld, M. J. Wright, I. Agartz, S. Djurovic, L. T. Westlye, H. Stefansson, K. Stefansson, S. Jacquemont, P. M. Thompson, O. A. Andreassen and E.-C. w. group (2021). "1q21.1 distal copy number variants are associated with cerebral and cognitive alterations in humans." Transl Psychiatry **11**(1): 182.

Stepniak, B., S. Papiol, C. Hammer, A. Ramin, S. Everts, L. Hennig, M. Begemann and H. Ehrenreich (2014). "Accumulated environmental risk determining age at schizophrenia onset: a deep phenotyping-based study." The Lancet Psychiatry **1**(6): 444-453.

Stilo, S. A. and R. M. Murray (2019). "Non-Genetic Factors in Schizophrenia." Curr Psychiatry Rep **21**(10): 100.

Sullivan, P. F. (2010). "The psychiatric GWAS consortium: big science comes to psychiatry." Neuron **68**(2): 182-186.

Tandon, R., H. A. Nasrallah and M. S. Keshavan (2009). "Schizophrenia, "just the facts" 4. Clinical features and conceptualization." Schizophr Res **110**(1-3): 1-23.

Tang, B., C. Capitao, B. Dean and E. A. Thomas (2012). "Differential age- and disease-related effects on the expression of genes related to the arachidonic acid signaling pathway in schizophrenia." Psychiatry Res **196**(2-3): 201-206.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Tibshirani, R. (2011). "Regression shrinkage and selection via the lasso: a retrospective." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(3): 273-282.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu and K. Knight (2005). "Sparsity and smoothness via the fused lasso." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(1): 91-108.

Tibshirani, R. J. (2013). "The lasso problem and uniqueness." Electronic Journal of Statistics **7**(0): 1456-1490.

Vilhjalmsson, B. J., J. Yang, H. K. Finucane, A. Gusev, S. Lindstrom, S. Ripke, G. Genovese, P. R. Loh, G. Bhatia, R. Do, T. Hayeck, H. H. Won, D. B. Schizophrenia Working Group of the Psychiatric Genomics Consortium, s. Risk of Inherited Variants in Breast Cancer, S. Kathiresan, M. Pato, C. Pato, R. Tamimi, E. Stahl, N. Zaitlen, B. Pasaniuc, G. Belbin, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, P. M. Visscher, P. Kraft, N. Patterson and A. L. Price (2015). "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." Am J Hum Genet **97**(4): 576-592.

Vos, T., A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, J. Abraham, I. Ackerman, R. Aggarwal, S. Y. Ahn, M. K. Ali, M. Alvarado, H. R. Anderson, L. M. Anderson, K. G. Andrews, C. Atkinson, L. M. Baddour, A. N. Bahalim, S. Barker-Collo, L. H. Barrero, D. H. Bartels, M. G. Basanez, A. Baxter, M. L. Bell, E. J. Benjamin, D. Bennett, E. Bernabe, K. Bhalla, B. Bhandari, B. Bikbov, A. Bin Abdulhak, G. Birbeck, J. A. Black, H. Blencowe, J. D. Blore, F. Blyth, I. Bolliger, A. Bonaventure, S. Boufous, R. Bourne, M. Boussinesq, T. Braithwaite, C. Brayne, L. Bridgett, S. Brooker, P. Brooks, T. S. Brugha, C. Bryan-Hancock, C. Bucello, R. Buchbinder, G. Buckle, C. M. Budke, M. Burch, P. Burney, R. Burstein, B. Calabria, B. Campbell, C. E. Canter, H. Carabin, J. Carapetis, L. Carmona, C. Cella, F. Charlson, H. Chen, A. T. Cheng, D. Chou, S. S. Chugh, L. E. Coffeng, S. D. Colan, S. Colquhoun, K. E. Colson, J. Condon, M. D. Connor, L. T. Cooper, M. Corriere, M. Cortinovis, K. C. de Vaccaro, W. Couser, B. C. Cowie, M. H. Criqui, M. Cross, K. C. Dabhadkar, M. Dahiya, N. Dahodwala, J. Damsere-Derry, G. Danaei, A. Davis, D. De Leo, L. Degenhardt, R. Dellavalle, A. Delossantos, J. Denenberg, S. Derrett, D. C. Des Jarlais, S. D. Dharmaratne, M. Dherani, C. Diaz-Torne, H. Dolk, E. R. Dorsey, T. Driscoll, H. Duber, B. Ebel, K. Edmond, A. Elbaz, S. E. Ali, H. Erskine, P. J. Erwin, P. Espindola, S. E. Ewoigbokhan, F. Farzadfar, V. Feigin, D. T. Felson, A. Ferrari, C. P. Ferri, E. M. Fevre, M. M. Finucane, S. Flaxman, L. Flood, K. Foreman, M. H. Forouzanfar, F. G. Fowkes, R. Franklin, M. Fransen, M. K. Freeman, B. J. Gabbe, S. E. Gabriel, E. Gakidou, H. A. Ganatra, B. Garcia, F. Gaspari, R. F. Gillum, G. Gmel, R. Gosselin, R. Grainger, J. Groeger, F. Guillemin, D. Gunnell, R. Gupta, J. Haagsma, H. Hagan, Y. A. Halasa, W. Hall, D. Haring, J. M. Haro, J. E. Harrison, R. Havmoeller, R. J. Hay, H. Higashi, C. Hill, B. Hoen, H. Hoffman, P. J. Hotez, D. Hoy, J. J. Huang, S. E. Ibeanusi, K. H. Jacobsen, S. L. James, D. Jarvis, R. Jasrasaria, S. Jayaraman, N. Johns, J. B. Jonas, G. Karthikeyan, N. Kassebaum, N. Kawakami, A. Keren, J. P. Khoo, C. H. King, L. M. Knowlton, O. Kobusingye, A. Koranteng, R. Krishnamurthi, R. Lalloo, L. L. Laslett, T. Lathlean, J. L. Leasher, Y. Y. Lee, J. Leigh, S. S. Lim, E. Limb, J. K. Lin, M. Lipnick, S. E. Lipshultz, W. Liu, M. Loane, S. L. Ohno, R. Lyons, J. Ma, J. Mabweijano, M. F. MacIntyre, R. Malekzadeh, L. Mallinger, S. Manivannan, W. Marcenes, L. March, D. J. Margolis, G. B. Marks, R. Marks, A. Matsumori, R. Matzopoulos, B. M. Mayosi, J. H. McAnulty, M. M. McDermott, N. McGill, J. McGrath, M. E. Medina-Mora, M. Meltzer, G. A. Mensah, T. R. Merriman, A. C. Meyer, V. Miglioli, M. Miller, T. R. Miller, P. B. Mitchell, A. O. Mocumbi, T. E. Moffitt, A. A. Mokdad, L. Monasta, M. Montico, M. Moradi-

Lakeh, A. Moran, L. Morawska, R. Mori, M. E. Murdoch, M. K. Mwaniki, K. Naidoo, M. N. Nair, L. Naldi, K. M. Narayan, P. K. Nelson, R. G. Nelson, M. C. Nevitt, C. R. Newton, S. Nolte, P. Norman, R. Norman, M. O'Donnell, S. O'Hanlon, C. Olives, S. B. Omer, K. Ortblad, R. Osborne, D. Ozgediz, A. Page, B. Pahari, J. D. Pandian, A. P. Rivero, S. B. Patten, N. Pearce, R. P. Padilla, F. Perez-Ruiz, N. Perico, K. Pesudovs, D. Phillips, M. R. Phillips, K. Pierce, S. Pion, G. V. Polanczyk, S. Polinder, C. A. Pope, 3rd, S. Popova, E. Porrini, F. Pourmalek, M. Prince, R. L. Pullan, K. D. Ramaiah, D. Ranganathan, H. Razavi, M. Regan, J. T. Rehm, D. B. Rein, G. Remuzzi, K. Richardson, F. P. Rivara, T. Roberts, C. Robinson, F. R. De Leon, L. Ronfani, R. Room, L. C. Rosenfeld, L. Rushton, R. L. Sacco, S. Saha, U. Sampson, L. Sanchez-Riera, E. Sanman, D. C. Schwebel, J. G. Scott, M. Segui-Gomez, S. Shahraz, D. S. Shepard, H. Shin, R. Shivakoti, D. Singh, G. M. Singh, J. A. Singh, J. Singleton, D. A. Sleet, K. Sliwa, E. Smith, J. L. Smith, N. J. Stapelberg, A. Steer, T. Steiner, W. A. Stolk, L. J. Stovner, C. Sudfeld, S. Syed, G. Tamburlini, M. Tavakkoli, H. R. Taylor, J. A. Taylor, W. J. Taylor, B. Thomas, W. M. Thomson, G. D. Thurston, I. M. Tleyjeh, M. Tonelli, J. A. Towbin, T. Truelsen, M. K. Tsilimbaris, C. Ubeda, E. A. Undurraga, M. J. van der Werf, J. van Os, M. S. Vavilala, N. Venketasubramanian, M. Wang, W. Wang, K. Watt, D. J. Weatherall, M. A. Weinstock, R. Weintraub, M. G. Weisskopf, M. M. Weissman, R. A. White, H. Whiteford, S. T. Wiersma, J. D. Wilkinson, H. C. Williams, S. R. Williams, E. Witt, F. Wolfe, A. D. Woolf, S. Wulf, P. H. Yeh, A. K. Zaidi, Z. J. Zheng, D. Zonies, A. D. Lopez, C. J. Murray, M. A. AlMazroa and Z. A. Memish (2012). "Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010." Lancet **380**(9859): 2163-2196.

Wang, D., S. Liu, J. Warrell, H. Won, X. Shi, F. C. P. Navarro, D. Clarke, M. Gu, P. Emani, Y. T. Yang, M. Xu, M. J. Gandal, S. Lou, J. Zhang, J. J. Park, C. Yan, S. K. Rhie, K. Manakongtreecheep, H. Zhou, A. Nathan, M. Peters, E. Mattei, D. Fitzgerald, T. Brunetti, J. Moore, Y. Jiang, K. Girdhar, G. E. Hoffman, S. Kalayci, Z. H. Gumus, G. E. Crawford, E. C. Psych, P. Roussos, S. Akbarian, A. E. Jaffe, K. P. White, Z. Weng, N. Sestan, D. H. Geschwind, J. A. Knowles and M. B. Gerstein (2018). "Comprehensive functional genomic resource and integrative model for the human brain." Science **362**(6420).

Wang, H., F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen and I. Alzheimer's Disease Neuroimaging (2012). "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort." Bioinformatics **28**(2): 229-237.

Wang, X., C. Zhang and Z. Zhang. (2009). Boosted multi-task learning for face verification with applications to web image and video search. Proceedings of IEEE Computer Society Conference on Computer Vision and Patter Recognition.

Warnat-Herresthal, S., H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers, N. A. Aziz, S. Ktena, C. Siever, M. Kraut, M. Desai, B. Monnet, M. Saridaki, C. M. Siegel, A. Drews, M. Nuesch-Germano, H. Theis, M. G. Netea, F. Theis, A. C. Aschenbrenner, T. Ulas, M. M. B. Breteler, E. J. Giamarellos-Bourboulis, M. Kox, M. Becker, S. Cheran, M. S. Woodacre, E. L. Goh and J. L. Schultze (2020). "Swarm Learning as a privacy-preserving machine learning approach for disease classification."

Warnat-Herresthal, S., H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Handler, P. Pickkers, N. A. Aziz, S. Ktena, F. Tran, M. Bitzer, S. Ossowski, N.

Casadei, C. Herr, D. Petersheim, U. Behrends, F. Kern, T. Fehlmann, P. Schommers, C. Lehmann, M. Augustin, J. Rybniker, J. Altmuller, N. Mishra, J. P. Bernardes, B. Kramer, L. Bonaguro, J. Schulte-Schrepping, E. De Domenico, C. Siever, M. Kraut, M. Desai, B. Monnet, M. Saridaki, C. M. Siegel, A. Drews, M. Nuesch-Germano, H. Theis, J. Heyckendorf, S. Schreiber, S. Kim-Hellmuth, C.-A. Study, J. Nattermann, D. Skowasch, I. Kurth, A. Keller, R. Bals, P. Nurnberg, O. Riess, P. Rosenstiel, M. G. Netea, F. Theis, S. Mukherjee, M. Backes, A. C. Aschenbrenner, T. Ulas, C.-O. I. Deutsche, M. M. B. Breteler, E. J. Giamarellos-Bourboulis, M. Kox, M. Becker, S. Cheran, M. S. Woodacre, E. L. Goh and J. L. Schultze (2021). "Swarm Learning for decentralized and confidential clinical machine learning." Nature **594**(7862): 265-270.

Watanabe, K., S. Stringer, O. Frei, M. Umicevic Mirkov, C. de Leeuw, T. J. C. Polderman, S. van der Sluis, O. A. Andreassen, B. M. Neale and D. Posthuma (2019). "A global overview of pleiotropy and genetic architecture in complex traits." Nat Genet **51**(9): 1339-1348.

Whelan, R., R. Watts, C. A. Orr, R. R. Althoff, E. Artiges, T. Banaschewski, G. J. Barker, A. L. Bokde, C. Buchel, F. M. Carvalho, P. J. Conrod, H. Flor, M. Fauth-Buhler, V. Frouin, J. Gallinat, G. Gan, P. Gowland, A. Heinz, B. Ittermann, C. Lawrence, K. Mann, J. L. Martinot, F. Nees, N. Ortiz, M. L. Paillere-Martinot, T. Paus, Z. Pausova, M. Rietschel, T. W. Robbins, M. N. Smolka, A. Strohle, G. Schumann, H. Garavan and I. Consortium (2014). "Neuropsychosocial profiles of current and future adolescent alcohol misusers." Nature **512**(7513): 185-189.

Widmer, C. Multitask Learning in Computational Biology, International Machine Learning Society.

Widmer, C., M. Kloft, N. Görnitz and G. Rätsch (2012). "Efficient Training of Graph-Regularized Multitask SVMs." **7523**: 633-647.

Widmer, C. and G. Ratsch (2012). "Multitask learning in computational biology." JMLR(27): 207–216.

Wilson, R. C., O. W. Butters, D. Avraam, J. Baker, J. A. Tedds, A. Turner, M. Murtagh and P. R. Burton (2017). "DataSHIELD – New Directions and Dimensions." Data Science Journal **16**.
Wittchen, H.-U., F. Jacobi, J. Rehm, A. Gustavsson, M. Svensson, B. Jönsson, J. Olesen, C. Allgulander, J. Alonso and C. Faravelli (2011). "The size and burden of mental disorders and other disorders of the brain in Europe 2010." European neuropsychopharmacology **21**(9): 655-679.

Wolfers, T., J. K. Buitelaar, C. F. Beckmann, B. Franke and A. F. Marquand (2015). "From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics." Neurosci Biobehav Rev **57**: 328-349.

Wu, Q., Z. Wang, C. Li, Y. Ye, Y. Li and N. Sun (2015). "Protein functional properties prediction in sparsely-label PPI networks through regularized non-negative matrix factorization." BMC Syst Biol **9 Suppl 1**: S9.

Wu, T., E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo and G. Yu (2021). "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data." Innovation (N Y) **2**(3): 100141.

Wu, Z., C. Valentini-Botinhao and O. Watts (2015). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP).

Wu, Z., C. Valentini-Botinhao, O. Watts and S. King (2015). "Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis." 4460-4464.
Xia, C. H., Z. Ma, R. Ciric, S. Gu, R. F. Betzel, A. N. Kaczkurkin, M. E. Calkins, P. A. Cook, A. Garcia de la Garza, S. N. Vandekar, Z. Cui, T. M. Moore, D. R. Roalf, K. Ruparel, D. H. Wolf, C. Davatzikos, R. C. Gur, R. E. Gur, R. T. Shinohara, D. S. Bassett and T. D. Satterthwaite (2018). "Linked dimensions of psychopathology and connectivity in functional brain networks." Nat Commun **9**(1): 3003.

Xiaogang, W., Z. Cha and Z. Zhengyou (2009). "Boosted multi-task learning for face verification with applications to web image and video search." 142-149.

Xie, L., I. M. Baytas, K. Lin and J. Zhou (2017). "Privacy-Preserving Distributed Multi-Task Learning with Asynchronous Updates." 1195-1204.

Xu, Q., S. Pan, H. Xue and Q. Yang (2011). "Multitask learning for protein subcellular location prediction." IEEE/ACM Trans Comput Biol Bioinform (8): 748–759.

Xu, Q., S. J. Pan, H. H. Xue and Q. Yang (2011). "Multitask Learning for Protein Subcellular Location Prediction." IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS.

Xu, Q., H. Xue and Q. Yang (2011). "Multi-platform gene-expression mining and marker gene analysis." Int J Data Min Bioinform **5**(5): 485-503.

Xue, G., C. Chen, Z. L. Lu and Q. Dong (2010). "Brain Imaging Techniques and Their Applications in Decision-Making Research." Xin Li Xue Bao **42**(1): 120-137.

Yang, E. and P. Ravikumar (2013). Dirty Statistical Models. NIPS.

Yang, S., L. Yuan, Y. C. Lai, X. Shen, P. Wonka and J. Ye (2012). "Feature Grouping and Selection Over an Undirected Graph." KDD: 922-930.

Yang, T., J. Liu, P. Gong, R. Zhang, X. Shen and J. Ye (2016). "Absolute Fused Lasso and Its Application to Genome-Wide Association Studies." 1955-1964.

Yang, Y. and T. Hospedales (2016). "Deep multi-task representation learning: A tensor factorisation approach." arXiv preprint arXiv:1605.06391.

Yang, Z. and G. Michailidis (2016). "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data." Bioinformatics **32**(1): 1-8.

Yuan, H., I. Paskov, H. Paskov, A. J. Gonzalez and C. S. Leslie (2016). "Multitask learning improves prediction of cancer drug sensitivity." Sci Rep **6**: 31619.

Zhang, C. and J. Liu (2020). "Distributed Learning Systems with First-Order Methods." Foundations and Trends® in Databases **9**(1): 1-100.

Zhang, Y. and D.-Y. Yeung (2012). "A convex formulation for learning task relationships in multi-task learning." arXiv preprint arXiv:1203.3536.

Zhang, Z., P. Luo, C. C. Loy and X. Tang (2014). "Facial Landmark Detection by Deep Multi-task Learning." **8694**: 94-108.

Zheng, J., A. M. Erzurumluoglu, B. L. Elsworth, J. P. Kemp, L. Howe, P. C. Haycock, G. Hemani, K. Tansey, C. Laurin, G. Early, C. Lifecourse Epidemiology Eczema, B. S. Pourcain, N. M. Warrington, H. K. Finucane, A. L. Price, B. K. Bulik-Sullivan, V. Anttila, L. Paternoster, T. R. Gaunt, D. M. Evans and B. M. Neale (2017). "LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis." Bioinformatics **33**(2): 272-279.

Zhou, J., J. Chen and J. Ye (2011). Clustered Multi-Task Learning Via Alternating Structure Optimization. Advances in Neural Information Processing Systems 24 (NIPS 2011).

Zhou, J., J. Chen and J. Ye (2011). "Malsar: Multi-task learning via structural regularization." Arizona State University **21**.

Zhou, J., J. Chen and J. Ye (2012). MALSAR: Multi-tAsk Learning via StructurAl Regularization, Arizona State University.

Zhou, J., J. Liu, V. A. Narayan and J. Ye (2012). "Modeling Disease Progression via Fused Sparse Group Lasso." KDD **2012**: 1095-1103.

Zhou, J., J. Liu, V. A. Narayan, J. Ye and I. Alzheimer's Disease Neuroimaging (2013). "Modeling disease progression via multi-task learning." Neuroimage **78**: 233-248.

Zhou, J., L. Yuan, J. Liu and J. Ye (2011). "A multi-task learning formulation for predicting disease progression." 814.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2): 301-320.

8 PUBLICATIONS

Cao, H., Zhang, Y., Baumbach, J., Burton, P.R., Dwyer, D., Koutsouleris, N., Matschinske, J., Marcon, Y., Rajan, S., Rieg, T. and Ryser-Welch, P., 2021. dsMTL-a computational framework for privacy-preserving, distributed multi-task machine learning. bioRxiv. (**In Review**)

Cao, H., Zhou, J. and Schwarz, E., 2019. RMTL: an R library for multi-task learning. Bioinformatics, 35(10), pp.1797-1798.

Cao, H., Meyer-Lindenberg, A. and Schwarz, E., 2018. Comparative evaluation of machine learning strategies for analyzing big data in psychiatry. International journal of molecular sciences, 19(11), p.3387.

Cao, H., Chen, J., Meyer-Lindenberg, A. and Schwarz, E., 2017. A polygenic score for schizophrenia predicts glycemic control. *Translational psychiatry*, *7*(12), pp.1-9.

Cao, H. and Schwarz, E., 2020. Opportunities and challenges of machine learning approaches for biomarker signature identification in psychiatry. In *Personalized Psychiatry* (pp. 117-126). Academic Press.

Cao, H. and Schwarz, E., 2019. Computational Approaches for Identification of Pleiotropic Biomarker Profiles in Psychiatry. *Reviews on Biomarker Studies of Metabolic and Metabolism-Related Disorders*, pp.111-128.

Cao, H., Hong, X., Tost, H., Meyer-Lindenberg, A., Schwarz, E., 2021. Advancing translational research in psychiatry through multi-task learning: a systematic review (**In Review**)

Chen, J., **Cao, H.**, Kaufmann, T., Westlye, L.T., Tost, H., Meyer-Lindenberg, A. and Schwarz, E., 2020. Identification of reproducible BCL11A alterations in schizophrenia through individual-level prediction of coexpression. *Schizophrenia bulletin*, *46*(5), pp.1165-1171.

Chen, J., **Cao, H.**, Meyer-Lindenberg, A. and Schwarz, E., 2018. Male increase in brain gene expression variability is linked to genetic risk for schizophrenia. *Translational psychiatry*, *8*(1), pp.1-10.

Chen J, Zang Z, Braun U, Schwarz K, Harneit A, Kremer T, Ma R, Schweiger J, Moessnang C, Geiger L, **Cao H.** Association of a reproducible epigenetic risk profile for schizophrenia with brain methylation and function. JAMA psychiatry. 2020 Jun 1;77(6):628-36.

Chen, J., Schwarz, K., Zang, Z., Braun, U., Harneit, A., Kremer, T., Ma, R., Schweiger, J., Moessnang, C., Geiger, L. and **Cao, H.**, 2021. Hyper-coordinated DNA methylation is altered in schizophrenia and associated with brain function. *Schizophrenia Bulletin Open*.

Schwarz, E., Alnæs, D., Andreassen, O.A., **Cao, H.**, Chen, J., Degenhardt, F., Doncevic, D., Dwyer, D., Eils, R., Erdmann, J. and Herrmann, C., 2021. Identifying multimodal signatures underlying the

somatic comorbidity of psychosis: the COMMITMENT roadmap. *Molecular Psychiatry*, *26*(3), pp.722-724.

Schwarz, K., Moessnang, C., Schweiger, J.I., Harneit, A., Schneider, M., Chen, J., **Cao, H.**, Schwarz, E., Witt, S.H., Rietschel, M. and Nöthen, M., 2021. Ventral striatal-hippocampus coupling during reward processing as a (stratification) biomarker for psychotic disorders. *Biological Psychiatry*.

Gass, N., Peterson, Z., Sartorius, A., Weber-Fahr, W., Reinwald, J.R., Sack, M., Chen, J., **Cao, H.**, Didriksen, M., Stensbøl, T.B. and Klemme, G., 2021. Identifying Polygenic Contributions to Differential Resting-State Connectivity in a Mouse Model of 22q11. 2 Deletion. *Biological Psychiatry*, *89*(9), pp.S291-S292.

Gass, N., Peterson, Z., Reinwald, J., Sartorius, A., Weber-Fahr, W., Sack, M., Chen, J., **Cao, H**., Didriksen, M., Stensbøl, T.B. and Klemme, G., 2021. Differential resting-state patterns across networks are spatially associated with Comt and Trmt2a gene expression patterns in a mouse model of 22q11. 2 deletion. *NeuroImage*, p.118520.

9 CURRICULUM VITAE

Personal Information

Name: Han Cao
Date of Birth: 20.09.1987
Place of Birth: Henan, China
Nationality: Chinese

Academic Education

- 02/2016 – present

PhD student, Department of Psychiatry and Psychotherapy, research group Systems Neuroscience in Psychiatry; Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Germany
Supervisor: Dr. Emanuel Schwarz

- 09/2019 – 06/2014

Master student, Software Engineering, Department of Computer and Information Science, University of Macau, Macau, China
Supervisor: Dr. Shirley Weng In Siu

- 09/2007 – 06/2011

Bachelor student, Software Engineering, Department of Computer and Information Science, University of Macau, Macau, China
Supervisor: Dr. Wen Wu

10 ACKNOWLEDGEMENT

First, I would like to thank my supervisor Dr. Emanuel Schwarz who gives me this opportunity to pursue my scientific career, always supports my ideas and work even when I was a research rookie, gives a lot of constructive suggestions when I got stuck and so on. Dr. Schwarz also played an essential role in forming my research taste, working habits and attitudes for collaboration. Besides, Dr. Schwarz also helped me expand the research network and get me involved in many academic collaborations.

Second, I want to thank Dr. Junfang Chen as well as members of SNiP group. I am appreciated Junfang introduced me to this great group and helped me with his broad knowledge of genetics and biology. I also want to thank my great co-authors Ms Kristina Schwartz and Ms Anais Harneit. They explained the data setting of neuroimaging and genetics in SNiP group and the related technical knowledge in detail. I also appreciated Ms Mirjam Melzer, who helped me a lot with administrative stuff.

Third, I appreciated the company and support of my friends Ren Ma, Xiaolong Zhang, Zhenxiang Zang, my girlfriend, and my parents. Without your patience and support, I cannot finish this.

Last, I want to thank all co-authors of my research work and members of the COMMITMENT consortium. It is my honor to collaborate with you, and I learned a lot from your expertise.

07.09.2021
Mannheim, Germany